# Data Assimilation

William Lahoz · Boris Khattatov ·
Richard Ménard
Editors

# Data Assimilation

## Making Sense of Observations

Springer

*Editors*
Dr. William Lahoz
Norsk Institutt for Luftforskning, Norwegian Institute for Air Research
Atmospheric and Climate Research
Instituttveien 18
Kjeller 2007
Norway
wal@nilu.no

Dr. Boris Khattatov
373 Arapahoe Avenue
Boulder CO 80302
USA
boris@fusionnumerics.com

Dr. Richard Ménard
Environment Canada
Atmospheric Science & Technology Directorate
2121 Trans-Canada Highway
Dorval QC H9P 1J3
Canada
richard.menard@ec.gc.ca

ISBN 978-3-540-74702-4 e-ISBN 978-3-540-74703-1
DOI 10.1007/978-3-540-74703-1
Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2010925150



*Cover design*: Bauer, Thomas

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

# Contents

# Contributors

**Erik Andersson** European Centre for Medium-Range Weather Forecasts (ECMWF), Reading, UK, erik.andersson@ecmwf.int

**Michael G. Bosilovich** NASA Goddard Space Flight Center, Greenbelt, MD, USA, Michael.Bosilovich@nasa.gov

**Mark Buehner** Meteorological Research Division, Data Assimilation and Satellite Meteorology Section, Environment Canada, Canada, mark.buehner@ec.gc.ca

**Andrew Charlton-Perez** Department of Meteorology, University of Reading, Reading, UK, a.j.charlton@reading.ac.uk

**Stephen E. Cohn** Global Modeling and Assimilation Office, NASA Goddard Space Flight Center, Greenbelt, MD 20771, USA, stephen.e.cohn@nasa.gov

**Gabriëlle J.M. De Lannoy** George Mason University, Fairfax, VA, USA; Faculty of Bioscience Engineering, Ghent University, Ghent B-9000, Belgium, gdlannoy@cola.iges.org

**Hendrik Elbern** Research Centre Jülich, Rhenish Institute for Environmental Research at the University of Cologne (RIU), Cologne, Germany; Helmholtz Virtual Institute for Inverse Modelling of Atmospheric Chemical Composition (IMACCO), Cologne, Germany, he@eurad.uni-koeln.de

**G. David Emmitt** Simpson Weather Associates (SWA), Charlottesville, VA, USA, gde@swa.com

**Quentin Errera** Belgian Institute for Space Aeronomy, BIRA-IASB, Brussels, Belgium, quentin@aeronomie.be

**Ronald M. Errico** NASA/GSFC, Greenbelt, MD, USA; Goddard Earth Science and Technology Center, University of Maryland, Baltimore, MD, USA, Ronald.M.Errico@nasa.gov

**Keith Haines** Environmental Systems Science Centre, University of Reading, Reading RG6 6AL, UK, kh@mail.nerc-essc.ac.uk

**Paul R. Houser** George Mason University, Fairfax, VA, USA, phouser@gmu.edu

**Xiang-Yu Huang** National Center for Atmospheric Research, Boulder, CO, USA, huangx@ucar.edu

**Eugenia Kalnay** University of Maryland, College Park, MD 20742-2425, USA, ekalnay@atmos.umd.edu

**Boris Khatattov** Fusion Numerics Inc, Boulder, CO, USA, boris@fusionnumerics.com

**William Lahoz** Norsk Institutt for Luftforskning, Norwegian Institute for Air Research, NILU, Kjeller, Norway, wal@nilu.no

**Stephen R. Lewis** Department of Physics and Astronomy, The Open University, Milton Keynes MK7 6AA, UK, S.R.Lewis@open.ac.uk

**Stephen J. Lord** NOAA/NWS/NCEP/EMC, Camp Springs, MD, USA, Stephen.Lord@noaa.gov

**Peter Lynch** University College Dublin, Dublin, Ireland, peter.lynch@ucd.ie

**Michiko Masutani** NOAA/NWS/NCEP/EMC, Camp Springs, MD, USA; Wyle Information Systems, El Segundo, CA, USA, Michiko.Masutani@noaa.gov

**Richard Ménard** Air Quality Research Division, Meteorological Service of Canada, Environment Canada, Dorval, Canada, richard.menard@ec.gc.ca

**Nancy K. Nichols** Department of Mathematics, University of Reading, Reading, UK, n.k.nichols@reading.ac.uk

**Lars Nieradzik** Research Centre Jülich, Rhenish Institute for Environmental Research at the University of Cologne (RIU), Cologne, Germany; Helmholtz Virtual Institute for Inverse Modelling of Atmospheric Chemical Composition (IMACCO), Cologne, Germany, e-mail: ln@eurad.unikoeln.de

**Lars-Peter Riishøjgaard** NASA/GSFC, Greenbelt, MD, USA; Goddard Earth Science and Technology Center, University of Maryland, Baltimore, MD, USA; Joint Center for Satellite Data Assimilation, Camp Springs, MD, USA, Lars.P.Riishojgaard@nasa.gov

**Richard B. Rood** University of Michigan, Ann Arbor, MI, USA, rbrood@umich.edu

**Thomas W. Schlatter** NOAA/Earth System Research Laboratory, Boulder, CO, USA, Tom.Schlatter@noaa.gov

**Ad Stoffelen** Royal Dutch Meteorological Institute (KNMI), DeBilt, The Netherlands, Ad.Stoffelen@knmi.nl

**Achim Strunk** Research Centre Jülich, Rhenish Institute for Environmental Research at the University of Cologne (RIU), Cologne, Germany; Helmholtz

Virtual Institute for Inverse Modelling of Atmospheric Chemical Composition (IMACCO), Cologne, Germany, as@eurad.uni-koeln.de

**Richard Swinbank** Met Office, Exeter, UK, richard.swinbank@metoffice.gov.uk

**Olivier Talagrand** Laboratoire de Météorologie Dynamique/CNRS, École Normale Supérieure, Paris, France, Talagrand@lmd.ens.fr

**Jean-Noël Thépaut** European Centre for Medium-Range Weather Forecasts, ECMWF, Shinfield, UK, jean-noel.thepaut@ecmwf.int

**Jeffrey P. Walker** Department of Civil and Environmental Engineering, The University of Melbourne, Victoria, Australia, j.walker@unimelb.edu.au

**John S. Woollen** NOAA/NWS/NCEP/EMC, Camp Springs, MD, USA; Science Applications International Corporation (SAIC), Mclean, VA, USA, Jack.Woollen@noaa.gov

**Valery Yudin** SAIC, Global Modeling Assimilation Office, Code 610.1, Goddard Space Flight Center, Greenbelt, MD 20771, USA; Atmospheric Chemistry Division, National Center for Atmospheric Research, Boulder, CO, USA, vyudin@ucar.edu

# Introduction

**William Lahoz, Boris Khattatov, and Richard Ménard**

This book came from a request from Springer to the editors to update knowledge on the science of data assimilation and incorporate developments during the last 5 years. It is designed to update the science of data assimilation since the NATO (North Atlantic Treaty Organization) Science Series Book "Data Assimilation for the Earth System" (R. Swinbank, V. Shutyaev, W.A. Lahoz, eds.) came out in 2003, and fill in some of the gaps in that book. The NATO Science Series Book was based on a set of lectures presented at the NATO Advanced Study Institute (ASI) on *Data Assimilation for the Earth System*, which was held at Maratea, Italy during May–June 2002. That ASI grew out of a concern that there was little teaching available in data assimilation, even though it had become central to modern weather forecasting, and was becoming increasingly important in a range of other Earth disciplines such as the ocean, land and chemistry.

Many changes have happened in the science of data assimilation over the last 5 years. They include the development of chemical data assimilation systems at several centres world-wide, both research and operational; the increased interaction between the research and operational communities; the use of data assimilation to evaluate research satellite data; the use of data assimilation ideas, long applied to weather forecast models, to evaluate climate models; the combination of theoretical notions from variational methods and ensemble Kalman filter methods to improve data assimilation performance; and the increased extension of data assimilation to areas beyond the atmosphere and dynamics: chemistry, ionosphere, and other planets, e.g., Mars and Venus. There has also been a consolidation in the use of data assimilation to evaluate future observations, and in the use of data assimilation in areas such as the ocean and the land.

Parallel to these changes in the science of data assimilation, another remarkable change over the last 5 years has been the increased presence of data assimilation in teaching initiatives such as Summer Schools. These include the now biennial ESA (European Space Agency) Earth Observation Summer School

W. Lahoz (✉)
Norsk Institutt for Luftforskning, Norwegian Institute for Air Research, NILU, Kjeller, Norway
e-mail: wal@nilu.no

(http://envisat.esa.int/envschool_2008/) and several others. It can now be said that data assimilation has become a mainstream topic in the teaching of Earth Observation.

The NATO Science Series book, although useful and a feature in many university lecture courses, has some gaps. These include, for example, an overview of data assimilation and its relationship to information, either in observations or models; a discussion of ensemble Kalman filter methods; a discussion of Observing System Simulation Experiments (OSSEs); a discussion of tropospheric chemical data assimilation; and a discussion of meteorology and dynamics.

This book is intended to build on the material from the NATO Science Series book, address the above changes, and fill in the above gaps. Although there will be inevitable gaps in this book, we think it will provide a useful addition to the literature on data assimilation. To achieve this, we have asked world-leading data assimilation scientists to contribute to the chapters. We hope we succeed, at least until the next data assimilation book along these lines comes out in 5 years! Finally, we dedicate this book to Andrew Crook (1958–2006) who was one of the original chapter authors.

November 2009

# Part I
# Theory

# Data Assimilation and Information

**William Lahoz, Boris Khattatov, and Richard Ménard**

## 1 Introduction

In this introductory chapter we provide an overview of the connection between the *data assimilation* methodology and the concept of *information*, whether embodied in *observations* or *models*. In this context, we provide a step by step introduction to the need for data assimilation, culminating in an easy to understand description of the data assimilation methodology. Schematic diagrams and simple examples form a key part of this chapter.

The plan is to first discuss the need for *information*; then discuss sources of information; discuss the characteristics of this information, in particular the presence of "information gaps"; provide an objective underpinning to methods to fill in these information gaps; and discuss the benefits of combining different sources of information, in this case from *observations* that sample in space and time the system of interest (e.g. the atmosphere, the ocean, the land surface, the ionosphere, other planets), and *models* that embody our understanding of the system observed. Finally, we bring together these ideas under the heading of "data assimilation", provide a schematic of the methodology, and provide three simple examples highlighting how data assimilation *adds value*, the impact of *spatial resolution* on information, and the impact of *temporal sampling* on information.

At the end of this chapter we identify the foci of this book and the order in which they are presented in the book.

## 2 Need for Information

The main challenges to society, for example, *climate change*, *impact of extreme weather*, *environmental degradation* and *ozone loss*, require information for an intelligent response, including making choices on future action. Regardless of its

W. Lahoz (✉)
Norsk Institutt for Luftforskning, Norwegian Institute for Air Research, NILU, Kjeller, Norway
e-mail: wal@nilu.no

W. Lahoz et al. (eds.), *Data Assimilation*, DOI 10.1007/978-3-540-74703-1_1,

source, we wish to be able to use this information to make predictions for the future, test hypotheses, and attribute cause and effect. In this way, we are able to take action according to information provided on the future behaviour of the system of interest, and in particular future events (*prediction*); test our understanding of the system, and adjust this understanding according to new information (*hypothesis testing*); and understand the cause of events, and obtain information on possible ways of changing, mitigating or adjusting to the course of events (*attribute cause and effect*).

We can identify a generic *chain of information processing*:

- Gather information;
- Test hypotheses based on this information;
- Build methods to use this information to attribute cause and effect;
- Use these methods to make predictions.

However, we still need two ingredients: a means of gathering information, and methods to build on this information gathered. Roughly speaking, *observations* (measurements) provide the first ingredient, and *models* (conceptual, numerical or otherwise) provide the second ingredient. Note, however, that from the point of view of information, observations and models are not distinct; it is the mechanism of obtaining this information that is distinct: observations have a roughly *direct link* with the system of interest via the measurement process; models have a roughly *indirect link* with the system of interest, being an embodiment of information received from measurements, experience and theory.

## 3 Sources of Information

We have two broad sources of information: *measurements* of the system of interest ("observations"); and *understanding* of the temporal and spatial evolution of the system of interest ("models"). Further details about observations and models can be found in Part II, *Observations*, and Part III, *Meteorology and Atmospheric Dynamics*, respectively.

Observations (or measurements) sample the system of interest in space and time, with spatial and temporal scales dependent on the technique used to make the measurements. These measurements provide information on the system of interest and contribute to building an understanding of how the system evolves in space and time.

Understanding can be *qualitative*, e.g., how variables roughly "connect" or are related, or *quantitative*, commonly expressed in equations. A rough, qualitative connection can indicate that if the velocity of a particle increases, its kinetic energy also increases. A quantitative connection based on equations assigns a numerical relationship between the velocity and the kinetic energy, so that we can make precise

(subject to the accuracy of the calculation) the increase in kinetic energy given an increase in velocity of the particle. Equations can come from *general laws* (e.g. Newton's laws of motion), or *relations between parameters* (e.g. empirical or statistical). In general, quantification on the basis of laws tends to be more rigorous than quantification on the basis of empirical or statistical relations, mainly because laws have broad (if not universal) application, whereas empirical or statistical relations tend to apply only to specific cases.

## 4 Characteristics of Information

To make use of the information embodied in observations and models it is necessary to understand the characteristics of this information. In particular, we must recognize that both observations and models have *errors*. We now discuss briefly the nature of these errors.

Observations have errors which are characterized as *random* (also known as precision), *systematic* (also known as bias) and of *representativeness* (or representativity). The sum of these errors is sometimes known as the *accuracy*. Random errors have the property that they are reduced by averaging. Systematic errors, by contrast, are not reduced by averaging; if known, they can be subtracted from an observation. The representativeness error is associated with differences in the resolution of observational information and the resolution of the model interpreting this information.

Models also have errors. These errors arise through the construction of models, as models can be incomplete due to a lack of understanding or due to processes being omitted to make the problem tractable; and through their imperfect simulation of the "real world", itself sampled by observations or measurements. Thus, information, whether in the form of observations or models has errors, and these have to be taken into account. Further details about the nature of observational and model errors can be found in the following chapters in Part I, *Theory*.

Another key feature of observations (or measurements) is that they are discrete in space and time, with the result that the information provided by observations has *gaps* (Fig. 1).

It is desirable to fill gaps in the information provided by observations: first, to make this information more *complete*, and hence more useful; second, to provide information at a *regular scale* to quantify the characteristics of this information. Information at an *irregular scale* can be quantified, but this procedure is more tractable when done with a regular scale.

Assuming a need to fill in the gaps in the observational information, the question is how to do so. Conceptually, it is desirable to use information on the behaviour of the system to extend the observations and fill in the gaps. This information is provided by a model of how the system behaves; this model then allows one to organize, summarize and propagate the information from observations. Note that

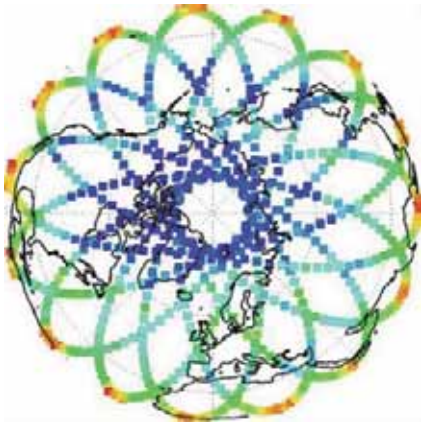


**Fig. 1** Plot representing ozone data at 10 hPa (approximately 30 km in altitude) for 1 February 1997 based on the observational geometry of ozone measurements from the MLS (Microwave Limb Sounder) instrument onboard the National Aeronautics and Space Administration (NASA) UARS (Upper Atmosphere Research Satellite) satellite. For information on UARS, see http://mls.jpl.nasa.gov/uars/science.php. *Blue* denotes relatively low ozone values; *red* denotes relatively high ozone values. Note the gaps between the satellite orbits. Thanks to Finn Bjørklid (NILU) for improving this figure

there can be differences in the resolution of the observations, and the resolution of the models used to propagate the information in observations. This will introduce errors when filling in the information gaps.

We now discuss algorithms to fill in the information gaps. The idea is that the algorithm, embedded in a model, provides a set of *consistent* (i.e., mathematically, physically or otherwise) and *objective* (i.e., based on impartial principles) rules which when followed fill in the information gaps associated with observations.

## 5 Objective Ways of Filling in Information Gaps

What algorithm should one use to fill in the information gaps associated with observations? There are a number of features that such an algorithm should have. The most important ones are that it be *feasible* and that it be *objective* (and *consistent*). From the point of view of feasibility, one could build a hierarchy of algorithms of increasing complexity, starting, for example, with linear interpolation between observations. A simple approach such as linear interpolation is feasible (because simple) and, in cases where observations are dense enough, could be expected to be reasonably accurate. However, although in principle consistent, it is not objective (because not general) and, for example, in general it will not reflect how it is understood systems such as the atmosphere behave. A more realistic approach would be to fill in the gap using a model of how the system behaved. For example, for the

atmosphere, we could use a model that embodies the *equations of motion*; *radiative transfer*; *physical processes* such as convection; and *chemistry*. Such a model would be more expensive to apply than a simple linear interpolation, but in principle would provide a more accurate (and more objective) approach to filling in the information gaps in the observations. In practice, one strikes a balance between using a model that is feasible and using a model that is objective and consistent. Practically, one seeks a model that is *tractable* and *realistic*.

We would like to find methods that allow the interpolation, i.e., filling in of the observational information gaps using a model, to be done in an "intelligent" way. By intelligent, we mean an "objective" way which makes use of concepts for combining information that can be quantified. For example, by finding the *minimum* or *maximum* value of a quantity that can be calculated from the information available. In this way, we can think of the model as an intelligent interpolator of the observation information: intelligent because it embodies our understanding of the system; intelligent because the combination of the observational and model information is done in an objective way. Note that in practice, the model (like the observations) provides information that is discrete in space and time.

*Mathematics* provides rules for combining information objectively, based on principles which aim to maximize (or minimize) a quantity (e.g. a "penalty function"), or on established *statistical concepts* which relate *prior information* (understanding, which comes from prior combination of observations and models) with *posterior information* (which comes from making an extra observation).

In particular, mathematics provides a foundation to address questions such as: "What combination of the observation and model information is optimal?", and provides an estimate of the errors of the "optimum" or "best" estimate. This is known as "data assimilation" (also as Earth Observation data/model fusion), and has strong links to several mathematical disciplines, including *control theory* and *Bayesian statistics*. The data assimilation methodology adds value to the observations by filling in the observational gaps and to the model by constraining it with observations (Fig. 2 below). In this way, the data assimilation allows one to "make sense" of the observations. Further details about the theory of data assimilation can be found in the following chapters in Part I, *Theory*.

Mathematics also provides an algorithmic basis for applying data assimilation to real problems, including, for example, *weather forecasting*, where data assimilation has been very successful. In particular, over the last 25 years, the skill of weather forecasts has increased – the skill of today's 5-day forecast is comparable to the skill of the 3-day forecast 25 years ago. Furthermore, the skill of forecasts for the Southern Hemisphere is now comparable to that of the Northern Hemisphere (Simmons and Hollingsworth 2002).

Mathematics also provides a *theoretical* and *algorithmic* basis for studying the problem of data assimilation, notably by using simpler models to test ideas. The results using these simpler models can then be used to inform data assimilation developments with complex systems, such as those used for weather forecasting.

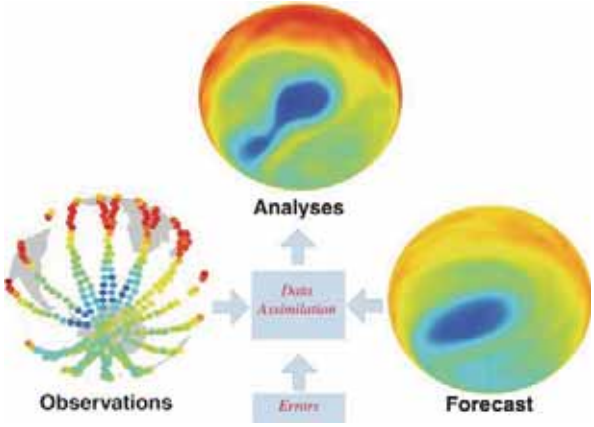



**Fig. 2** Schematic of how data assimilation (DA) works and adds value to observational and model information. The data shown are various representations of ozone data at 10 hPa (about 30 km in height) on 23 September 2002. *Lower left panel*, "observations": plot representing the day's ozone data based on the observational geometry of ozone measurements from the MIPAS (Michelson Interferometer for Passive Atmospheric Sounding) instrument onboard the European Space Agency (ESA) Envisat satellite; for information on MIPAS, see http://envisat.esa.int/instruments/mipas/. *Lower right panel*, "forecast": plot representing a 6-day ozone forecast (1200 UTC) based on output from a DA system. *Top panel*, "analyses": plot representing an ozone analysis (1200 UTC) based on output from a DA system. The DA system associated with the lower right plot and the top plot is based on that at the Met Office, and is described in Geer et al. (2006). *Blue* denotes relatively low ozone values; *red* denotes relatively high ozone values. The DA method combines the observations with a model forecast (commonly short-term, e.g., 6 or 12 h), including their errors to produce an ozone analysis. Note how the analysis (*top panel*) fill in the gaps in the observations (*lower left panel*), and the analysis captures the Antarctic ozone hole split (verified using independent data not used in the assimilation) whereas the 6-day forecast (*lower right panel*) does not. In this sense, the DA method adds value to both the observations and the model. Thanks to Alan Geer for providing the basis of this figure and for Finn Bjørklid for improving the figure

## 6 Simple Examples of Data Assimilation

We now provide three simple examples highlighting how data assimilation *adds value* (Example 1); the impact of *spatial resolution* on information (Example 2); and the impact of *temporal sampling* on information (Example 3).

*Example 1* Combining observations with understanding of a system, where both pieces of information have finite errors, should, intuitively, increase the information about the system. There are several ways of quantifying this increase in information, one of them being the error embodied in the information, quantified by the *standard deviation*. We discuss this using a simple example where information from two scalar quantities with *Gaussian* (i.e., normally distributed) errors is combined.

Consider two observations $(x_1, x_2)$ of variable $x$, with associated variances $({\sigma_1}^2, {\sigma_2}^2)$. Now assume that the observation errors are random, unbiased and normally distributed. It can be shown that the *optimum estimate* ("most probable" value) is given by:

$$x = \frac{\left(\frac{x_1}{\sigma_1^2} + \frac{x_2}{\sigma_2^2}\right)}{\left(\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}\right)},$$

with variance:

$$\sigma^2 = \left(\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}\right)^{-1}.$$

We can also see from this example that:

$$\sigma_1 \to \infty, x \to x_2;$$
$$\sigma^2 \leq \min\{\sigma_1^2, \sigma_2^2\}.$$

We can see from this simple example that the error (variance) associated with the combined information is generally lower than the error associated with any of the two pieces of information being combined and that, at worse, it is equal to the minimum of the errors of the individual pieces of information, but never larger. We can also see obvious limiting cases, when the error of one of the pieces of information being combined becomes infinitely large, i.e., the information from this piece becomes vanishingly small. The result in this example can be generalized to two observations $(\mathbf{x}_1, \mathbf{x}_2)$ of a vector variable $\mathbf{x}$, with associated matrix error covariances $(\mathbf{S}_1, \mathbf{S}_2)$.

Although this simple example encapsulates how information is increased, this result concerning variances only holds for Gaussian errors. For errors that are not Gaussian, the variance of the combined information can be larger than that of one of the pieces of information being combined. This apparently counter-intuitive result indicates that variance is not the best way of measuring increases in information. In fact, one must use the concept of *entropy* to consider errors with general probability distributions.

*Example 2* Consider a large square room, where temperature measurements are made at each corner. What is the temperature at the centre of the room? What is the temperature representative for the room? These questions concern the *spatial resolution* of information, and how the latter changes as the former changes.

To estimate the temperature at the centre of the room we could average the four corner temperatures, giving each measurement equal weight. This gives the same result assuming the temperature varies linearly between opposite corners and taking an average of the two resulting measurements. Regardless of how the final value is

computed, a *model* of how the temperature varies in the room is needed to compute the temperature at the centre of the room.

To estimate the temperature representative for the room we could proceed as above. In this case we would be averaging the "point" temperature information from each corner to provide "area" temperature information for the whole room. When we use this estimate of the "area" temperature (or any other estimate) as representative of the room temperature, we incur an error of *representativeness*. This was introduced in Sect. 4 above.

The impact of spatial resolution on the estimate for the temperature at the centre of the room can be seen as follows. If we increase the number of measurements in the room, for example along the walls or toward the centre, we tend to get a better estimate of the temperature at the centre of the room, either because we are sampling closer to the room centre, and/or we are obtaining more information of how the temperature varies in the room. Higher spatial observational sampling generally provides (at least initially) better information on the system by reducing the observational gaps. However, there comes a point where we do not get further information, e.g., sampling the temperature at close enough locations in the room gives essentially an unchanged temperature within the error of the measuring device. This illustrates the concept of observational information *saturation* with respect to other observations, where the measurement is no longer *independent* and provides no new information.

The impact of spatial resolution on the estimate for the "area" temperature of the room can be seen as follows. Assume the spatial resolution of the algorithm (i.e., model) used to estimate the "area" temperature remains fixed. As we *reduce* the spatial dimensions of the room the observational gaps become smaller, and the estimate of the "area" temperature as calculated above (or generally using any algorithm or model) initially tends to become *more accurate*. However, there comes a point where, within the error of the algorithm, we do not get further information if we continue reducing the spatial dimension of the observational gaps. We have observational information *saturation* with respect to the model.

Through representation of errors, data assimilation takes account of the spatial resolutions in the model and the observations, and the information saturation between observations, and between the observations and the model.

*Example 3* Consider a person walking along a path in the forest, gathering information about their surroundings through their eyes, and keeping their eyes closed for regular intervals. How does this person keep on the path when their eyes are closed? How does the time the person keeps their eyes closed affect their progress along the path? These questions concern the rate at which information is sampled in time, i.e., *temporal sampling*.

The person gathers *observational* information about their surroundings through their eyes: "the path is straight"; "the path curves to the left". This provides information of the path to the person, who then incorporates it into a model of their surroundings. This allows the person to keep along the path when their eyes are closed: "keep straight ahead"; "turn left". When the person next opens their eyes

they can adjust (*correct*) their model of their surroundings depending on the new observational information: “turn right”; “bend down to avoid a low tree branch”. The combination of observational and model information allows the person to walk along the path.

However, the amount of time the person keeps their eyes closed affects the quality of observational information they get about their surroundings. If the amount of time is relatively short, say 1 s, the quality of observational information will be relatively high and the person should be able to walk along the path without mishap. By contrast, if the amount of time is relatively long, say 1 min, the quality of observational information will be relatively low and the person would be expected to have problems walking along the path (note, however, that this depends on the nature of the path, see later). This shows how temporal sampling can affect the quality of observational information received, which in turn allows the correction of model information.

If the path is straight, the amount of time the person keeps their eyes closed can be relatively long and still allow them to be able to keep along the path without mishap. This is because the model of the path (built from observational information) is relatively simple: “keep on a straight line”, and does not need relatively high temporal sampling to adjust it. Conversely, if the path has many bends without pattern in their handedness, the model of the path (again, built from observational information) is relatively complex: “keep turning in the direction of the path”, and needs relatively high temporal sampling to adjust it. This shows how the complexity of the system affects the temporal sampling of observational information needed to adjust (i.e., keep “on track”) a model describing the system. The appropriate complexity of a model describing the system depends on the character of the observational information gathered (observation types, errors, spatial resolution, temporal sampling).

Data assimilation, by confronting the model with observations in time and space, keeps the model on track.

## 7 Benefits of Combining Information

As seen in Fig. 2 above, and the examples in Sect. 6, combining information from observations and a model adds value to both the observations and the model: the information gaps in the observations are filled in; the model is constrained by the observations. Other benefits accrue from “confronting” observations and models, as is done in the data assimilation method. These benefits include the *evaluation* of both the observations and the model. This evaluation of information is crucial in *Earth Observation* (observational information); *Earth System Modelling* (model information, i.e., information which embodies our understanding); and in *melding* observations with a model, which we call “data assimilation” (merging of information). By evaluating information, shortcomings can be identified and remedied, with a consequent improvement in the collection, propagation and use of information.

## 8 What This Book Is About

This book develops the theme introduced in this chapter, namely, the use of data assimilation to make sense of observations. It has six foci:

- *Theory* (the eight chapters in Part I following this chapter);
- *Observations* (the three chapters in Part II);
- *Meteorology and Atmospheric Dynamics* (the three chapters in Part III);
- *Atmospheric Chemistry* (the four chapters in Part IV);
- *Wider Applications* (the three chapters in Part V);
- *The Longer View* (the three chapters in Part VI).

These foci span several cross-cutting axes: (i) the mathematics of data assimilation; (ii) observations and models; (iii) the activities of the weather centres and the activities of the research community; (iv) the different elements of the Earth System: atmosphere, ocean, land and chemistry; (v) evaluation and production of added-value analyses; and (vi) the success of the data assimilation method and future developments. These are exciting times for data assimilation and we hope this book conveys this excitement.

## References

Geer, A.J., W.A. Lahoz, S. Bekki, et al., 2006. The ASSET intercomparison of ozone analyses: Method and first results. *Atmos. Chem. Phys.*, **6**, 5445–5474.

Simmons, A.J. and A. Hollingsworth, 2002. Some aspects of the improvement in skill of numerical weather prediction. *Q. J. R. Meteorol. Soc*., **128**, 647–677.

# Mathematical Concepts of Data Assimilation

**N.K. Nichols**

## 1 Introduction

Environmental systems can be realistically described by mathematical and numerical models of the system dynamics. These models can be used to predict the future behaviour of the system, provided that the initial states of the system are known. Complete data defining all of the states of a system at a specific time are, however, rarely available. Moreover, both the models and the available initial data contain inaccuracies and random noise that can lead to significant differences between the predicted states and the actual states of the system. In this case, observations of the system over time can be incorporated into the model equations to derive "improved" estimates of the states and also to provide information about the "uncertainty" in the estimates.

The problem of state-estimation is an inverse problem and can be treated using observers and/or filters derived by feedback design techniques (see, for example, Barnett and Cameron 1985). For the very large non-linear systems arising in the environmental sciences, however, many traditional state-estimation techniques are not practicable and new "data assimilation" schemes have been developed to generate accurate state-estimates (see, for example, Daley 1993; Bennett 1992). The aim of such schemes can be stated as follows.

> The aim of a data assimilation scheme is to use measured observations in combination with a dynamical system model in order to derive accurate estimates of the current and future states of the system, together with estimates of the uncertainty in the estimated states.

The most significant properties of the data assimilation problem are that the models are very large and non-linear, with order $O(10^7–10^8)$ state variables. The dynamics are multi-scale and often unstable and/or chaotic. The number of observations is also large, of order $O(10^5–10^6)$ for a period of 6 h, but the data are not evenly distributed in time or space and generally have "holes" where there are no

N.K. Nichols (✉)
Department of Mathematics, University of Reading, Whiteknights, Reading, RG6 6AX UK
e-mail: n.k.nichols@reading.ac.uk

W. Lahoz et al. (eds.), *Data Assimilation*, DOI 10.1007/978-3-540-74703-1_2,

observations (see chapter *Data Assimilation and Information*, Lahoz et al.). In practice the assimilation problem is generally ill-posed and the state estimates may be sensitive to errors.

There are two basic approaches to this problem. The first uses a "dynamic observer," which gives a *sequential data assimilation scheme*, and the second uses a "direct observer," which gives a *four-dimensional data assimilation scheme*. In the first case, the observations are "fed-back" into the model at each time these are available and a best estimate is produced and used to predict future states. In the second case a feasible state trajectory is found that best fits the observed data over a time window, and the estimated states at the end of the window are used to produce the next forecast. Under certain mathematical assumptions these processes solve the same "optimal" state-estimation problem. In operational systems, solving the "optimal" problem in "real-time" is not always possible, and many different approximations to the basic assimilation schemes are employed.

In the next section the data assimilation problem is formulated mathematically. In subsequent sections various techniques for solving the assimilation problem are discussed.

## 2 Data Assimilation for Non-linear Dynamical Systems

A variety of models is used to describe systems arising in environmental applications, as well as in other physical, biological and economic fields. These range from simple linear, deterministic, continuous ordinary differential equation models to sophisticated non-linear stochastic partial-differential continuous or discrete models. The data assimilation schemes, with minor modifications, can be applied to any general model.

We begin by assuming that for any given initial states and given inputs, the equations modelling the dynamical system uniquely determine the states of the system at all future times. This is known as the "perfect" model assumption. In the following subsections we define the data assimilation problem for this case and examine its properties. Next we determine a best linear estimate of the solution to the non-linear assimilation problem. The data assimilation scheme is then interpreted in a stochastic framework and the "optimal" state-estimate is derived using statistical arguments. We consider the case where the model includes errors in the system equations in a later section of this chapter.

### *2.1 Basic Least-Squares Formulation for Perfect Models*

Data assimilation schemes are described here for a system modelled by the discrete non-linear equations

$$\mathbf{x}_{k+1} = \mathcal{M}_{k,k+1}(\mathbf{x}_k), \quad k = 0, \ldots, N-1, \tag{1}$$

where $\mathbf{x}_k \in \mathbb{R}^n$ denotes the vector of $n$ model states at time $t_k$ and $\mathcal{M}_{k,k+1} : \mathbb{R}^n \to \mathbb{R}^n$ is a non-linear operator describing the evolution of the states from time $t_k$ to time $t_{k+1}$. The operator contains known inputs to the system including known external forcing functions that drive the system and known parameters describing the system.

Prior estimates, or "background estimates," $\mathbf{x}_0^b$, of the initial states $\mathbf{x}_0$ at time $t_0$ are assumed to be known, usually provided by a previous forecast.

The observations are assumed to be related to the system states by the equations

$$\mathbf{y}_k = \mathcal{H}_k(\mathbf{x}_k) + \boldsymbol{\varepsilon}_k^{\mathrm{o}}, \quad k = 0, \ldots, N, \tag{2}$$

where $\mathbf{y}_k \in \mathbb{R}^{p_k}$ is a vector of $p_k$ observations at time $t_k$ and $\mathcal{H}_k : \mathbb{R}^n \to \mathbb{R}^{p_k}$ is a non-linear operator that includes transformations and grid interpolations. The observational errors $\boldsymbol{\varepsilon}_k^{\mathrm{o}} \in \mathbb{R}^{p_k}$ consist of instrumentation errors and representativity (or representativeness) errors (see chapter *Data Assimilation and Information*, Lahoz et al.).

For the "optimal" analysis, we aim to find the best estimates $\mathbf{x}_k^a$ for the system states $\mathbf{x}_k$, $k = 0, \ldots, N$, to fit the observations $\mathbf{y}_k$, $k = 0, \ldots, N$, and the background state $\mathbf{x}_0^b$, subject to the model equations (1). We write the problem as a weighted non-linear least-squares problem constrained by the model equations.

**Problem 1** Minimize, with respect to $\mathbf{x}_0$, the objective function

$$\begin{aligned} J = &\frac{1}{2}\left(\mathbf{x}_0 - \mathbf{x}_0^b\right)^T \mathbf{B}_0^{-1}\left(\mathbf{x}_0 - \mathbf{x}_0^b\right) + \\ &+ \frac{1}{2}\sum_{k=0}^{N}(\mathcal{H}_k(\mathbf{x}_k) - \mathbf{y}_k)^T \mathbf{R}_k^{-1}(\mathcal{H}_k(\mathbf{x}_k) - \mathbf{y}_k), \end{aligned} \tag{3}$$

subject to $\mathbf{x}_k$, $k = 1, \ldots, N$, satisfying the system equations (1) with initial states $\mathbf{x}_0$.

The model is assumed here to be "perfect" and the system equations are treated as *strong constraints* on the minimization problem. The states $\mathbf{x}_k$ that satisfy the model equations (1) are uniquely determined by the initial states and therefore can be written explicitly in terms of $\mathbf{x}_0$. Substituting into the objective function (3) then allows the optimization problem to be expressed in terms of the initial states alone. The assimilation problem, Problem 1, thus becomes an unconstrained weighted least-squares problem where the initial states are the required control variables in the optimization.

The weighting matrices $\mathbf{B}_0 \in \mathbb{R}^{n \times n}$ and $\mathbf{R}_k \in \mathbb{R}^{p_k \times p_k}$, $k = 0, 1 \ldots, N$, are taken to be symmetric and positive definite and are chosen to give the problem a "smooth" solution. They represent, respectively, the uncertainty in the background states (prior estimates) and the observations. The objective function (3) can then be written in the compact form:

$$J(\mathbf{x}_0) = \frac{1}{2}\|\mathbf{f}(\mathbf{x}_0)\|_2^2 \equiv \frac{1}{2}\mathbf{f}(\mathbf{x}_0)^T \mathbf{f}(\mathbf{x}_0), \tag{4}$$

where

$$\mathbf{f}(\mathbf{x}_0) = \begin{pmatrix} \mathbf{B}_0^{-1/2}\left(\mathbf{x}_0 - \mathbf{x}_0^b\right) \\ \mathbf{R}_0^{-1/2}(\mathcal{H}_0(\mathbf{x}_0) - \mathbf{y}_0) \\ \vdots \\ \mathbf{R}_N^{-1/2}(\mathcal{H}_N(\mathbf{x}_N) - \mathbf{y}_N) \end{pmatrix}, \tag{5}$$

and $\mathbf{x}_k = \mathcal{M}_{0,k}(\mathbf{x}_0)$, $k = 1, \ldots, N$, satisfy the system equations (1) with initial states $\mathbf{x}_0$ at time $t_0$ (see Lawless et al. 2005). The matrices $\mathbf{B}_0^{-1/2}$ and $\mathbf{R}_k^{-1/2}$ denote the inverses of the symmetric square roots of $\mathbf{B}_0$ and $\mathbf{R}_k$, respectively.

In this approach the initial states are treated as parameters that must be selected to minimize the weighted mean square errors between the observations predicted by the model and the measured observations over the time window and between the initial and background states. The initial state is adjusted to different positions in order to achieve the best fit, using an efficient iterative minimization algorithm.

## *2.2 Properties of the Basic Least-Squares Formulation*

The solution $\mathbf{x}_0^a$ to the least-squares problem (4) is known as the *analysis*. The analysis may not be well-defined if $\mathbf{B}_0^{-1} = 0$, that is, if no background state is specified. In that case the number and locations of the observations may not be sufficient to determine all the degrees of freedom in the optimization problem; in other words, the system may not be "observable." If the weighting matrix $\mathbf{B}_0$ is non-singular, however, then, provided the operators $\mathcal{M}_{0,k}$ and $\mathcal{H}_k$ are continuously differentiable, the stationary points of the least-squares problem are well-defined. The weighted background term acts as a "regularization" term, ensuring the existence of a solution and also damping the sensitivity of the solution to the observational errors (Johnson et al. 2005a, b).

Under these conditions, the stationary points of the objective function (4) satisfy the gradient equation, given by

$$\nabla_{\mathbf{x}_0} J = \mathbf{J}^T \mathbf{f}(\mathbf{x}_0) = 0, \tag{6}$$

where $\mathbf{J}$ is the Jacobian of the vector function $\mathbf{f}$ defined in (5). The Jacobian can be written in the compact form

$$\mathbf{J} = \begin{pmatrix} \mathbf{B}_0^{-1/2} \\ \hat{\mathbf{R}}^{-1/2}\hat{\mathbf{H}} \end{pmatrix}, \; \hat{\mathbf{H}} = \begin{pmatrix} \mathbf{H}_0 \\ \mathbf{H}_1\mathbf{M}_{0,1} \\ \vdots \\ \mathbf{H}_N\mathbf{M}_{0,N} \end{pmatrix}, \tag{7}$$

where $\hat{\mathbf{R}} = \text{diag}\{\mathbf{R}_k\}$ is a block diagonal matrix containing the weighting matrices $\mathbf{R}_k$ on the diagonal. The matrices $\mathbf{M}_{0,k}$ and $\mathbf{H}_k$ denote the Jacobians of the model and observation operators $\mathcal{M}_{0,k}$ and $\mathcal{H}_k$, respectively; that is,

$$\mathbf{M}_{0,k} = \left.\frac{\partial \mathcal{M}_{0,k}}{\partial \mathbf{x}}\right|_{\mathbf{x}_0}, \quad \mathbf{H}_k = \left.\frac{\partial \mathcal{H}_k}{\partial \mathbf{x}}\right|_{\mathcal{M}_{0,k}(\mathbf{x}_0)}.$$

If $\mathbf{B}_0$ is non-singular, then the Jacobian $\mathbf{J}$, given by (7), is of full rank and the stationary points satisfying the gradient equation (6) are well-defined. Stationary points are not unique, however, and may not yield a minimum of the non-linear assimilation problem. If a stationary point is such that the Hessian $\nabla^2_{\mathbf{x}_0} J$, of the objective function (3) (or equivalently (4)) is positive-definite at that point, then the stationary point is a local minimum of the assimilation problem (see Gratton et al. 2007). It should be noted that multiple local minima of the assimilation problem may exist.

We remark that the sensitivity of the analysis to small perturbations in the data depends on the "conditioning" of the Hessian, $\nabla^2_{\mathbf{x}_0} J$, that is, on the sensitivity of the inverse of the Hessian to small perturbations. If small errors in the Hessian lead to large errors in its inverse, then the computed solution to the data assimilation problem may be very inaccurate. In designing data assimilation schemes, it is important, therefore, to ensure that the conditioning of the Hessian is as small as feasible, or to use "preconditioning" techniques to improve the conditioning.

## *2.3 Best Linear Least-Squares Estimate*

In general, explicit solutions to the non-linear data assimilation problem, Problem 1, cannot be found. A "best" *linear* estimate of the solution to the non-linear problem can, however, be derived explicitly. We assume that the departure of the estimated analysis $\mathbf{x}_0^a$ from the background $\mathbf{x}_0^b$ is a *linear* combination of the innovations $\mathbf{d}_k = \mathbf{y}_k - \mathcal{H}_k\left(\mathbf{x}_k^b\right)$, $k = 0, 1, \ldots, N$, and find the estimate for $\mathbf{x}_0^a$ that solves the least-squares data assimilation problem as accurately as possible.

To determine the estimate, we linearize the assimilation problem about the non-linear background trajectory $\mathbf{x}_k^b = \mathcal{M}_{0,k}\left(\mathbf{x}_0^b\right)$, $k = 1, \ldots, N$. We denote by the matrices $\mathbf{H}_k$ and $\mathbf{M}_{0,k}$ the linearizations of the observation and model operators $\mathcal{H}_k$ and $\mathcal{M}_{0,k}$, respectively, about the background trajectory; that is,

$$\mathbf{H}_k = \left.\frac{\partial \mathcal{H}_k}{\partial \mathbf{x}}\right|_{\mathbf{x}_k^b}, \quad \mathbf{M}_{0,k} = \left.\frac{\partial \mathcal{M}_{0,k}}{\partial \mathbf{x}}\right|_{\mathbf{x}_0^b}.$$

The linearized least-squares objective function is then given by

$$\tilde{J} = \frac{1}{2}\boldsymbol{\delta}\mathbf{x}_0^T \mathbf{B}_0^{-1} \boldsymbol{\delta}\mathbf{x}_0 + \frac{1}{2}\sum_{k=0}^{N} (\mathbf{H}_k \mathbf{M}_{0,k} \boldsymbol{\delta}\mathbf{x}_0 - \mathbf{d}_k)^T \mathbf{R}_k^{-1} (\mathbf{H}_k \mathbf{M}_{0,k} \boldsymbol{\delta}\mathbf{x}_0 - \mathbf{d}_k), \qquad (8)$$

where $\boldsymbol{\delta}\mathbf{x}_0 = \left(\mathbf{x}_0 - \mathbf{x}_0^b\right)$. Using the compact form of the Jacobian (7), the gradient equation of the linearized problem may be written

$$\begin{aligned}\nabla_{\mathbf{x}_0}\tilde{J} &= \mathbf{B}_0^{-1}\left(\mathbf{x}_0 - \mathbf{x}_0^b\right) + \\ &\quad + \sum_{k=0}^{N}(\mathbf{H}_k\mathbf{M}_{0,k})^T\mathbf{R}_k^{-1}\left(\mathbf{H}_k\mathbf{M}_{0,k}\left(\mathbf{x}_0 - \mathbf{x}_0^b\right) - \left(\mathbf{y}_k - \mathcal{H}_k\left(\mathbf{x}_k^b\right)\right)\right) \\ &= \left(\mathbf{B}_0^{-1} + \hat{\mathbf{H}}^T\hat{\mathbf{R}}^{-1}\hat{\mathbf{H}}\right)\left(\mathbf{x}_0 - \mathbf{x}_0^b\right) + \hat{\mathbf{H}}^T\hat{\mathbf{R}}^{-1}\hat{\mathbf{d}} \\ &= 0,\end{aligned} \tag{9}$$

where $\hat{\mathbf{d}} = \left(\mathbf{d}_0^T, \mathbf{d}_1^T, \ldots, \mathbf{d}_N^T\right)^T$ is the vector of innovations.

The optimal *linear* state-estimate for $\mathbf{x}_0^a$ is then the solution to the gradient equation (9) and is given by

$$\mathbf{x}_0^a = \mathbf{x}_0^b + \hat{\mathbf{K}}\hat{\mathbf{d}}, \tag{10}$$

where

$$\hat{\mathbf{K}} = \left(\mathbf{B}_0^{-1} + \hat{\mathbf{H}}^T\hat{\mathbf{R}}^{-1}\hat{\mathbf{H}}\right)^{-1}\hat{\mathbf{H}}^T\hat{\mathbf{R}}^{-1} \equiv \mathbf{B}_0\hat{\mathbf{H}}^T\left(\hat{\mathbf{H}}\mathbf{B}_0\hat{\mathbf{H}}^T + \hat{\mathbf{R}}\right)^{-1}. \tag{11}$$

The matrix $\hat{\mathbf{K}}$ is known as the *gain* matrix.

For systems where the model and observation operators are linear, the analysis (10) and (11) is an exact, unique, stationary point of the data assimilation problem, Problem 1. For non-linear systems multiple stationary points of the objective function (3) may exist and the analysis (10) and (11) is only a first order approximation to an optimal solution, due to the linearization of the non-linear model and observation operators.

The Hessian of the linearized objective function (8) at the analysis (10) and (11) is given by

$$\nabla^2_{\mathbf{x}_0}\tilde{J} = \left(\mathbf{B}_0^{-1} + \hat{\mathbf{H}}^T\hat{\mathbf{R}}^{-1}\hat{\mathbf{H}}\right). \tag{12}$$

If $\mathbf{B}_0$ is non-singular, then the matrix (12) is symmetric and positive-definite and (10) and (11) provides the "best" linear estimate of the minimum of the data assimilation problem, Problem 1, in a region of the state space near to the background.

## *2.4 Statistical Interpretation*

The data assimilation problem, as formulated in Problem 1, determines a least-squares fit of the model predictions to the observations, subject to constraints. An estimate of the "uncertainty" in this analysis would be valuable. If additional assumptions about the stochastic nature of the errors in the initial state estimates and

the observations are made, then the solution to the data assimilation problem can be interpreted in statistical terms and the uncertainty in the analysis can be derived.

To obtain a statistical formulation of the data assimilation problem, we assume that the errors $\left(\mathbf{x}_0–\mathbf{x}_0^b\right)$ between the true initial states $\mathbf{x}_0$ and the prior background estimates $\mathbf{x}_0^b$ are randomly distributed with mean zero and covariance matrix $\mathbf{B}_0 \in \mathbb{R}^{n\times n}$. The observational errors $\boldsymbol{\varepsilon}_k^{\mathrm{o}} \in \mathbb{R}^{p_k}$, $k = 0, \ldots, N$, defined in (2), are assumed to be unbiased, serially uncorrelated, randomly distributed vectors with zero means and covariance matrices $\mathbf{R}_k \in \mathbb{R}^{p_k\times p_k}$. The observational errors and the errors in the prior estimates are assumed to be uncorrelated.

Under these basic statistical assumptions, given the prior estimates $\mathbf{x}_0^b$, and the observations $\mathbf{y}_k$, $k = 0, \ldots, N$, the "best linear unbiased estimate," or BLUE, of the true state $\mathbf{x}_0$ at time $t_0$ equals the best least-squares estimate (10) and (11) for the analysis $\mathbf{x}_0^a$. The uncertainty in this estimate is described by the analysis error covariance, which is given by

$$\mathbf{A} = (\mathbf{I}_n - \hat{\mathbf{K}}\hat{\mathbf{H}})\mathbf{B}_0. \tag{13}$$

Over all linear combinations of the innovations of form (10), the BLUE minimizes the analysis error covariance and is thus the solution to the assimilation problem with *minimum variance*. The analysis given by (10) and (11) is therefore the "optimal" linear estimate in this sense.

In addition to the basic statistical assumptions, the errors in the prior estimates and in the observations are commonly assumed to have Gaussian probability distributions, which are fully defined by the means and covariances specified. In this case, the solution to the data assimilation problem, Problem 1, is equal to the *maximum a posteriori Bayesian estimate* of the system states at the initial time. From Bayes Theorem we have that the posterior probability of $(\mathbf{x}_0 - \mathbf{x}_0^b)$, given the departures from the observations $(\mathbf{y}_k - \mathcal{H}_k(\mathbf{x}_k))$, $k = 0, \ldots, N$, satisfies

$$\begin{aligned}\rho\Big(\mathbf{x}_0 - \mathbf{x}_0^b | \mathbf{y}_k - \mathcal{H}_k(\mathbf{x}_k),\ k = 0, \ldots, N\Big) = \\ = \alpha\rho\Big(\mathbf{x}_0 - \mathbf{x}_0^b\Big)\,\rho\Big(\mathbf{y}_k - \mathcal{H}_k(\mathbf{x}_k),\ k = 0, \ldots, N | \mathbf{x}_0 - \mathbf{x}_0^b\Big),\end{aligned} \tag{14}$$

where $\rho\left(\mathbf{x}_0 - \mathbf{x}_0^b\right)$ is the prior probability of $\left(\mathbf{x}_0 - \mathbf{x}_0^b\right)$ and $\rho(\mathbf{y}_k - \mathcal{H}_k(\mathbf{x}_k)$, $k = 0, \ldots, N | \mathbf{x}_0 - \mathbf{x}_0^b\big)$ is the conditional joint probability of $(\mathbf{y}_k - \mathcal{H}_k(\mathbf{x}_k))$, $k = 0, \ldots, N$, given $\left(\mathbf{x}_0 - \mathbf{x}_0^b\right)$. The scalar $\alpha$ is a normalizing constant that ensures that the value of the posterior probability is not greater than unity. The "optimal" analysis is then the initial state that maximizes the posterior probability.

From the assumption that the probability distributions are Gaussian, we have that

$$\rho\left(\mathbf{x}_0 - \mathbf{x}_0^b\right) \propto \exp\left[-\frac{1}{2}\left(\mathbf{x}_0 - \mathbf{x}_0^b\right)^T \mathbf{B}^{-1}\left(\mathbf{x}_0 - \mathbf{x}_0^b\right)\right]$$

and

$$\rho(\mathbf{y}_k - \mathcal{H}_k(\mathbf{x}_k)) \propto \exp\left[-\frac{1}{2}(\mathbf{y}_k - \mathcal{H}_k(\mathbf{x}_k))^T \mathbf{R}_k^{-1}(\mathbf{y}_k - \mathcal{H}_k(\mathbf{x}_k))\right],$$

for $k = 0, 1, \ldots, N$. Taking the log of the posterior probability and using the assumptions that the observational errors are uncorrelated in time and uncorrelated with the background errors, we find that

$$\begin{aligned} J(\mathbf{x}_0) &\equiv -\ln\left[\rho\left(\mathbf{x}_0 - \mathbf{x}_0^b | \mathbf{y}_k - \mathcal{H}_k(\mathbf{x}_k), k = 0, \ldots, N\right)\right] \\ &= -\ln\left[\rho\left(\mathbf{x}_0 - \mathbf{x}_0^b\right)\right] - \sum_{k=0}^{N} \ln\left[\rho\left(\mathbf{y}_k - \mathcal{H}_k(\mathbf{x}_k)\right)\right]. \end{aligned} \tag{15}$$

(See Lorenc 1986, 1988.) The solution $\mathbf{x}_0$ to the data assimilation problem, Problem 1, that minimizes $J(\mathbf{x}_0)$ is therefore equivalent to the maximum Bayesian a posteriori likelihood estimate.

If the model and observation operators are linear and the errors are normally distributed (i.e., Gaussian), then the *maximum a posteriori Bayesian estimate* and the *minimum variance estimate* are equivalent. The BLUE, given explicitly by (10) and (11), with zero mean and covariance (13), is thus the unique optimal in both senses.

In practice the error distributions may not be Gaussian and the assumptions underlying the estimates derived here may not hold. Ideally, we would like to be able to determine the full probability distributions for the true states of the system given the prior estimates and the observations. This is a major topic of research and new approaches based on sampling methods and particle filters are currently being developed.

Techniques used in practice to solve the data assimilation problem, Problem 1, include sequential assimilation schemes and variational assimilation schemes. These methods are described in the next two sections.

## 3 Sequential Data Assimilation Schemes

We describe sequential assimilation schemes for discrete models of the form (1), where the observations are related to the states by the Eq. (2). We make the *perfect model assumption* here. We assume that at some time $t_k$, *prior* background estimates $\mathbf{x}_k^b$ for the states are known. The differences between the observations of the true states and the observations predicted by the background states at this time, $\left(\mathbf{y}_k - \mathcal{H}\left(\mathbf{x}_k^b\right)\right)$, known as the innovations, are then used to make a correction to the background state vector in order to obtain improved estimates $\mathbf{x}_k^a$, known as the analysis states. The model is then evolved forward from the analysis states to the next time $t_{k+1}$ where observations are available. The evolved states of the system at the time $t_{k+1}$ become the background (or forecast) states and are denoted by $\mathbf{x}_{k+1}^b$. The background is then corrected to obtain an analysis at this time and the process is repeated.

Mathematically this procedure may be written

$$\mathbf{x}_k^a = \mathbf{x}_k^b + \mathbf{K}_k \left( \mathbf{y}_k - \mathcal{H}_k \left( \mathbf{x}_k^b \right) \right), \tag{16}$$

$$\mathbf{x}_{k+1}^b = \mathcal{M}_{k,k+1} \left( \mathbf{x}_k^a \right). \tag{17}$$

The matrix $\mathbf{K}_k \in \mathbb{R}^{n \times p}$, known as the "gain matrix," is chosen to ensure that the analysis states converge to the true states of the system over time. This is possible if the system is "observable." Conditions for this property to hold are known (see, for example, Barnett and Cameron 1985).

The system (16) and (17) forms a modified dynamical system for the analysis states that can be written

$$\mathbf{x}_{k+1}^a = \mathcal{M}_{k,k+1} \left( \mathbf{x}_k^a \right) - \mathbf{K}_{k+1} \mathcal{H}_{k+1} \left( \mathcal{M}_{k,k+1} \left( \mathbf{x}_k^a \right) \right) + \mathbf{K}_{k+1} \mathbf{y}_{k+1}. \tag{18}$$

This system is driven by the observations and has different properties from the original discrete system model (1). The evolution of the analysed states from time $t_k$ to time $t_{k+1}$ is described by a modified non-linear operator and the response of the system depends generally upon the spectrum of its Jacobian, given by the matrix $(\mathbf{M}_{k,k+1} + \mathbf{K}_{k+1} \mathbf{H}_{k+1} \mathbf{M}_{k,k+1})$, where $\mathbf{H}_k = \frac{\partial \mathcal{H}_k}{\partial \mathbf{x}} \Big|_{\mathbf{x}_k^a}$ and $\mathbf{M}_{k,k+1} = \frac{\partial \mathcal{M}_{k,k+1}}{\partial \mathbf{x}} \Big|_{\mathbf{x}_k^a}$. The choice of the gain matrices $\mathbf{K}_k$, $k = 0, 1, \ldots$, therefore determines the behaviour of the analysed states over time and this choice characterizes the data assimilation scheme.

### *3.1 Optimal Sequential Assimilation Scheme*

For the "optimal" sequential assimilation scheme, the analysis $\mathbf{x}_k^a$, given by (16), is taken to be the *best linear estimate* of the solution to the least-squares assimilation problem

$$\min_{\mathbf{x}} \left[ \frac{1}{2} \left( \mathbf{x} - \mathbf{x}_k^b \right)^T \mathbf{B}_k^{-1} \left( \mathbf{x} - \mathbf{x}_k^b \right) + \frac{1}{2} (\mathcal{H}_k(\mathbf{x}) - \mathbf{y}_k)^T \mathbf{R}_k^{-1} (\mathcal{H}_k(\mathbf{x}) - \mathbf{y}_k) \right] \tag{19}$$

at time $t_k$. The gain matrix $\mathbf{K}_k$ is then given by

$$\mathbf{K}_k = \mathbf{B}_k \mathbf{H}_k^T \left( \mathbf{H}_k \mathbf{B}_k \mathbf{H}_k^T + \mathbf{R}_k \right)^{-1}, \tag{20}$$

with $\mathbf{H}_k = \frac{\partial \mathcal{H}_k}{\partial \mathbf{x}} \Big|_{\mathbf{x}_k^b}$.

If we assume that the background errors are randomly distributed with mean zero and error covariance matrix

$$\mathbf{B}_k = \mathcal{E} \left( \left( \mathbf{x} - \mathbf{x}_k^b \right) \left( \mathbf{x} - \mathbf{x}_k^b \right)^T \right), \tag{21}$$

then the optimal analysis is equal to the BLUE, or best linear unbiased estimate, and minimizes the analysis error variance given, at the optimum, by

$$\mathbf{A}_k \equiv \mathcal{E}\left(\left(\mathbf{x}-\mathbf{x}_k^a\right)\left(\mathbf{x}-\mathbf{x}_k^a\right)^T\right) = (\mathbf{I}_n - \mathbf{K}_k\mathbf{H}_k)\mathbf{B}_k. \tag{22}$$

If the random background error vector has a Gaussian distribution, then the analysis is the maximum posterior Bayesian estimate. For linear systems, the solution (16) and (20) gives the exact optimal analysis, but for non-linear systems this solution gives only a first order approximation to the optimal due to the linearization $\mathbf{H}_k$ of the non-linear observation operator that is used.

In evolving the "optimal" BLUE analysis sequentially, two computational difficulties arise. The first is that the background covariance matrices $\mathbf{B}_k$ are required at each time step. These matrices can be propagated forward in time from the initial background error covariance matrix $\mathbf{B}_0$ using an extended Kalman filter (EKF) technique (Kalman 1961). It is assumed that, at time $t_0$, *prior* background estimates $\mathbf{x}_0^b$ for the states are known and the errors between the true initial states and the background estimates are randomly distributed with mean zero and error covariance $\mathbf{B}_0$. The steps of the extended Kalman filter assimilation scheme are then given as follows. For $k = 0, 1, \ldots$ find

$$\mathbf{x}_k^a = \mathbf{x}_k^b + \mathbf{K}_k\left(\mathbf{y}_k - \mathcal{H}_k\left(\mathbf{x}_k^b\right)\right), \tag{23}$$

$$\text{where} \quad \mathbf{K}_k = \mathbf{B}_k\mathbf{H}_k^T\left(\mathbf{H}_k\mathbf{B}_k\mathbf{H}_k^T + \mathbf{R}_k\right)^{-1}, \tag{24}$$

$$\mathbf{A}_k = (\mathbf{I} - \mathbf{K}_k\mathbf{H}_k)\mathbf{B}_k, \tag{25}$$

$$\mathbf{x}_{k+1}^b = \mathcal{M}_{k,k+1}\left(\mathbf{x}_k^a\right), \tag{26}$$

$$\mathbf{B}_{k+1} = \mathbf{M}_{k,k+1}\mathbf{A}_k\mathbf{M}_{k,k+1}^T. \tag{27}$$

For systems where the model and observation operators are linear, the analysis $\mathbf{x}_N^a$ produced by the Kalman filter at time $t_N$ is exactly equal to the solution $\mathbf{x}_N^a = \mathcal{M}_{0,N}\left(\mathbf{x}_0^a\right)$ to the least-squares data assimilation problem, Problem 1, at the end of the time window. Furthermore, the analysis states produced by the Kalman filter converge over time to the expected values of the true states. For non-linear systems, however, the EKF only gives approximations to the optimal solution and the EKF may even become unstable as a dynamical system. The EKF is also sensitive to computational round-off errors (Bierman 1977).

For large geophysical and environmental systems the extended Kalman filter is, in any case, impractical to implement due to the size of the covariance matrices that need to be propagated. For example, for global weather and ocean systems, the EKF requires the computation of matrices containing of the order of $10^{14}$ elements at

every time step, making it computationally much too expensive to use for real-time state estimation.

The second difficulty in implementing the optimal assimilation scheme (16) and (20) sequentially is that in order to compute the analysis $\mathbf{x}_k^a$ at each time step, we must find $\mathbf{B}_k\mathbf{H}_k^T\mathbf{w}_k^a$, where $\mathbf{w}_k^a$ solves the linear equations

$$\left(\mathbf{H}_k\mathbf{B}_k\mathbf{H}_k^T + \mathbf{R}_k\right)\mathbf{w}_k^a = \left(\mathbf{y}_k - \mathcal{H}_k\left(\mathbf{x}_k^b\right)\right). \tag{28}$$

This is a very large inverse problem with $O(10^5$–$10^6)$ variables to find. Moreover, the solution may be sensitive to small errors in the data if the matrix $\left(\mathbf{H}_k\mathbf{B}_k\mathbf{H}_k^T + \mathbf{R}_k\right)$ is ill-conditioned.

In practice most operational sequential assimilation schemes avoid these two difficulties by using approximations that can be implemented efficiently. A summary of these methods is given in the next subsection.

## *3.2 Practical Implementation*

A variety of sequential data assimilation schemes has been developed for practical implementation. These differ mainly in the detailed steps of the procedures. Sequential assimilation schemes used operationally include (Nichols 2003a):

– *Successive Correction.* In these schemes, the feedback gain $\mathbf{K}_k$ is not chosen optimally, but is designed to smooth observations into the states at all spatial grid points within some radius of influence of each observation (Bergthorsson and Döös 1955). An iterative process is used to determine the analysis. The Cressman scheme is an example (Cressman 1959). The iterations converge to a result that is consistent with observational error but may not be consistent with the dynamical system equations. Over time the analysis states may not converge to the expected values of the true states. These schemes are generally not effective in data sparse regions.
– *Optimal Interpolation or Statistical Interpolation.* These schemes approximate the optimal solution by replacing the background error covariance matrix $\mathbf{B}_k$ by a constant matrix $\tilde{\mathbf{B}}$, which has a "fixed" structure for all $k$. The gain matrix $\mathbf{K}_k$ in (16) is then taken to be

$$\mathbf{K}_k = \tilde{\mathbf{B}}\mathbf{H}_k^T\left(\mathbf{H}_k\tilde{\mathbf{B}}\mathbf{H}_k^T + \mathbf{R}_k\right)^{-1}. \tag{29}$$

(see Ghil and Malanotte-Rizzoli 1991). The matrix $\tilde{\mathbf{B}}$ is generally defined by an isotropic correlation function (dependent only on the distance between spatial grid points and observational points), with the correlation lengths adjusted empirically. To simplify the inversion step, the gain is further modified to have a block structure by using innovations only in small regions around grid points to obtain the analysis

states. The inversion problem then reduces to solving a number of much smaller systems of equations.

– *Analysis Correction.* In these schemes, approximations to the optimal analysis states are computed iteratively, as in the Successive Correction method. The procedure is designed, however, to ensure that the iterates converge to the approximate "optimal" analysis that is obtained by replacing the optimal gain matrix (20) by the gain matrix (29), as in the optimal interpolation scheme (Bratseth 1986; Lorenc et al. 1991). This scheme is effective across data sparse regions and the analysis produced remains consistent with the dynamical equations.
– *3D-Var.* These schemes apply iterative minimization methods directly to the variational problem (19) (Rabier et al. 1993). The covariance matrix $\mathbf{B}_k$ is replaced by the approximation $\tilde{\mathbf{B}}$, as defined for optimal interpolation. The solution converges to the analysis obtained by replacing the optimal gain (20) by (29) in (16). Minimization techniques used commonly are pre-conditioned conjugate gradient methods and quasi-Newton methods. The properties of the analysis are similar to those obtained by the Analysis Correction method, but the iteration procedure is more efficient.
– *3D-PSAS and 3D-Representer.* In these schemes iterative minimization methods are applied to the dual variational problem

$$\min_{\mathbf{w}} \left[ \frac{1}{2} \left( \mathbf{w}^T \mathbf{H}_k \tilde{\mathbf{B}} \mathbf{H}_k^T + \mathbf{R}_k \right) \mathbf{w} - \mathbf{w}^T \left( \mathcal{H}_k(\mathbf{x}) - \mathbf{y}_k \right) \right].$$

The iterates converge to the solution $\mathbf{w}_k^a$ of the system (28) with $\mathbf{B}_k$ replaced by $\tilde{\mathbf{B}}$. The resulting analysis states converge to $\mathbf{x}_k^a = \tilde{\mathbf{B}} \mathbf{H}_k^T \mathbf{w}_k^a$, which approximates the "optimal" solution to the variational problem (19), as in the 3D-Var scheme (Cohn et al. 1998; Daley and Barker 2001). The advantage is that this scheme operates in the "observation space," which is, in general, of lower dimension than the state space. Additional work is needed, however, in order to reconstruct the analysis states.

In summary, most operational sequential data assimilation schemes aim to approximate the optimal analysis by replacing the background error covariance matrix by an approximation that is fixed over time and by simplifying the inversion problem and/or solving the inversion iteratively. Examples illustrating the application of these schemes to simplified models can be found in Martin et al. (1999) and on the website of the Data Assimilation Research Centre at http://darc.nerc.ac.uk/.

## *3.3 Ensemble Filters and Sampling Methods*

Newer approaches to sequential data assimilation known as *ensemble filter* methods, based on classical Kalman or square-root filtering, have recently received much attention. These methods use reduced rank estimation techniques to approximate the

classical filters and make the implementation feasible in real time. With these methods an ensemble consisting of a small number of analysis vectors (much less than the number of states $n$) is propagated simultaneously by the non-linear model from one observation time to the next in order to provide an ensemble of background states. The background ensemble is updated with the observations to give a new ensemble of analysis vectors and the "optimal" analysis state and its error covariance matrix are determined using a filter similar to the classical filters. An advantage of these methods is that the model and observation operators are not approximated linearly. The accuracy of the estimated states depends, however, on the spread of the ensemble, which must be sufficient to capture the true behaviour of the system.

There are many variants of this technique under development; see, for example, Anderson (2001); Bishop et al. (2001); Burgers et al. (1998); Evensen (2003); Houtekamer and Mitchell (1998); Nerger et al. (2005); Ott et al. (2004); Tippett et al. (2003); Zupanski (2005). Although the implementations may suffer from some difficulties (Livings et al. 2008), these methods retain the advantages of the classical Kalman and square-root filters while remaining feasible for application to large systems. Details of these techniques are described in a later chapter (chapter *Ensemble Kalman Filter: Cument Status and Potential*, Kalnay).

Sampling and particle filter methods aim to determine the full probability distributions for the true states of the system. These methods allow for non-Gaussian behaviour of the errors in the prior estimates and the observations and are closely related to the ensemble methods; see for example, Anderson and Anderson (1999); Pham (2001); Kim et al. (2003); van Leeuwen (2003); Apte et al. (2007). Although these methods are not yet efficient for very large geophysical problems, these approaches are promising and provide new directions for research.

## 4 Four-Dimensional Variational Assimilation Schemes

The least-squares data assimilation problem, Problem 1, is currently treated in many operational centres using four-dimensional variational schemes (4D-Var) (Sasaki 1970; Talagrand 1981; Rabier et al. 2000; Chapter *Variational Assimilation*, Talagrand). In these schemes the constrained minimization problem, Problem 1, is solved iteratively by a gradient optimization method where the gradients are determined using an adjoint method.

### *4.1 4D-Var and the Adjoint Method*

To solve the least-squares assimilation problem iteratively, the constrained problem, Problem 1, is first written as an unconstrained problem using the method of Lagrange. Necessary conditions for the solution to the unconstrained problem then require that a set of adjoint equations together with the system equations (1) must be satisfied. The adjoint equations are given by

$$\boldsymbol{\lambda}_{N+1} = 0, \tag{30}$$

$$\boldsymbol{\lambda}_k = \mathbf{M}_{k,k+1}^T \boldsymbol{\lambda}_{k+1} + \mathbf{H}_k^T \mathbf{R}_k^{-1} (\mathbf{y}_k - \mathcal{H}_k(\mathbf{x}_k)),\ k = N, \ldots, 0, \tag{31}$$

where $\boldsymbol{\lambda}_k \in \mathbb{R}^n$, $k = 0, \ldots, N$, are the adjoint variables and $\mathbf{M}_{k,k+1} \in \mathbb{R}^{n \times n}$ and $\mathbf{H}_k \in \mathbb{R}^{n \times p_k}$ are the Jacobians of $\mathcal{M}_{k,k+1}$ and $\mathcal{H}_k$ with respect to $\mathbf{x}_k$. The adjoint variables $\boldsymbol{\lambda}_k$ measure the sensitivity of the objective function (3) to changes in the solutions $\mathbf{x}_k$ of the state equations for each value of $k$.

The gradient of the objective function (3) with respect to the initial data $\mathbf{x}_0$ is then given by

$$\nabla_{\mathbf{x}_0} J = \mathbf{B}_0^{-1} \left( \mathbf{x}_0 - \mathbf{x}_0^b \right) - \boldsymbol{\lambda}_0. \tag{32}$$

At the optimum, the gradient (32) is required to be equal to zero. Otherwise this gradient provides the local descent direction needed in the iteration procedure to find an improved estimate for the optimal initial states. Each step of the gradient iteration process requires one forward solution of the model equations, starting from the current best estimate of the initial states, and one backward solution of the adjoint equations. The estimated initial conditions are then updated using the computed gradient direction. This process is expensive, but it is operationally feasible, even for very large systems.

A dual approach, used in 4D-PSAS and 4D-Representer methods, in which the minimization is performed in observation space, is also possible (Courtier 1997; Xu et al. 2005; Rosmond and Xu 2006). In these schemes, as in the three dimensional 3D-PSAS and 3D-Representer methods, a dual four-dimensional variational problem is solved using a gradient iteration method, and the analysis states are then reconstructed from the dual variables.

The primary difficulty in implementing variational assimilation schemes is the need to develop an adjoint model for the system. The adjoint equations are related theoretically to the linearized state equations, and the system matrix of the adjoint model is given directly by $\mathbf{M}_{k,k+1}^T$, where $\mathbf{M}_{k,k+1}$ is the system matrix of the linearized model. The adjoint equations can thus be generated directly from the linearized system equations. Automatic differentiation techniques can be applied to the forward solution code to generate the adjoint code (Griewank and Corliss 1991; Giering and Kaminski 1998). Alternatively an approximate adjoint system can be obtained by discretizing a continuous linear or adjoint model of the non-linear dynamics (Lawless et al. 2003). This approach has the advantage that additional approximations can be incorporated into the linearization of the system equations.

Other issues arising in the use of variational schemes are the need to cycle the scheme from one analysis time to the next and the length of the window to be used in each cycle. For each new cycle, the initial background weighting, or covariance, matrix $\mathbf{B}_0$ should depend on the current best estimate of the state, which is taken to be the optimal solution of the variational problem at the end of the previous

assimilation window. The Hessian of the objective function at the end of the previous cycle can provide this information, but this information is expensive to extract. In practice a climatological or seasonal average is used for the weighting matrix to start each cycle. New research is now becoming available on flow dependent covariance matrices and on longer assimilation windows, in which the initial weighting matrix is expected to have less influence on the analysis (see ECMWF 2007).

## *4.2 Incremental Variational Methods*

To make the variational methods more efficient, an "incremental" approach is generally used in which the non-linear assimilation problem is replaced by a sequence of approximate linear least-squares problems (Courtier et al. 1994).

At each step $i$ of this method, a linear variational problem is solved to find an increment $\delta\mathbf{x}_0^{(i)}$ to the current best estimate of the analysis $\mathbf{x}_0^{(i)}$. From the analysis $\mathbf{x}_0^{(i)}$ we solve the non-linear model equations (1) in order to determine the analysis states $\mathbf{x}_k^{(i)} = \mathcal{M}_{0,k}\left(\mathbf{x}_0^{(i)}\right)$ and the corresponding innovations $\mathbf{d}_k^{(i)} = \mathbf{y}_k - \mathcal{H}_k\left(\mathbf{x}_k^{(i)}\right)$ at time $t_k$. We then linearize the non-linear assimilation problem about the analysis state trajectory. Initially we set $\mathbf{x}_0^{(0)} = \mathbf{x}_0^b$, for $i = 0$. The linearized variational problem becomes

$$\begin{aligned}\min_{\delta\mathbf{x}_0^{(i)}} \frac{1}{2}\left(\delta\mathbf{x}_0^{(i)} - \left[\mathbf{x}_0^b - \mathbf{x}_0^{(i)}\right]\right)^T \mathbf{B}_0^{-1}\left(\delta\mathbf{x}_0^{(i)} - \left[\mathbf{x}_0^b - \mathbf{x}_0^{(i)}\right]\right) \\ + \frac{1}{2}\sum_{k=0}^{N}\left(\mathbf{H}_k\delta\mathbf{x}_k^{(i)} - \mathbf{d}_k^{(i)}\right)^T \mathbf{R}_k^{-1}\left(\mathbf{H}_k\delta\mathbf{x}_k^{(i)} - \mathbf{d}_k^{(i)}\right),\end{aligned} \tag{33}$$

subject to the tangent linear model (TLM) equations

$$\delta\mathbf{x}_{k+1}^{(i)} = \mathbf{M}_{k,k+1}\delta\mathbf{x}_k^{(i)}, \tag{34}$$

where $\mathbf{M}_{k,k+1} \in \mathbb{R}^{n\times n}$ and $\mathbf{H}_k \in \mathbb{R}^{n\times p_k}$ are linearizations of the operators $\mathcal{M}_{k,k+1}$ and $\mathcal{H}_k$ about the states $\mathbf{x}_k^{(i)}$. A new estimate for the analysis $\mathbf{x}_0^{(i+1)} = \mathbf{x}_0^{(i)} + \delta\mathbf{x}_0^{(i)}$ is obtained by updating the current estimate of the analysis with the solution to the linear variational problem (33) and the process is then repeated.

The linearized problem (33) is solved by an "inner" iteration process. Each inner iteration requires one forward solution of the tangent linear model equations (34), and one backward solution of the corresponding linear adjoint equations to determine the gradient of the objective function. The full incremental variational procedure thus consists of an inner and outer iteration process. In practice, the inner linear least-squares problem is solved only approximately, using a relatively small number of inner iterations, and only a few outer loops of the process are carried out, due to computational time constraints.

The incremental approach is also used in the implementation of the 4D-Representer method (Xu et al. 2005). The dual of the inner linear minimization problem is solved in observation space. The increments in physical space are then reconstructed from the dual variables at the end of the inner iteration and the outer loop is repeated.

Recently the incremental procedure has been shown to be equivalent to an approximate Gauss-Newton method and conditions for its convergence have been established (Lawless et al. 2005; Gratton et al. 2007). Approximations to the tangent linear model and to the corresponding adjoint may be used in the inner iteration without loss of convergence. Furthermore, the inner linear minimization problem does not need to be solved to full accuracy in each outer loop, thus avoiding unnecessary computation. Appropriate stopping criteria for the inner iteration process are presented in Lawless and Nichols (2006).

Additional techniques for increasing the efficiency of the four-dimensional variational methods are discussed in the next subsections.

## *4.3 Control Variable Transforms*

In the incremental variational assimilation scheme, transformations of the "control variables" may be applied in order to "decouple" the state variables, to simplify the computational work and to improve the conditioning of the minimization problem. The assimilation problem is written in terms of new variables $\boldsymbol{\chi}_0$, where

$$\left(\mathbf{x}_0 - \mathbf{x}_0^b\right) = \mathbf{U}\boldsymbol{\chi}_0. \tag{35}$$

The transformed linear variational problem (33) becomes

$$\min_{\boldsymbol{\chi}_0}\left[\frac{1}{2}\|\mathbf{B}_0^{-1/2}\mathbf{U}\boldsymbol{\chi}_0\|_2^2 + \frac{1}{2}\|\hat{\mathbf{R}}^{-1/2}\hat{\mathbf{H}}\mathbf{U}\boldsymbol{\chi}_0 - \hat{\mathbf{R}}^{-1/2}\hat{\mathbf{d}}\|_2^2\right], \tag{36}$$

where $\hat{\mathbf{H}}$, $\hat{\mathbf{R}}$ are defined as in (7) and $\hat{\mathbf{d}}$ is the vector comprised of the innovations. The conditioning of the optimization problem then depends on the Hessian of the objective function. Transforming the control variables alters the Hessian and changes the convergence properties of the inner iteration of the incremental method. The transformation thus acts as a preconditioner on the inner linearized least-squares problem. The transformation does not, however, affect the convergence of the outer loop of the incremental process.

If we choose $\mathbf{U} = \mathbf{B}_0^{1/2}$, where $\mathbf{B}_0^{1/2}$ is the symmetric square root of $\mathbf{B}_0$, the transformed problem (36) takes the form of a classical Tikhonov regularized inverse problem. The Hessian is then given by

$$\mathbf{I} + \mathbf{B}_0^{1/2}\hat{\mathbf{H}}\hat{\mathbf{R}}^{-1}\hat{\mathbf{H}}\mathbf{B}_0^{1/2}, \tag{37}$$

which is essentially a low-rank update of the identity matrix. The matrix $\hat{\mathbf{R}}^{-1/2}\hat{\mathbf{H}}\mathbf{B}_0^{1/2}$ is the *observability* matrix of the system and is key to the assimilation of information from the observations (Johnson et al. 2005a, b). In the transformed optimization problem (36), the state variables in the background (or regularization) term are weighted by the identity matrix and thus are decoupled. From a statistical point of view, this means that the transformed variables are uncorrelated, identically distributed random variables. From a practical point of view, the computational work needed in the inversion of the Hessian is simplified and the inner iteration may be implemented more efficiently. Additional preconditioners may also be applied to the gradient minimization algorithm in the incremental method to give further increases in the rates of convergence.

Operationally, control variable transforms may be used implicitly to define the background weighting, or covariance, matrix $\mathbf{B}_0$ in the least-squares formulation of the assimilation problem. A set of control variables is selected that are assumed from physical arguments to be uncorrelated. An appropriate transformation $\mathbf{U}$ from these variables to the original variables $\left(\mathbf{x}_0 - \mathbf{x}_0^b\right)$ is then defined and the matrix $\mathbf{B}_0$ is implicitly constructed from this transformation together with information about the spatial autocorrelations of each control variable. By this method additional constraints can be built into the transformations to ensure balance relations hold between the variables, and spectral and other transformations can also be applied implicitly. Flow dependence is also introduced into the weighting matrices by this technique. The validity of this approach depends, however, on the assumption that the transformed control variables $\boldsymbol{\chi}_0$ are truly decoupled, or uncorrelated (see, for example, Bannister et al. 2008; Katz 2007; Wlasak et al. 2006; Cullen 2003; Weaver and Courtier 2001). Good choices for the control variables and appropriate preconditioners for the gradient minimization algorithms continue to be major topics of research and development.

## *4.4 Model Reduction*

To increase the efficiency of the incremental methods further, the inner linear minimization problem is often approximated using low dimensional models. The simplest approach is to obtain the increments using a low-resolution linearized model for the dynamical system on a coarse grid. A prolongation operator is then used to map the low-resolution increments to the high-resolution model. Different resolutions can be used at each outer iteration of the procedure, leading to a multi-level approach (Trémolet 2005; Radnoti et al. 2005). These methods are now applied in practice, but theory to support their use is needed.

An alternative technique uses projection operators determined by methods from control theory to produce "optimal" reduced order models that most accurately capture the response of the full dynamic system. This approach allows much smaller system models to be used for the same computational accuracy, but currently these are expensive to derive (Lawless et al. 2008). More efficient approaches using subspace iteration methods and rational interpolation techniques are currently under

development. The latter approaches are promising as they allow for the practical reduction of unstable systems (Boess 2008; Bunse-Gerstner et al. 2007). Efficient new approximation methods based on proper orthogonal decomposition (POD) have also been developed recently for constructing the optimal projection operators (Willcox and Peraire 2002).

Other new approaches aim to solve the full non-linear variational problem in a low dimensional subspace spanned by basis functions generated using POD schemes from control theory or other similar methods (see Cao et al. 2007, and references therein). The accuracy and efficiency of these methods depends on how well the dynamics of the system can be captured in the low dimensional space. Similar techniques, which are adjoint free, have been developed for parameter estimation and model calibration (Vermeulen and Heemink 2006). Research in this area is currently active.

In summary, four-dimensional variational data assimilation schemes are in operational use at major numerical weather forecasting centres and new theory and new implementation techniques for these schemes continue to be major areas for research. Examples illustrating the use of these schemes on simplified models can be found in Griffith (1997) and Lawless et al. (2005). Tutorial examples are also available on the website of the Data Assimilation Research Centre at http://darc.nerc.ac.uk/.

## 5 Data Assimilation for Dynamical Systems with Model Errors

In the previous sections of this chapter, we have made the "perfect" model assumption that the initial states of the model equations uniquely determine the future states of the system. In practice, however, the non-linear dynamical model equations describing geophysical and environmental systems do not represent the system behaviour exactly and model errors arise due to lack of resolution (representativity errors) and inaccuracies in physical parameters, boundary conditions and forcing terms. Errors also occur due to discrete approximations and random disturbances. Model errors can be taken into account by treating the model equations as weak constraints in the assimilation problem.

A general least-squares formulation of the data assimilation problem for systems with model errors is introduced in this section. A statistical interpretation of the problem is presented and techniques for solving the assimilation problem for models with random forcing errors are discussed. In reality, model errors are comprised of both systematic and random components. A framework for treating both types of model error using the technique of state augmentation is developed (Nichols 2003b) and applications are reviewed.

### *5.1 Least-Squares Formulation for Models with Errors*

We assume that the evolution of the dynamical system, taking into account model errors, is described by the discrete non-linear equations

$$\mathbf{x}_{k+1} = \mathcal{M}_{k,k+1}(\mathbf{x}_k) + \boldsymbol{\eta}_k, \quad k = 0, \ldots, N-1, \tag{38}$$

where $\boldsymbol{\eta}_k \in \mathbb{R}^n$ denotes model errors at time $t_k$. Prior estimates, or "background" estimates, $\mathbf{x}_0^b$, of the initial states $\mathbf{x}_0$ are assumed to be known and the observations are assumed to be related to the system states by the Eq. (2).

For the "optimal" analysis, we aim to find the best estimates $\mathbf{x}_k^a$ of the true states of the system, $\mathbf{x}_k$, given observations $\mathbf{y}_k$, $k = 0, \ldots, N$, subject to the model equations (38) and prior estimates $\mathbf{x}_0^b$. The "optimal" assimilation problem is written as a weighted non-linear least-squares problem where the square errors in the model equations, together with the square errors between the model predictions and the observed system states and between the background and initial states are minimized. The data assimilation problem is defined mathematically as follows.

**Problem 2** Minimize, with respect to $\mathbf{x}_0$ and $\boldsymbol{\eta}_k$, $k = 0, \ldots, N-1$, the objective function

$$\begin{aligned} J = &\frac{1}{2}\left(\mathbf{x}_0 - \mathbf{x}_0^b\right)^T \mathbf{B}_0^{-1}\left(\mathbf{x}_0 - \mathbf{x}_0^b\right) + \\ &+\frac{1}{2}\sum_{k=0}^{N}(\mathcal{H}_k(\mathbf{x}_k) - \mathbf{y}_k)^T\mathbf{R}_k^{-1}(\mathcal{H}_k(\mathbf{x}_k) - \mathbf{y}_k) + \frac{1}{2}\sum_{k=0}^{N-1}\boldsymbol{\eta}_k^T\mathbf{Q}_k^{-1}\boldsymbol{\eta}_k, \end{aligned} \tag{39}$$

subject to $\mathbf{x}_k$, $k = 0, \ldots, N$, satisfying the system equations (38).

The model equations (38) are treated here as *weak constraints* on the objective function. The initial states of the system and the model errors at every time step are the control parameters that must be determined. The weighting matrices $\mathbf{B}_0 \in \mathbb{R}^{n\times n}$ and $\mathbf{R}_k \in \mathbb{R}^{p_k\times p_k}$, $\mathbf{Q}_k \in \mathbb{R}^{n\times n}$, $k = 0, 1 \ldots, N$, are taken to be symmetric and positive definite and are chosen to give the problem a "smooth" solution. The choices of the weights should reflect the relative confidence in the accuracy of the background, the observations and the model dynamics and also the structure of the model errors over the time window of the assimilation.

If we assume that the errors in the prior estimates, in the observations and in the model equations are random variables, then the "optimal" solution to the weakly constrained data assimilation problem, Problem 2, can be interpreted in a statistical sense. We assume that the probability distribution of the random errors $\left(\mathbf{x}_0 - \mathbf{x}_0^b\right)$ between the true initial states and the prior background estimates is Gaussian with mean zero and covariance matrix $\mathbf{B}_0 \in \mathbb{R}^{n\times n}$. The observational errors $\boldsymbol{\varepsilon}_k^{\mathrm{o}} \in \mathbb{R}^{p_k}$, defined in (2), are assumed to be unbiased, serially uncorrelated, Gaussian random vectors with covariance matrices $\mathbf{R}_k \in \mathbb{R}^{p_k\times p_k}$. The model errors $\boldsymbol{\eta}_k$, defined in (38), are also assumed to be randomly distributed variables that are unbiased and serially uncorrelated, with zero means and covariance matrices given by $\mathbf{Q}_k \in \mathbb{R}^{n\times n}$. The model errors, the observational errors and the errors in the prior estimates are assumed to be uncorrelated. Under these statistical assumptions, the optimal analysis $\mathbf{x}_0$ that solves the data assimilation problem, Problem 2, is equal to the *maximum* a posteriori *Bayesian estimate* of the system states, given the observations and the prior estimates of the initial states.

## 5.2 Optimal Solution of the Assimilation Problem

In order to find the "optimal" analysis that solves Problem 2, either sequential schemes that use the extended Kalman filter (EKF) or variational schemes that solve the minimization problem iteratively can be applied. The EKF propagates the analysis and the covariance matrices forward together, taking into account the model error statistics, in order to produce the "optimal" linear unbiased state estimate at each time step, conditional on the current and all previous observations. For linear models, the solution obtained using the EKF is the exact optimal and is equal to the solution to the assimilation problem at the end of the time period. For non-linear systems, approximate linearizations of the model and observation operators are introduced in the extended filter, and the optimality property is not retained.

Variational techniques, in contrast, solve the optimal assimilation problem, Problem 2, for all the analysis states in the assimilation window simultaneously. A direct gradient iterative minimization procedure is applied to the objective function (39), where the descent directions are determined from the associated adjoint equations. The full set of adjoint equations provides gradients of the objective function with respect to the initial states and with respect to all of the model errors at each time step. A forward solve of the model equations, followed by a reverse solve of the adjoint equations is needed to determine the gradients. Alternatively, the optimal assimilation problem can be solved by a dual variational approach in which the minimization is performed in observation space.

For very large stochastic systems, such as weather and ocean systems, these techniques for treating model errors are not practicable for "real-time" assimilation due to computational constraints. The four-dimensional variational and extended Kalman filter data assimilation schemes are both generally too expensive for operational use due to the enormous cost of estimating all of the model errors in the variational approach or, alternatively, propagating the error covariance matrices in the Kalman filter.

Promising practical approaches to solving the assimilation problem for models with stochastic forcing errors include the sequential ensemble filter methods and the dual variational methods. The ensemble methods take the model errors into account in the low order equations for propagating the ensemble statistics. The dual variational methods solve the assimilation problem in observational space and estimate the model errors implicitly during the reconstruction of the states from the dual variables. Reduced order approaches to solving the variational problem in physical space also allow model errors to be taken into account.

In practice, model errors do not, however, satisfy the statistical assumptions made here. The model error is expected to depend on the model state and hence to be *systematic* and *correlated in time*. A more general form of the model error that includes both systematic and random elements is described in the next subsection.

## 5.3 Systematic Model Error and State Augmentation

The problem of accounting for systematic model errors in a cost-effective way has recently received more attention. Techniques for treating bias errors in the forecast using sequential and four-dimensional variational assimilation schemes (Dee and da Silva 1998; Derber 1989; chapter *Bias Estimation*, Ménard) and for treating time-correlated stochastic errors (Zupanski 1997) have been investigated. A general formulation for the treatment of systematic model errors has also been derived (Griffith and Nichols 1996).

We present here a framework for treating systematic, time-correlated model errors based on the formulation of Griffith and Nichols (1996, 2000). Simple assumptions about the evolution of the errors are made, enabling the systematic error to be estimated as part of the assimilation process. The model equations are augmented by the evolution equations for the error and standard data assimilation techniques can then be applied to the augmented state system.

To take into account the systematic components of the model errors, we assume that the evolution of the errors in the model equations (38) is described by the equations

$$\boldsymbol{\eta}_k = T_k(\mathbf{e}_k) + \mathbf{q}_k, \tag{40}$$

$$\mathbf{e}_{k+1} = \mathcal{G}_{k,k+1}(\mathbf{x}_k, \mathbf{e}_k), \tag{41}$$

where the vectors $\mathbf{e}_k \in \mathbb{R}^r$ represent time-varying systematic components of the model errors and $\mathbf{q}_k \in \mathbb{R}^n$ are random errors. The random errors are commonly assumed to be unbiased, serially uncorrelated, and normally distributed with known covariances. The effect of the systematic errors on the model equations is defined by the operators $T_k : \mathbb{R}^r \to \mathbb{R}^n$. The operators $\mathcal{G}_{k,k+1} : \mathbb{R}^n \times \mathbb{R}^r \to \mathbb{R}^r$, describing the systematic error dynamics, are to be specified. The evolution of the errors may depend on the current states of the system.

In practice little is known about the form of the model errors and a simple form for the error evolution that reflects any available knowledge needs to be prescribed. The most common assumption is that the errors constitute a constant bias in each of the model equations. In this case the evolution of the errors is given by $\mathbf{e}_{k+1} = \mathbf{e}_k$, with $T_k = I$. Other forms include linearly evolving error and spectral forms varying on a given time-scale (see Griffith 1997; Griffith and Nichols 2000). These forms are expected to be appropriate, respectively, for representing average errors in source terms or in boundary conditions, for representing discretization error in models that approximate continuous dynamical processes by discrete-time systems, and for approximating the first order terms in a Fourier or spherical harmonic expansion of the model error.

Together the system equations and the model error equations (38), (40) and (41) constitute an *augmented* state system model. The aim of the data assimilation problem for the augmented system is to estimate the values of the augmented states $(\mathbf{x}_k^T, \mathbf{e}_k^T)^T$, for $k = 0, \ldots, N-1$, that best fit the observations, subject to the

augmented state equations. Assuming that the errors in the initial states, the observations and the random components of the model errors, are unbiased, normally distributed, serially uncorrelated and uncorrelated with each other, then the solution delivers the maximum a posteriori estimate of the augmented system states. Although this formulation takes into account the evolution of the systematic model errors, the data assimilation problem remains intractable for operational use. If, however, the augmented system is treated as a "perfect" deterministic model, then solving the augmented data assimilation problem becomes feasible. The aim of the data assimilation, in this case, is to estimate the systematic components of the model error simultaneously with the model states.

The "perfect" augmented system equations are written

$$\mathbf{x}_{k+1} = \mathcal{M}_{k,k+1}(\mathbf{x}_k) + T_k(\mathbf{e}_k), \tag{42}$$

$$\mathbf{e}_{k+1} = \mathcal{G}_{k,k+1}(\mathbf{x}_k, \mathbf{e}_k) \tag{43}$$

for $k = 0, \ldots, N-1$, where the observations are related to the model states by the Eq. (2), as previously. It is assumed that prior estimates, or "background estimates," $\mathbf{x}_0^b$ and $\mathbf{e}_0^b$ of $\mathbf{x}_0$ and $\mathbf{e}_0$ are known.

The augmented data assimilation problem is to minimize the weighted square errors between the model predictions and the observed system states, over the assimilation interval. The problem is written

**Problem 3** Minimize, with respect to $(\mathbf{x}_0^T,\ \mathbf{e}_0^T)^T$, the objective function

$$\begin{aligned} J = &\frac{1}{2}((\mathbf{x}_0 - \mathbf{x}_0^b)^T, (\mathbf{e}_0 - \mathbf{e}_0^b)^T)\mathbf{W}_0^{-1}((\mathbf{x}_0 - \mathbf{x}_0^b)^T, (\mathbf{e}_0 - \mathbf{e}_0^b)^T)^T \\ &+\frac{1}{2}\sum_{k=0}^{N}\left(\mathbf{y}_k - \mathcal{H}_k(\mathbf{x}_k)\right)^T \mathbf{R}_k^{-1}\left(\mathbf{y}_k - \mathcal{H}_k(\mathbf{x}_k)\right), \end{aligned} \tag{44}$$

subject to the augmented system equations (42) and (43).

The augmented system equations (42) and (43) are treated as strong constraints on the problem. The initial values $\mathbf{x}_0$ and $\mathbf{e}_0$ of the model states and model errors completely determine the response of the augmented system and are taken to be the control variables in the optimization. The weighting matrices $\mathbf{W}_0 \in \mathbb{R}^{(n+r)\times(n+r)}$ and $\mathbf{R}_k \in \mathbb{R}^{p_k \times p_k}$, $k = 0, 1, \ldots, N$, are assumed to be symmetric and positive definite. Since the matrix $\mathbf{W}_0$ is non-singular, the problem is well-posed and may be solved by any of the standard data assimilation schemes described in this chapter.

We remark that if a sequential method is used, then the initial weighting matrix $\mathbf{W}_0$ must contain cross-variable terms relating the states and the model errors or the observations may have no effect on the error estimates. In the variational methods, the weighting matrices (or covariance matrices) are implicitly propagated and this is not a problem. Furthermore, in the sequential methods, since the error estimates are updated at every observation point, the error estimates may not behave

smoothly. The variational method tends to average the analysis updates over time, automatically smoothing the estimates. For the variational methods, however, an additional set of adjoint equations must be solved to determine the gradient of the objective function with respect to the initial model errors $\mathbf{e}_0$.

Various applications of this approach to model error estimation, using both sequential and four-dimensional assimilation methods, are described in the literature for simplified models (Griffith 1997; Martin 2001; Martin et al. 1999; Griffith and Nichols 1996, 2000). These techniques have been applied successfully in practice to estimate systematic errors in operational equatorial ocean models (Martin et al. 2001; Bell et al. 2004).

## *5.4 Data Assimilation for Parameter Estimation*

Model errors also arise from inaccurate parameters in the model equations. The parameters generally enter the problem non-linearly, but since the required parameters are constants, the dynamics of the model errors in this case are simple. The error vector is usually also of small dimension relative to the dimension of the state variables. Using augmented forms of the equations, data assimilation can be applied directly to the estimation and calibration of the parameters. The augmented model equations take the form

$$\mathbf{x}_{k+1} = \mathcal{M}_{k,k+1}(\mathbf{x}_k, \mathbf{e}_k), \tag{45}$$

$$\mathbf{e}_{k+1} = \mathbf{e}_k, \tag{46}$$

where the vector $\mathbf{e}_0$ represents the unknown parameters in the model. The estimation problem is then to minimize the objective function (44), subject to the model equations (45) and (46).

The standard sequential and variational assimilation schemes can be applied to solve the problem. In the sequential methods, the form of the weighting (or covariance) matrices becomes important due to the non-linearity of the system equations. On the other hand, in the variational methods, the adjoint equations take a simple form and only the adjoints of the states are needed in order to find the gradients of the objective function with respect to both the states and the model errors. An application of a sequential scheme to the estimation of parameters in a simplified morphodynamic model for forecasting coastal bathymetry is described in Smith et al. (2008).

In summary, assimilation techniques for estimating random and systematic components of model errors along with the model states are described here. These techniques are effective and can lead to significantly improved forecasts (see chapter *Assimilation of Operational Data,* Andersson and Thépaut). For different types of error, different forms for the model error evolution are appropriate. Efficient methods for taking into account both random and systematic model errors are currently major topics of research.

## 6 Conclusions

The aims and basic concepts of data assimilation for geophysical and environmental systems are described here. Two approaches to the problem of data assimilation, sequential and variational assimilation, are introduced. A variety of assimilation schemes for discrete non-linear system models is derived and practical implementation issues are discussed. For all of these schemes, the model equations are assumed to be "perfect" representations of the true dynamical system. In practice the models contain both systematic errors and random noise. In the final section of the chapter we discuss data assimilation techniques for treating model errors of both types. Significant approximations are needed in order to implement these methods in "real-time," due to computational constraints. Further research on data assimilation schemes is needed and there remain many open problems for investigation. Details of current work on data assimilation schemes are given in subsequent chapters of this book.

## References

Anderson, J.L., 2001. An ensemble adjustment Kalman filter for data assimilation. *Mon. Weather Rev.*, **129**, 2884–2903.

Anderson, J.L. and S.L. Anderson, 1999. A Monte Carlo implementation of the non-linear filtering problem to produce ensemble assimilations and forecasts. *Mon. Weather Rev.*, **127**, 2741–2758.

Apte, A., M. Hairer, A.M. Stuart and J. Voss, 2007. Sampling the posterior: An approach to non-Gaussian data assimilation. *Physica D*, **230**, 50–64.

Bannister, R.N., D. Katz, M.J.P. Cullen, A.S. Lawless and N.K. Nichols, 2008. Modelling of forecast errors in geophysical fluid flows. *Int. J. Numeric. Methods Fluids*, **56**, 1147–1153, doi:10.1002/fld.1618.

Barnett, S. and R.G. Cameron, 1985. *Introduction to the Mathematical Theory of Control*, 2nd edition, Clarendon Press, Oxford, UK.

Bennett, A.F., 1992. *Inverse Methods in Physical Oceanography*. Cambridge University Press, Cambridge, UK.

Bell, M.J., M.J. Martin and N.K. Nichols, 2004. Assimilation of data into an ocean model with systematic errors near the equator. *Q. J. R. Meteorol. Soc.*, **130**, 873–894.

Bergthorsson, P. and B.R. Döös, 1955. Numerical weather map analysis. *Tellus*, **7**, 329–340.

Bierman, G.L., 1977. *Factorization Methods for Discrete Sequential Estimation*, Mathematics in Science and Engineering, vol. 128, Academic Press, New York.

Bishop, C.H., B.J. Etherton and S.J. Majumdar, 2001. Adaptive sampling with the ensemble trasform Kalman filter. Part I: Theoretical aspects, *Mon. Weather Rev.*, **129**, 420–436.

Boess, C., 2008. Using model reduction techniques within the incremental 4D-Var method, PhD Thesis, Fachbereich 3 – Mathematik und Informatik, Universitaet Bremen.

Bratseth, A.M., 1986. Statistical interpolation by means of successive corrections. *Tellus*, **38A**, 439–447.

Bunse-Gerstner, A., D. Kubalinska, G. Vossen and D. Wilczek, 2007. *h*2-norm optimal model reduction for large-scale discrete dynamical MIMO systems, Universitaet Bremen, Zentrum fur Technomathematik, Technical Report 07-04.

Burgers, G., P.J. van Leeuwen and G. Evensen, 1998. Analysis scheme in the ensemble Kalman filter. *Mon. Weather Rev.*, **126**, 1719–1724.

Cao, Y., J. Zhu, I.M. Navon and Z. Luo, 2007. A reduced-order approach to four-dimensional variational data assimilatio using proper orthogonal decomposition. *Int. J. Numeric. Meth. Fluids*, **53**, 1571–1583.

Cohn, S.E., A. da Silva, J. Guo, M. Sienkiewicz and D. Lamich, 1998. Assessing the effects of data selection with the DAO physical-space statistical analysis system. *Mon. Weather Rev.*, **126**, 2913–2926.

Courtier, P., 1997. Dual formulation of four-dimensional variational assimilation. *Q. J. R. Meteorol. Soc.*, **123**, 2449–2461.

Courtier, P., J.-N. Thépaut and A. Hollingsworth, 1994. A strategy for operational implementation of 4D-Var, using an incremental approach. *Q. J. R. Meteorol. Soc.*, **120**, 1367–1387.

Cressman, G., 1959. An optimal objective analysis system. *Mon. Weather Rev.*, **87**, 367–374.

Cullen, M.J.P., 2003. Four-dimensional variational data assimilation: A new formulation of the background covariance matrix based on a potential vorticity representation. *Q. J. R. Meteorol. Soc.*, **129**, 2777–2796.

Daley, R., 1993. *Atmospheric Data Analysis*. Cambridge University Press, Cambridge, UK.

Daley, R. and E. Barker, 2001. NAVDAS: Formulation and diagnostics. *Mon. Weather Rev.*, 129, 869–883.

Dee, D.P. and A.M. da Silva, 1998. Data assimilation in the presence of forecast bias. *Q. J. R. Meteorol. Soc.*, **117**, 269–295.

Derber, J.C., 1989. A variational continuous assimilation technique. *Mon. Weather Rev.*, 117, 2437–2446.

ECMWF., 2007. *Proceedings of the ECMWF Workshop on Flow Dependent Aspects of Data Assimilation* 11–13 June, 2007, ECMWF, UK.

Evensen, G., 2003. The ensemble Kalman filter: Theoretical formulation and practical implementation. *Ocean Dyn.*, 53, 343–367.

Ghil, M. and P. Malanotte-Rizzoli, 1991. Data assimilation in meteorology and oceanography. *Adv. Geophys.*, **33**, 141–266.

Giering, R. and T. Kaminski, 1998. Recipes for adjoint code construction. *ACM Trans. Math. Software*, **24**, 437–474.

Gratton, S., A.S. Lawless and N.K. Nichols, 2007. Approximate Gauss-Newton methods for nonlinear least-squares problems. *SIAM J. Optim.*, **18**, 106–132.

Griewank, A. and G.F. Corliss, 1991. *Automatic Differentiation of Algorithms*, SIAM, PA.

Griffith, A.K., 1997. *Data Assimilation for Numerical Weather Prediction Using Control Theory*, The University of Reading, Department of Mathematics, PhD Thesis. http://www.reading.ac.uk/maths/research/maths-phdtheses.asp#1997.

Griffith, A.K. and N.K. Nichols, 1996. Accounting for model error in data assimilation using adjoint methods. In *Computational Differentiation: Techniques, Applications and Tools,* Berz, M., C. Bischof, G. Corliss and A. Greiwank (eds.), SIAM, PA, pp 195–204.

Griffith, A.K. and N.K. Nichols, 2000. Adjoint techniques in data assimilation for estimating model error. *J. Flow, Turbulence Combustion*, **65**, 469–488.

Houtekamer, P.L. and H.L. Mitchell, 1998. Data assimilation using an ensemble Kalman filter technique. *Mon. Weather Rev.*, **126**, 796–811.

Johnson, C., B.J. Hoskins and N.K. Nichols, 2005a. A singular vector perspective of 4-DVar: Filtering and interpolation. *Q. J. R. Meteorol. Soc.*, 131, 1–20.

Johnson, C., N.K. Nichols and B.J. Hoskins, 2005b. Very large inverse problems in atmosphere and ocean modelling. *Int. J. Numeric. Methods Fluids*, **47**, 759–771.

Kalman, R.E., 1961. A new approach to linear filtering and prediction problems. *Trans. ASME, Series D*, **83**, 35–44.

Katz, D., 2007. The application of PV-based control variable transformations in variational data assimilation, PhD Thesis, Department of Mathematics, University of Reading. http://www.reading.ac.uk/maths/research/maths-phdtheses.asp#2007.

Kim, S., G.L. Eyink, J.M. Restrepo, F.J. Alexander and G. Johnson, 2003. Ensemble filtering for non-linear dynamics. *Mon. Weather Rev.*, **131**, 2586–2594.

Lawless, A.S., S. Gratton and N.K. Nichols, 2005. An investigation of incremental 4D-Var using non-tangent linear models. *Q. J. R. Meteorol. Soc.*, **131**, 459–476.

Lawless, A.S. and N.K. Nichols, 2006. Inner loop stopping criteria for incremental fourdimensional variational data assimilation. *Mon. Weather Rev.*, **134**, 3425–3435.

Lawless, A.S., N.K. Nichols and S.P. Ballard, 2003. A comparison of two methods for developing the linearization of a shallow water model. *Q. J. R. Meteorol. Soc.*, **129**, 1237–1254.

Lawless, A.S., N.K. Nichols, C. Boess and A. Bunse-Gerstner, 2008. Using model reduction methods within incremental four-dimensional variational data assimilation. *Mon. Weather Rev.*, **136**, 1511–1522.

Livings, D.M., S.L. Dance and N.K. Nichols, 2008. Unbiased ensemble square root filters. *Physica D*, **237**, 1021–1028.

Lorenc, A.C., 1986. Analysis methods for numerical weather prediction. *Q. J. R. Meteorol. Soc.*, **112**, 1177–1194.

Lorenc, A.C., 1988. Optimal non-linear objective analysis. *Q. J. R. Meteorol. Soc.*, **114**, 205–240.

Lorenc, A.C., R.S. Bell and B. Macpherson, 1991. The Met. Office analysis correction data assimilation scheme. *Q. J. R. Meteorol. Soc.*, **117**, 59–89.

Martin, M.J., 2001. *Data Assimilation in Ocean Circulation Models with Systematic Errors*, The University of Reading, Department of Mathematics, PhD Thesis. http://www.reading.ac.uk/maths/research/maths-phdtheses.asp#2001.

Martin, M.J., M.J. Bell and N.K. Nichols, 2001. Estimation of systematic error in an equatorial ocean model using data assimilation. In *Numerical Methods for Fluid Dynamics VII*, Baines, M.J. (ed.), ICFD, Oxford, pp 423–430.

Martin, M.J., N.K. Nichols and M.J. Bell, 1999. *Treatment of Systematic Errors in Sequential Data Assimilation*, Meteorological Office, Ocean Applications Division, Tech. Note, No. 21.

Nerger, L., W. Hiller and J. Schroeter, 2005. A comparison of error subspace Kalman filters. *Tellus*, **57A**, 715–735.

Nichols, N.K., 2003a. Data assimilation: Aims and basic concepts. In *Data Assimilation for the Earth System*, NATO Science Series: IV. Earth and Environmental Sciences 26, Swinbank, R., V. Shutyaev and W.A. Lahoz (eds.), Kluwer Academic Publishers, Dordrecht, The Netherlands, pp 9–20, 378pp.

Nichols, N.K., 2003b. Treating model error in 3-D and 4-D data assimilation. In *Data Assimilation for the Earth System*, NATO Science Series: IV. Earth and Environmental Sciences 26, Swinbank, R., V. Shutyaev and W.A. Lahoz (eds.), Kluwer Academic Publishers, Dordrecht, The Netherlands, pp 127–135, 378pp.

Ott, E., B.R. Hunt, I. Szunyogh, A.V. Zimin, E.J. Kostelich, M. Corazza, E. Kalnay, D.J. Patil and J.A.l Yorke, 2004. A local ensemble Kalman filter for atmospheric data assimilation. *Tellus*, **56A**, 415–428.

Pham, D.T., 2001. Stochastic methods for sequential data assimilation in strongly non-linear systems. *Mon. Weather Rev.*, **129**, 1194–1207.

Rabier, F., P. Courtier, J. Pailleux, O. Talalgrand and D. Vasiljevic, 1993. A comparison between four-dimensional variational assimilation and simplified sequential assimilation relying on three-dimensionsal variational analysis. *Q. J. R. Meteorol. Soc.*, **119**, 845–880.

Rabier, F., H. Järvinen, E. Klinker, J.-F. Mahfouf and A. Simmons, 2000. The ECMWF operational implementation of four-dimensional variational assimilation. I: Experimental results with simplified physics. *Q. J. R. Meteorol. Soc.*, **126**, 1143–1170.

Radnoti, G., Y. Trémolet, E. Andersson, L. Isaksen, E. Hólm and M. Janiscova, 2005. Diagnostics of linear and incremental approximations in 4D-Var revisted for highter resolution analysis, ECMWF Technical Memorandum, No. 467, European Centre for Medium-Range Weather Forecasts, Reading, UK.

Rosmond, T. and L. Xu, 2006. Development of NAVDAS-AR: Non-linear formulation and outer loop tests. *Tellus*, **58A**, 45–58.

Sasaki, Y., 1970. Some basic formulisms on numerical variational analysis. *Mon. Weather Rev.*, **98**, 875–883.

Smith, P.J., M.J. Baines, S.L. Dance, N.K. Nichols and T.R. Scott, 2008. Data assimilation for parameter estimation with application to a simple morphodynamic model, University of Reading, Department of Mathematics, Mathematics Report 2/2008.

Talagrand, O., 1981. A study on the dynamics of four-dimensional data assimilation. *Tellus*, **33**, 43–60.

Tippett, M.K., J.L. Anderson, C.H. Bishop, T.M. Hamil and J.S. Whitaker, 2003. Ensemble square root filters. *Mon. Weather Rev.*, **131**, 1485–1490.

Trémolet, Y., 2005. Incremental 4D-Var Convergence Study, ECMWF Technical Memorandum, No. 469, European Centre for Medium-Range Weather Forecasts, Reading, UK.

van Leeuwen, P.J., 2003. A variance minimizing filter for large scale applications. *Mon. Weather Rev.*, **131**, 2071–2084.

Vermeulen, P.T.M. and A.W. Heemink, 2006. Model-reduced variational data assinilation. *Mon. Weather Rev.*, **134**, 2888–2899.

Weaver, A. and P. Courtier, 2001. Correlation modelling on the sphere using a generalized diffusion equation. *Q. J. R. Meteorol. Soc.*, **127**, 1815–1846.

Willcox, K. and J. Peraire, 2002. Model reduction via the proper orthogonal decomposition. *AIAA J.*, **40**, 2323–2330.

Wlasak, M.A., N.K. Nichols and I. Roulstone, 2006. Use of potential vorticity for incremental data assimilation. *Q. J. R. Meteorol. Soc.*, **132**, 2867–2886.

Xu, L., T. Rosmond and R. Daley, 2005. Development of NAVDAS-AR: Formulation and initial tests of the linear problem. *Tellus*, **57A**, 546–559.

Zupanski, D., 1997. A general weak constraint applicable to operational 4D-Var data assimilation systems. *Mon. Weather Rev.*, **123**, 1112–1127.

Zupanski, M., 2005. Maximum likelihood ensemble filter: Theoretical aspects. *Mon. Weather Rev.*, **133**, 1710–1726.

# Variational Assimilation

**Olivier Talagrand**

## 1 Introduction

The expression *variational assimilation* designates a class of assimilation algorithms in which the fields to be estimated are explicitly determined as minimizers of a scalar function, called the *objective function*, that measures the misfit to the available data. In particular, *four-dimensional variational assimilation*, usually abbreviated as *4D-Var*, minimizes the misfit between a temporal sequence of model states and the observations that are available over a given assimilation window. As such, and contrary to the standard Kalman filter and, more generally, to sequential algorithms for assimilation, it propagates the information contained in the data both forward and backward in time.

From a numerical point of view, variational algorithms require the minimization of a scalar function defined over a large dimensional space. That is possible in practice through the systematic use of the *adjoint* of the assimilating model.

We first describe variational assimilation in the context of statistical linear estimation, which also underlies the theory of the Kalman filter (Sect. 2). This leads to the definition of a general form for the objective function to be minimized. Minimization methods and the adjoint approach for computing gradients, are then succinctly described (Sect. 3), as well as practical implementation of variational assimilation (Sect. 4). A number of problems, associated in particular with the strong non-linearity of the governing equations, are discussed (Sect. 5). The adjoint approach is further discussed, concerning in particular uses other than variational assimilation (Sect. 6). Conclusions follow in Sect. 7.

A large part of what follows is derived in the framework of Bayesian and statistical estimation. $\mathcal{E}[\,]$ will denote statistical expectation, and $\mathcal{N}(a, C)$ the Gaussian probability distribution (either scalar or vector) with expectation $a$ and covariance $C$. The superscript $^T$ will denote transposition.

O. Talagrand (✉)
Laboratoire de Météorologie Dynamique/CNRS, École Normale Supérieure, Paris, France
e-mail: Talagrand@lmd.ens.fr

W. Lahoz et al. (eds.), *Data Assimilation*, DOI 10.1007/978-3-540-74703-1_3,

## 2 Variational Assimilation in the Context of Statistical Linear Estimation

For an elementary introduction, consider the following situation. One wants to determine an unknown scalar quantity $x^t$ (i.e. true state) from two observations of the form

$$z_1 = x^t + \varepsilon_1 \tag{1a}$$

$$z_2 = x^t + \varepsilon_2. \tag{1b}$$

In these expressions, $\varepsilon_1$ and $\varepsilon_2$ are observational errors, whose exact values are unknown, but whose statistical properties are known. More precisely, it is assumed that these errors are centred ($\mathcal{E}[\varepsilon_1] = \mathcal{E}[\varepsilon_2] = 0$), mutually uncorrelated ($\mathcal{E}[\varepsilon_1\varepsilon_2] = 0$), and have respective variances $\mathcal{E}[{\varepsilon_1}^2] = s_1$ and $\mathcal{E}[{\varepsilon_2}^2] = s_2$. We look for an estimate of $x$, of the form $x^a \equiv \alpha_1 z_1 + \alpha_2 z_2$, ($\alpha_1$+$\alpha_2 = 1$), with $\alpha_1$ and $\alpha_2$ chosen so as to minimize the statistical quadratic estimation error $s \equiv \mathcal{E}[(x^a - x)^2]$. The answer is

$$x^a = \frac{s_2 z_1 + s_1 z_2}{s_1 + s_2} \tag{2}$$

that is each of the two measurements is weighted in inverse proportion to the variance of the error on that measurement. The corresponding quadratic estimation error, which minimizes $s$, and which we denote $s_a$, is given by

$$\frac{1}{s_a} = \frac{1}{s_1} + \frac{1}{s_2}. \tag{3}$$

The same estimate $x^a$ would be obtained by considering $z_1$ as a "background" estimate for $x$, and $z_2$ as an "observation" (or the reverse), and then applying the standard formulas for the Kalman filter.

The same estimate can also be obtained as the minimizer of the function

$$x \to J(x) \equiv \frac{1}{2}\left[\frac{(x - z_1)^2}{s_1} + \frac{(x - z_2)^2}{s_2}\right]. \tag{4}$$

The meaning of this expression is clear. The squared deviation of $x$ from either one of the two observations is weighted in inverse proportion of the variance of the error on that observation. Minimization of $J(x)$ therefore imposes that $x$ must fit either observation to within its own accuracy. This leads to the estimate given by Eqs. (2) and (3).

Variational assimilation, as it is implemented at present in meteorological and oceanographical applications (see chapters *Numerical Weather Prediction*, Swinbank; *Ocean Data Assimilation*, Haines), minimizes a function which generalizes Eq. (4). In particular, in the linear case, and as in the elementary example

above, it minimizes the statistical quadratic estimation error (on any component of the estimated fields individually), and is actually another algorithm for solving the same problem as the Kalman filter.

Consider the following more general estimation problem. Estimate an unknown vector $\mathbf{x}^t$ (with components $x_i^t$, $i = 1, \ldots, n$), belonging to *state space* $\mathcal{S}$, with dimension $n$, from a known *data vector* $\mathbf{z}$ (with components $z_j$, $j = 1, \ldots, m$), belonging to *data space* $\mathcal{D}$, with dimension $m$, of the form

$$\mathbf{z} = \boldsymbol{\Gamma}\mathbf{x}^t + \boldsymbol{\varepsilon}. \tag{5}$$

In Eq. (5), $\boldsymbol{\Gamma}$ is a known linear operator from $\mathcal{S}$ into $\mathcal{D}$, called the *data operator*, and represented by an $m \times n$-matrix. $\boldsymbol{\varepsilon}$ is a random vector in $\mathcal{D}$, called the *error vector*. The problem is, therefore, to invert the operator $\boldsymbol{\Gamma}$, taking into account, as far as possible, the statistical properties of the error $\boldsymbol{\varepsilon}$. The estimate of $\mathbf{x}^t$ is sought in the form of a linear (and a priori non-homogeneous) function of $\mathbf{z}$, viz.

$$\mathbf{x}^a = \mathbf{a} + \mathbf{A}\mathbf{z}, \tag{6}$$

where $\mathbf{a}$ is a vector of $\mathcal{S}$, and $\mathbf{A}$ is a linear operator from $\mathcal{D}$ into $\mathcal{S}$. $\mathbf{a}$ and $\mathbf{A}$ are to be determined under the following two conditions:

(i) The estimate $\mathbf{x}^a$ is invariant in a change of origin in state space (for instance, if the unknown $\mathbf{x}^t$ contains temperatures, the result must be independent of whether those temperatures are expressed in degrees Celsius or in Kelvins);
(ii) For any component $x_i^t$ of $\mathbf{x}^t$, the statistical expectation of the square of the corresponding estimation error $x_i^a - x_i^t$ is minimized.

The solution to this problem is given by

$$\mathbf{x}^a = (\boldsymbol{\Gamma}^T\mathbf{S}^{-1}\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}^T\mathbf{S}^{-1}(\mathbf{z} - \boldsymbol{\mu}) \tag{7}$$

[i.e., $\mathbf{A} = (\boldsymbol{\Gamma}^T\mathbf{S}^{-1}\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}^T\mathbf{S}^{-1}$ and $\mathbf{a} = -\mathbf{A}\boldsymbol{\mu}$], where $\boldsymbol{\mu} \equiv \mathcal{E}[\boldsymbol{\varepsilon}]$ and $\mathbf{S} \equiv \mathcal{E}[(\boldsymbol{\varepsilon} - \boldsymbol{\mu})(\boldsymbol{\varepsilon} - \boldsymbol{\mu})^T]$ are, respectively, the expectation and covariance matrix of the error $\boldsymbol{\varepsilon}$. It is seen that $\mathbf{A}$ is a left-inverse of $\boldsymbol{\Gamma}$ (i.e., $\mathbf{A}\boldsymbol{\Gamma} = \mathbf{I}_n$, where $\mathbf{I}_n$ is the unit matrix of order $n$), with the consequence that the estimate $\mathbf{x}^a$ is unbiased, ($\mathcal{E}[\mathbf{x}^a - \mathbf{x}^t] = 0$), and that the corresponding estimation error has covariance

$$\mathbf{P}^a \equiv \mathcal{E}\left[(\mathbf{x}^a - \mathbf{x}^t)(\mathbf{x}^a - \mathbf{x}^t)^T\right] = (\boldsymbol{\Gamma}^T\mathbf{S}^{-\mathbf{1}}\boldsymbol{\Gamma})^{-1}. \tag{8}$$

Condition (ii) above means that the trace of $\mathbf{P}^a$ is the minimum trace that can be obtained among all possible linear estimates of $\mathbf{x}$.

Equations (7) and (8) generalize Eqs. (2) and (3). The estimate $\mathbf{x}^a$ is called the *Best Linear Unbiased Estimate* (*BLUE*) of $\mathbf{x}$ from $\mathbf{z}$ (the term *Best Linear Unbiased Estimator* is also used). Its explicit determination requires the knowledge of (at most) the expectation $\boldsymbol{\mu}$ and the covariance matrix $\mathbf{S}$ of the data error $\boldsymbol{\varepsilon}$.

Taking Eq. (7) at face value, the unambiguous definition of the *BLUE* requires the matrix $\mathbf{S}$, and then the matrix $\mathbf{\Gamma}^T\mathbf{S}^{-1}\mathbf{\Gamma}$, to be invertible. The need for invertibility of $\mathbf{S}$ is only apparent (without going into full details, $\mathbf{S}$ is singular when some components of $\mathbf{x}$ are exactly observed; it then suffices to restrict the estimation to those components that are not exactly observed). The condition for invertibility of $\mathbf{\Gamma}^T\mathbf{S}^{-1}\mathbf{\Gamma}$, once $\mathbf{S}$ is invertible, is on the other hand real. It is equivalent to the condition that the null space of the data operator $\mathbf{\Gamma}$ is restricted to the 0-vector

$$\mathbf{\Gamma x} = \mathbf{0} \;\Leftrightarrow\; \mathbf{x} = \mathbf{0} \tag{9}$$

or, equivalently, that $\mathbf{\Gamma}$ has rank equal to the dimension $n$ of $\mathbf{x}$. This means that the data vector $\mathbf{z}$ contains information, either directly or indirectly, on every component of $\mathbf{x}$. The problem of determining $\mathbf{x}$ from $\mathbf{z}$ is *overdetermined*. This requires that $m \geq n$. There must be at least as many scalar data in $\mathbf{z}$ as there are scalar parameters to be determined. We will set $m = n + p$. The condition given by Eq. (9) will be called the *determinacy* condition.

The *BLUE* possesses a number of important properties.

- As already mentioned, the operator $\mathbf{A}$ is a left-inverse of $\mathbf{\Gamma}$. This means that, if the data are exact ($\boldsymbol{\varepsilon} = 0$ in Eq. 9), then so is the estimate $\mathbf{x}^a$;
- The *BLUE* is invariant in a change of origin in either data or state space. It is also invariant in any invertible linear change of coordinates in either space. This means, for instance, that a profile of observed temperatures can be transformed, through the hydrostatic equation, into a profile of geopotential values without altering the estimated fields. It also means that the horizontal wind can be estimated in terms of geometrical coordinates, or in terms of its divergence and vorticity. The result will be the same. This condition of invariance also means that the *BLUE* is independent of the choice of a scalar product, either in state or data space. For instance, for any symmetric definite positive matrix $\mathbf{C}$, the quantity $(\mathbf{x}^a–\mathbf{x})^T\mathbf{C}(\mathbf{x}^a–\mathbf{x})$, which is one (among infinitely many) measure of the magnitude of the estimation error $(\mathbf{x}^a–\mathbf{x})$, is minimized by the *BLUE*. The invariance of the *BLUE* in any invertible change of linear coordinates can also be expressed by saying that Eqs. (7) and (8) are more than vector-matrix equations. They are *tensor equations*, valid in any system of linear coordinates;
- When the data error $\boldsymbol{\varepsilon}$ is Gaussian, $\boldsymbol{\varepsilon} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{S})$, the *BLUE* achieves Bayesian estimation, in the sense that the conditional probability distribution for the state vector $\mathbf{x}$, given the data vector $\mathbf{z}$, is the Gaussian distribution with expectation $\mathbf{x}^a$ and covariance matrix $\mathbf{P}^a$, as given by Eqs. (7) and (8). In condensed notation,

$$\mathbf{P}(\mathbf{x}|\mathbf{z}) = \mathcal{N}(\mathbf{x}^a, \mathbf{P}^a).$$

It is easily verified that the *BLUE* $\mathbf{x}^a$ can be obtained as the minimizer of the following scalar function, defined over state space

$$\mathbf{x} \rightarrow J(\mathbf{x}) \equiv \frac{1}{2}[\mathbf{\Gamma x} - (\mathbf{z} - \boldsymbol{\mu})]^T\mathbf{S}^{-1}\left[\mathbf{\Gamma x} - (\mathbf{z} - \boldsymbol{\mu})\right]. \tag{10}$$

This expression generalizes Eq. (4). Its significance is clear. For any vector $\mathbf{x}$ in state space, $\mathbf{\Gamma x}$ is what the data operator $\mathbf{\Gamma}$ would produce if it was applied on $\mathbf{x}$. $J(\mathbf{x})$ is then a measure of the magnitude of the discrepancy between $\mathbf{\Gamma x}$ and the unbiased data vector $\mathbf{z}$–$\boldsymbol{\mu}$. Through the inverse covariance matrix $\mathbf{S}^{-1}$, that measure possesses two notable properties. First, it weights the data according to their accuracy. Second, it is physically non-dimensional, making it possible to combine in a consistent way data of a different physical nature.

Variational assimilation, as it exists at present in meteorology and oceanography, minimizes objective functions of the form of Eq. (10), with the only difference, to be discussed later, that moderately non-linear operators $\mathbf{\Gamma}$ are used. What follows is a more detailed description of how variational assimilation is implemented in practice, and of the main results it produces.

The first step in the minimization of a function such as that given by Eq. (10) is to remove the bias in the data by subtracting the error expectation $\boldsymbol{\mu}$ from the data vector. Unless specified otherwise, it will be assumed below that this has been done, and the expectation $\boldsymbol{\mu}$ will not appear any more explicitly in the equations. But it must be kept in mind that implementation of variational assimilation requires the prior knowledge, and subtraction from the data, of the error expectation, or *bias*. Failure to properly remove the bias in the data will, in general, result in the presence of residual biases in the estimated fields (chapter *Bias Estimation*, Ménard, discusses bias in data assimilation).

When the determinacy condition (Eq. 9) is verified, the data vector $\mathbf{z}$ can always be transformed, through linear invertible operations, into two components of the following form. First, an explicit estimate of the true state vector $\mathbf{x}^t$, of form

$$\mathbf{x}^b = \mathbf{x}^t + \boldsymbol{\varepsilon}^b, \tag{11}$$

where $\boldsymbol{\varepsilon}^b$ is an error; second, an additional set of data, of the form

$$\mathbf{y} = \mathbf{H}\mathbf{x}^t + \boldsymbol{\varepsilon}^o, \tag{12}$$

with dimension $p = m - n$. In this equation, $\mathbf{H}$ is a linear operator, represented by a $p \times n$-matrix, and $\boldsymbol{\varepsilon}^o$ is an error. In addition, the transformations that lead to Eqs. (11) and (12) can always be defined in such a way that the errors $\boldsymbol{\varepsilon}^b$ and $\boldsymbol{\varepsilon}^o$ are uncorrelated

$$\mathcal{E}\left[\boldsymbol{\varepsilon}^b(\boldsymbol{\varepsilon}^o)^T\right] = 0 \tag{13}$$

It is in the form of Eqs. (11) and (12) that data are most often available in meteorological and oceanographical applications. The component $\mathbf{x}^b$ is a prior, or *background* estimate of the unknown state vector $\mathbf{x}$ at a given time $k$ (usually a recent forecast, or a climatological estimate). As for the additional vector $\mathbf{y}$, it consists of observations depending on the state vector through the *observation operator* $\mathbf{H}$. The uncorrelation hypothesis (Eq. 13), although certainly disputable, is often (if not always) made. Equations (11) and (12), together with Eq. (13), are also assumed in

the standard Kalman filter. We stress here that Eqs. (11) and (13) are no more restrictive than, but exactly equivalent to, Eq. (5) together with the determinacy condition, Eq. (9).

Introducing the covariance matrices of the errors $\boldsymbol{\varepsilon}^b$ and $\boldsymbol{\varepsilon}^o$

$$\mathbf{P}^b \equiv \mathcal{E}\left[\boldsymbol{\varepsilon}^b(\boldsymbol{\varepsilon}^b)^T\right], \quad \mathbf{R} \equiv \mathcal{E}\left[\boldsymbol{\varepsilon}^o(\boldsymbol{\varepsilon}^o)^T\right], \tag{14}$$

Equations (7) and (8) take the following form, used in particular in the Kalman filter

$$\mathbf{x}^a = \mathbf{x}^b + \mathbf{P}^b\mathbf{H}^T[\mathbf{H}\mathbf{P}^b\mathbf{H}^T + \mathbf{R}]^{-1}(\mathbf{y} - \mathbf{H}\mathbf{x}^b) \tag{15a}$$

$$\mathbf{P}^a = \mathbf{P}^b - \mathbf{P}^b\mathbf{H}^T[\mathbf{H}\mathbf{P}^b\mathbf{H}^T + \mathbf{R}]^{-1}\mathbf{H}\mathbf{P}^b. \tag{15b}$$

We recall that the vector

$$\mathbf{d} \equiv \mathbf{y} - \mathbf{H}\mathbf{x}^b, \tag{16}$$

is called the *innovation vector*, and that the matrix $\mathbf{H}\mathbf{P}^b\mathbf{H}^T + \mathbf{R}$, the inverse of which appears in Eqs. (15a) and (15b), is the covariance matrix of $\mathbf{d}$

$$\mathbf{H}\mathbf{P}^b\mathbf{H}^T + \mathbf{R} = \mathcal{E}[\mathbf{d}\mathbf{d}^T]. \tag{17}$$

As for the objective function (Eq. 10), it takes under decomposition of Eqs. (11) and (12) the following form

$$J(\mathbf{x}) = \frac{1}{2}(\mathbf{x} - \mathbf{x}^b)^T[\mathbf{P}^b]^{-1}(\mathbf{x} - \mathbf{x}^b) + \frac{1}{2}(\mathbf{H}\mathbf{x} - \mathbf{y})^T\mathbf{R}^{-1}(\mathbf{H}\mathbf{x} - \mathbf{y}). \tag{18}$$

The meaning of this expression is clear. The first term on the right hand side of Eq. (18) is a measure of the deviation of $\mathbf{x}$ from the background, while the second term is a measure of the deviation from the observation.

Several situations are encountered in the practice of meteorology and oceanography, which we are going to describe in some detail, giving more explicit expressions for the general form (Eq. 18) of the objective function.

The simplest situation is when a background $\mathbf{x}^b$, of form given by Eq. (11), is available at some time $k$, together with observations, of form given by Eq. (12), that have been performed at the same time (or over a period of time short enough so that the flow can be considered stationary). Minimization of the objective function (Eq. 18) will produce an estimate of the state of the flow at time $t$. One then speaks in that case of *three-dimensional variational analysis*, often abbreviated as *3D-Var*.

A different, more complex, situation is encountered when one wants to assimilate observations that are distributed over a period of time over which the evolution of the flow cannot be neglected. Let us assume observations are available at successive times $k = 0, 1, \ldots, K$, of the form

$$\mathbf{y}_k = \mathbf{H}_k \mathbf{x}_k^t + \boldsymbol{\varepsilon}_k^o, \tag{19}$$

where $\mathbf{x}_k^t$ is the exact true state of the flow at time $k$, $\mathbf{H}_k$ is a linear observation operator, and $\boldsymbol{\varepsilon}_k^o$ an observational error with covariance matrix $\mathbf{R}_k$. The observational errors are assumed to be uncorrelated in time. It is assumed in addition that the temporal evolution of the flow is described by the equation

$$\mathbf{x}_{k+1}^t = \mathbf{M}_k \mathbf{x}_k^t + \boldsymbol{\eta}_k, \tag{20}$$

with known model linear operator $\mathbf{M}_k$, and random model error $\boldsymbol{\eta}_k$.

Assume in addition that a background ${\mathbf{x}_0}^b$, with error covariance matrix ${\mathbf{P}_0}^b$, and error uncorrelated with the observational errors in Eq. (19), is available at time $k = 0$.

If the model error is ignored, any initial condition $\mathbf{x}_0$ at time $k = 0$ defines a model solution

$$\mathbf{x}_{k+1} = \mathbf{M}_k \mathbf{x}_k \quad k = 0, \ldots, K-1. \tag{21}$$

The objective function

$$J(\mathbf{x}_0) = \frac{1}{2}\left(\mathbf{x}_0 - \mathbf{x}_0^b\right)^T \left[\mathbf{P}_0^b\right]^{-1} \left(\mathbf{x}_0 - \mathbf{x}_0^b\right) + \frac{1}{2}\sum_{k=0}^{K} (\mathbf{H}_k \mathbf{x}_k - \mathbf{y}_k)^T [\mathbf{R}_k]^{-1} (\mathbf{H}_k \mathbf{x}_k - \mathbf{y}_k) \tag{22}$$

which is of the general form given by Eq. (10), measures the distance between the model solution (Eq. 21) and the data. Minimization of $J(\mathbf{x}_0)$ will define the initial condition of the model solution that fits the data most closely. Following a terminology first introduced by Sasaki (1970a, b, c), this is called *strong constraint four-dimensional variational assimilation*, often abbreviated as *strong constraint 4D-Var*. The words "strong constraint" stress the fact that the model identified by Eq. (21) must be exactly verified by the sequence of estimated state vectors.

If the model error is taken into account, Eq. (20) defines an additional set of "noisy" data. We assume the model error $\boldsymbol{\eta}_k$ in Eq. (20) to have covariance matrix $\mathbf{Q}_k$, to be uncorrelated in time and to be uncorrelated with observation and background errors. Equation (10) then gives the following expression for the objective function defining the *BLUE* of the sequence of states $\{\mathbf{x}_k, k = 0, \ldots, K\}$

$$\begin{aligned} J(\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_K) = {} & \frac{1}{2}\left(\mathbf{x}_0 - \mathbf{x}_0^b\right)^T \left[\mathbf{P}_0^b\right]^{-1} (\mathbf{x}_0 - \mathbf{x}_0^b) \\ & + \frac{1}{2}\sum_{k=0}^{K} (\mathbf{H}_k \mathbf{x}_k - \mathbf{y}_k)^T [\mathbf{R}_k]^{-1} (\mathbf{H}_k \mathbf{x}_k - \mathbf{y}_k) \\ & + \frac{1}{2}\sum_{k=0}^{K-1} (\mathbf{x}_{k+1} - \mathbf{M}_k \mathbf{x}_k)^T [\mathbf{Q}_k]^{-1} (\mathbf{x}_{k+1} - \mathbf{M}_k \mathbf{x}_k). \end{aligned} \tag{23}$$

The objective function is now a function of the whole sequence of states $\{\mathbf{x}_k, k = 0, \ldots, K\}$. Minimization of an objective function of the form given by Eq. (23), where the model equations are present as noisy data to be fitted by the analysed fields like any other data, is called, again according to the terminology introduced by Sasaki (1970a, b, c), *weak constraint four-dimensional variational assimilation*, abbreviated as *weak constraint 4D-Var*.

Equations (22) and (23), with appropriate redefinition of the state and observation spaces, are particular cases of Eq. (10). Another type of variational algorithm can be defined from Eq. (15a), which can be written as

$$\mathbf{x}^a = \mathbf{x}^b + \mathbf{P}^b\mathbf{H}^T\mathbf{w}, \tag{24}$$

where the vector $\mathbf{w} \equiv [\mathbf{H}\mathbf{P}^b\mathbf{H}^T + \mathbf{R}]^{-1}\,\mathbf{d}$ minimizes the objective function

$$K(\mathbf{v}) \equiv \frac{1}{2}\mathbf{v}^T[\mathbf{H}\mathbf{P}\mathbf{H}^T + \mathbf{R}]\mathbf{v} - \mathbf{d}^T\mathbf{v}. \tag{25}$$

This function is defined on the dual of the observation space, which has dimension $p$. Minimization of Eq. (25) corresponds to the *dual* approach to variational assimilation, by opposition to the *primal* approach, given by Eq. (18). The dual approach is also known as defining the *Physical Space Assimilation System* (*PSAS*, pronounced "pizzazz"; the word *Physical* is historical). Just as Eqs. (18), (22), and (23) are particular forms of Eq. (10), the dual approach can be used in any of the situations corresponding to those three equations. Depending on the conditions of the problem, and especially on the relative dimension of the state and observation space, it may be more advantageous to use the primal or the dual approach. A significant difference is that the dual approach uses the error covariance matrices $\mathbf{P}^b$ and $\mathbf{R}$ in their direct forms, while the primal approach requires their inverses. Another difference is that the dual approach requires an explicit background $\mathbf{x}^b$, while the primal approach can be implemented, in the general form given by Eq. (10), without an explicit background (it only requires the determinacy condition, Eq. 9).

All forms of variational assimilation given by Eqs. (18), (22), (23), and (25) have been used, or at least extensively studied, for assimilation of meteorological and oceanographical observations. The theory of the *BLUE* requires the data operators ($\mathbf{\Gamma}$, $\mathbf{H}$ and $\mathbf{M}_k$ in the above notations) to be linear. In practice, this condition is rarely verified. In particular, variational assimilation of form given by Eq. (22) or (23) is almost always implemented with a non-linear model. From a heuristic point of view, it is clear that, if the non-linearity is in a sense sufficiently small, variational assimilation, even if it does not solve a clearly identified estimation problem, is likely to produce useful results (this point will be further discussed in Sect. 5 below). The dual approach, on the other hand, explicitly uses the transpose observation operator $\mathbf{H}^T$, and requires exact linearity.

# 3 Minimization Methods: The Adjoint Approach

## *3.1 Gradient Methods for Minimization*

Variational assimilation aims at minimizing an objective function of one of the forms defined in the previous section. The objective functions we will consider can be exactly quadratic or not. We will make a slight change of notation, and will systematically denote by $\mathbf{x}$, and will call *control variable*, the argument of the function to be minimized; in Eq. (23), the control variable is the whole sequence $\mathbf{x}_0, \ldots, \mathbf{x}_K$, while it is $\boldsymbol{\nu}$ in Eq. (25). The control variable belongs to the *control space*, whose dimension will be denoted by $N$. We will denote by $\partial J/\partial \mathbf{x}$ the gradient of $J$ with respect to $\mathbf{x}$, i.e., the $N$-vector whose components are the partial derivatives of $J$ with respect to the components $x_i$ of $\mathbf{x}$, viz.,

$$\frac{\partial J}{\partial \mathbf{x}} = \left( \frac{\partial J}{\partial x_i} \right)_{i=1,\ldots,N}. \tag{26}$$

The gradient is equal to 0 at the minimum of the objective function. One way to determine the minimum could conceivably be (as is actually often done in simple small dimension problems) to determine analytical expressions for the components of the gradient, and then to solve a system of $N$ scalar equations for the minimizing components of $\mathbf{x}$. In meteorological and oceanographical applications, the complexity of the computations defining the objective function (in 4D-Var, these calculations include the temporal integration of a numerical dynamical model of the flow over the assimilation window) makes it totally inconceivable even to obtain analytical expressions for the gradient. Another way to proceed is to implement an iterative minimization algorithm, which determines a sequence of successive approximations $\mathbf{x}^{(l)}$ of the minimizing value of $\mathbf{x}$, viz.,

$$\mathbf{x}^{(l+1)} = \mathbf{x}^{(l)} - \mathbf{D}^{(l)}, \tag{27}$$

where $\mathbf{D}^{(l)}$ is at every iteration an appropriately chosen vector in control space. One possibility is to choose $\mathbf{D}^{(l)}$ along the direction of the local gradient $\partial J/\partial \mathbf{x}$. Algorithms which are based on that choice, called *steepest descent* algorithms, turn out, however, not to be numerically very efficient. In other algorithms, the vector $\mathbf{D}^{(l)}$ is determined as a combination of the local gradient and of a number of gradients computed at previous steps of the iteration, Eq. (27) (see, e.g., Bonnans et al. 2003). All minimization methods that are efficient for large dimensions are of the form given by Eq. (27), and require the explicit determination, at each iteration step, of the local gradient $\partial J/\partial \mathbf{x}$. They are called *gradient methods*. Since one cannot hope to obtain an analytical expression for the gradient, it must be determined numerically. One possibility could be to determine it by finite differences, by imposing in turn a perturbation $\Delta x_i$ on all components $x_i$ of the control vector, and approximating the partial derivative $\partial J/\partial x_i$ by the difference quotient

$$\frac{\partial J}{\partial x_i} \approx \frac{J(\mathbf{x} + \Delta x_i) - J(\mathbf{x})}{\Delta x_i}. \tag{28}$$

This, however, would require $N$ explicit computations of the objective function, i.e., in the case of four-dimensional assimilation, $N$ integrations of the assimilating model. Although that has actually been done for variational assimilation of meteorological observations, in an experimental setting, and with a relatively small dimension model (Hoffman 1986), it would clearly be impossible in any practical application.

## *3.2 The Adjoint Method*

The *adjoint method* allows numerical computation of the gradient of a scalar function at a cost that is at most a few times the cost of the direct computation of that function. Adjoint equations are an extremely powerful mathematical and numerical tool. They are central to the theory of *optimal control*, i.e., the theory of how the behaviour of a physical system can be controlled by acting on some of its components (see for instance the book by Lions 1971). Adjoint equations can also be used for solving mathematical problems in their own right. The use of adjoint equations in meteorological and oceanographical applications was advocated by the Russian school of mathematics at an early stage of development of numerical modelling of the atmosphere and ocean (see, e.g., Marchuk 1974). We are going to demonstrate the method of adjoint equations in the special case of strong constraint 4D-Var (Eq. 22), in the most general case where the model and observation operators can be non-linear.

In order to stress the possible non-linearity of the model and observation operators, we now introduce the non-linear model operator, $\mathcal{M}_k(\,)$, and the non-linear observation operator, $\mathcal{H}_k(\,)$. The notation for operators used hitherto in this chapter, $\mathbf{M}_k$ and $\mathbf{H}_k$ (denoting linear model and observation operators, respectively), being reserved hereafter for the *Jacobians* (matrices of partial derivatives) of $\mathcal{M}_k(\,)$ and $\mathcal{H}_k(\,)$, respectively. We rewrite Eqs. (21) and (22) with non-linear operators as

$$J(\mathbf{x}_0) = \frac{1}{2}\left(\mathbf{x}_0 - \mathbf{x}_0^b\right)^T \left[\mathbf{P}_0^b\right]^{-1} \left(\mathbf{x}_0 - \mathbf{x}_0^b\right) + \frac{1}{2}\sum_{k=0}^{K} (\mathcal{H}_k(\mathbf{x}_k) - \mathbf{y}_k)^T [\mathbf{R}_k]^{-1} (\mathcal{H}_k(\mathbf{x}_k) - \mathbf{y}_k) \tag{29a}$$

with

$$\mathbf{x}_{k+1} = \mathcal{M}_k(\mathbf{x}_k), \quad k = 0, \ldots, K-1. \tag{29b}$$

Our purpose is to determine the gradient $\partial J/\partial \mathbf{x}_0$ of $J$ with respect to $\mathbf{x}_0$. That gradient is characterized by the property that, for any perturbation $\delta\mathbf{x}_0$ of $\mathbf{x}_0$, the corresponding variation of $J$ is, to first order with respect to $\delta\mathbf{x}_0$, equal to

$$\delta J = \left( \frac{\partial J}{\partial \mathbf{x}_0} \right)^T \delta \mathbf{x}_0. \tag{30}$$

The perturbation $\delta \mathbf{x}_0$ results at later times in perturbations which, through differentiation of Eq. (29b), are given to first order by

$$\delta \mathbf{x}_{k+1} = \mathbf{M}_k \delta \mathbf{x}_k, \quad k = 0, \ldots, K-1, \tag{31}$$

where, as said, $\mathbf{M}_k$ is the Jacobian of $\mathbf{x}_{k+1}$ with respect to $\mathbf{x}_k$. Equation (31) is called the *tangent linear equation* of Eq. (29b). Although the dependence is not explicit in Eq. (31), it must be kept in mind that the Jacobian $\mathbf{M}_k$ will, in general, depend in the non-linear case on the local value of $\mathbf{x}_k$.

As for the first order variation of the objective function $J$, it is given by differentiation of Eq. (29a), viz.,

$$\delta J = \left( \mathbf{x}_0 - \mathbf{x}_0^b \right)^T \left[ \mathbf{P}_0^b \right]^{-1} \delta \mathbf{x}_0 + \sum_{k=0}^{K} (\mathcal{H}_k(\mathbf{x}_k) - \mathbf{y}_k)^T \mathbf{R}_k^{-1} \mathbf{H}_k \delta \mathbf{x}_k, \tag{32}$$

where $\mathbf{H}_k$ is the local Jacobian of $\mathcal{H}_k$, and where the $\delta \mathbf{x}_k$'s are given by Eq. (31).

$\delta J$ is a compound function of $\delta \mathbf{x}_0$ through the $\delta \mathbf{x}_k$'s. Our purpose is to "skip" the intermediate $\delta \mathbf{x}_k$'s, and to obtain a direct dependence of $\delta J$ with respect to $\delta \mathbf{x}_0$ of form given by Eq. (30). To that end, we introduce at each time $k = 1,\ldots, K$ a vector $\boldsymbol{\lambda}_k$, belonging to the dual of state space (and therefore with dimension $n$), and to be defined more precisely later. We form the products $\boldsymbol{\lambda}_k^T (\delta \mathbf{x}_k - \mathbf{M}_{k-1} \delta \mathbf{x}_{k-1})$, which, according to Eq. (31), are equal to 0. Subtracting those products from the right-hand side of Eq. (32) yields

$$\begin{aligned} \delta J = \left( \mathbf{x}_0 - \mathbf{x}_0^b \right)^T \left[ \mathbf{P}_0^b \right]^{-1} \delta \mathbf{x}_0 + \sum_{k=0}^{K} (\mathcal{H}_k(\mathbf{x}_k) - \mathbf{y}_k)^T \mathbf{R}_k^{-1} \mathbf{H}_k \delta \mathbf{x}_k \\ - \sum_{k=0}^{K} \boldsymbol{\lambda}_k^T (\delta \mathbf{x}_k - \mathbf{M}_{k-1} \delta \mathbf{x}_{k-1}) \end{aligned} \tag{33}$$

(subtracting rather than adding the products is of course arbitrary, but convenient). We now transform Eq. (33) by first using the fact that the transpose of a matrix product is the product of the corresponding transposes, taken in reversed order. For instance, the product $(\mathcal{H}_k(\mathbf{x}_k) - \mathbf{y}_k)^T \mathbf{R}_k^{-1} \mathbf{H}_k$ is equal to $\left( \left[ \mathbf{H}_k^T \mathbf{R}_k^{-1} (\mathcal{H}_k(\mathbf{x}_k) - \mathbf{y}_k) \right] \right)^T$ (where use has been made of the fact that the covariance matrix $\mathbf{R}_k$ is symmetric), thus transforming the (scalar) quantity $(\mathcal{H}_k(\mathbf{x})_k - \mathbf{y}_k)^T \mathbf{R}_k^{-1} \mathbf{H}_k \delta \mathbf{x}_k$ into the scalar product of the two $n$-vectors $\mathbf{H}_k^T \mathbf{R}_k^{-1} (\mathcal{H}_k(\mathbf{x}_k) - \mathbf{y}_k)$ and $\delta \mathbf{x}_k$. Performing that operation on all terms in Eq. (33) and gathering all terms with common factor $\delta \mathbf{x}_k$ yields

$$\begin{aligned}\delta J = &\left\{\left[\mathbf{P}_0^b\right]^{-1}\left(\mathbf{x}_0 - \mathbf{x}_0^b\right) + \mathbf{H}_0^T\mathbf{R}_0^{-1}(\mathcal{H}_0(\mathbf{x}_0) - \mathbf{y}_0) + \mathbf{M}_0^T\boldsymbol{\lambda}_1\right\}^T \delta\mathbf{x}_0 \\ &+ \sum_{k=1}^{K-1}\left\{\mathbf{H}_k^T\mathbf{R}_k^{-1}(\mathcal{H}_k(\mathbf{x}_k) - \mathbf{y}_k) - \boldsymbol{\lambda}_k + \mathbf{M}_k^T\boldsymbol{\lambda}_{k+1}\right\}^T \delta\mathbf{x}_k \\ &+ \left\{\mathbf{H}_K^T\mathbf{R}_K^{-1}(\mathcal{H}_K(\mathbf{x}_K) - \mathbf{y}_K) - \boldsymbol{\lambda}_K\right\}^T \delta\mathbf{x}_K.\end{aligned} \tag{34}$$

This expression is valid for any choice of the $\boldsymbol{\lambda}_k$'s. It is seen that choosing

$$\boldsymbol{\lambda}_K = \mathbf{H}_K^T\mathbf{R}_K^{-1}\left(\mathcal{H}_K(\mathbf{x}_K) - \mathbf{y}_K\right), \tag{35a}$$

and then recursively

$$\boldsymbol{\lambda}_k = \mathbf{M}_k^T\boldsymbol{\lambda}_{k+1} + \mathbf{H}_k^T\mathbf{R}_k^{-1}\left(\mathcal{H}_k(\mathbf{x}_k) - \mathbf{y}_k\right) \text{ for } k = K-1, \ldots, 1, \tag{35b}$$

eliminates all $\delta\mathbf{x}_k$ terms in Eq. (34), except the $\delta\mathbf{x}_0$ term. Defining further

$$\boldsymbol{\lambda}_0 = \mathbf{M}_0^T\boldsymbol{\lambda}_1 + \left[\mathbf{P}_0^b\right]^{-1}\left(\mathbf{x}_0 - \mathbf{x}_0^b\right) + \mathbf{H}_0^T\mathbf{R}_0^{-1}(\mathcal{H}_0(\mathbf{x}_0) - \mathbf{y}_0) \tag{35c}$$

there remains

$$\delta J = \boldsymbol{\lambda}_0^T\delta\mathbf{x}_0, \tag{36}$$

which shows that $\boldsymbol{\lambda}_0$ is the required gradient of the objective function with respect to the initial condition $\mathbf{x}_0$ (see Eq. 30).

Equations (35a), (35b), and (35c) make up the *adjoint* of the tangent linear equation, Eq. (31). The word "adjoint" comes from the fact that Eqs. (35a), (35b), and (35c) are built on the transpose matrices $\mathbf{H}_k^T$ and $\mathbf{M}_k^T$, which are particular cases of the more general notion of adjoint operators. The adjoint equation is defined for the particular solution $\mathbf{x}_k$ of the basic equation (29b) for which the gradient is to be determined. It depends on that solution through the terms $\mathcal{H}_k(\mathbf{x}_k) - \mathbf{y}_k$ and, in the case of either a non-linear model operator $\mathcal{M}_k$ or a non-linear observation operator $\mathcal{H}_k$, through the transpose Jacobians $\mathbf{M}_k^T$ and/or $\mathbf{H}_k^T$. It is often said for convenience that Eqs. (35a), (35b), and (35c) define the adjoint of the basic model given by Eq. (29b), but it must be kept in mind that the adjoint equation is defined for a particular solution of that model.

The computations to be performed for determining the gradient $\partial J/\partial\mathbf{x}_0$ for given initial condition $\mathbf{x}_0$ are now clearly defined:

(1) Starting from $\mathbf{x}_0$, integrate the basic equation (29b). Store the corresponding solution $\mathbf{x}_k$ in memory;
(2) Starting from the "final" condition (Eq. 35a) at time $K$, integrate the adjoint equations (35b) and (35c) backward in time. The required gradient is $\boldsymbol{\lambda}_0$. The direct solution $\mathbf{x}_k$ is necessary for computing the terms $\mathbf{H}_k^T\mathbf{R}_k^{-1}(\mathcal{H}_k(\mathbf{x}_k) - \mathbf{y}_k)$

and, in case the basic model (Eq. 29b) is non-linear, for determining the transpose Jacobian $\mathbf{M}_k^T$.

The determination of the gradient therefore requires one forward integration of the basic model (Eq. 29b), followed by one backward integration of the adjoint model (Eqs. 35a, 35b, and 35c). The latter is a modified form of the direct model, and the corresponding cost must be of similar magnitude to the cost of integrating the direct model. It can be rigorously shown that, in terms of the number of arithmetic operations to be performed, the cost of one adjoint computation of the gradient $\partial J/\partial \mathbf{x}_0$ is at most four times the cost of one computation of the objective function, $J$. In meteorological and oceanographical applications, the cost of one adjoint integration (in terms of elapsed computer time) is typically twice the cost of one direct integration. This ratio is basically independent of the dimension $N$ of the control variable, and makes the adjoint computation of a gradient much more economical than the $N$ direct model integrations that would be required if the gradient was to be computed by explicit perturbations. It is this fact that made variational assimilation possible at all in the first place.

Not surprisingly, there is a price to be paid for this major reduction in computing time. The price, as seen above, is the necessity to store in memory the direct solution $\mathbf{x}_k$. More precisely, what has to be kept in memory (or else to be recomputed in the course of the adjoint integration) are all quantities that are arguments of non-linear operations in the direct integration. Relaxing the storage constraint, for instance by using a more economical approximate adjoint, is difficult. Experience shows that minimization algorithms, especially efficient ones, are very sensitive to even slight misspecification of the gradient. The question of how the cost of variational assimilation can be reduced will be discussed in the next section.

The description that has just been given of the adjoint method is fundamentally sufficient for 4D-Var. It obviously covers the case of 3D-Var (minimization of an objective function of form given by Eq. 18), which does not involve a dynamical model of the flow. In that case, of course, only the transpose Jacobian $\mathbf{H}^T$ of the observation operator is needed.

The first attempt at using the adjoint approach for variational assimilation of meteorological observations was made by Penenko and Obraztsov (1976), on a simple one-level linear atmospheric model, and with synthetic data. Later attempts were made by Lewis and Derber (1985), Le Dimet and Talagrand (1986) and Talagrand and Courtier (1987). Courtier and Talagrand (1987) first used real data, while Thacker and Long (1988) made the first attempt at using adjoint equations for variational assimilation of oceanographical observations. Thépaut and Courtier (1991) first used a full primitive equation meteorological model. These early works showed that variational assimilation of meteorological or oceanographical observations was numerically feasible at an acceptable cost, and produced physically realistic results. Variational assimilation was progressively applied to more and more complex numerical models. It was introduced in 1997 in operational prediction, in the strong-constraint formulation, at the European Centre for Medium-Range Weather

Forecasts, ECMWF (Klinker et al. 2000), and in 2000 at the French Meteorological Service (Météo-France). In both places, operational implementation of variational assimilation has resulted in significant improvements of the ensuing forecasts (see chapter *Assimilation of Operational Data*, Andersson and Thépaut). Some of these improvements were due to side effects not directly linked to the variational character of the assimilation, but others, especially in a number of specific meteorological situations, were due to better consistency between the assimilated states and the dynamics of the atmospheric flow. Since then, other meteorological services, such as the Japan Meteorological Agency, the Meteorological Office (United Kingdom), the Meteorological Service of Canada and the China Meteorological Administration, have introduced variational assimilation in their operational prediction system. All these schemes are of the strong-constraint form, and use a 6-h assimilation window (12-h in the case of ECMWF). In addition, ECMWF, after having produced several sets of reanalysed past observations, all based on sequential assimilation algorithms, is now running a new reanalysis project (the *ERA-Interim* project, http://www.ecmwf.int/research/era/do/get/era-interim) based on variational assimilation. A specific advantage of variational assimilation in the case of reanalysis of past data is that it propagates information both forward and backward in time, thus allowing the use of observations that have been performed after estimation time.

Similar developments have taken place in oceanography, and variational assimilation using the adjoint of oceanographic circulation models is now commonly used for many diverse applications (although not so far for operational oceanographic prediction). Those applications include determination of the initial conditions of the flow, as described above (see, e.g., Weaver and Anderson 1997; Vialard et al. 2003; Ricci et al. 2005), but also identification of "parameters", such as wind stress at the surface of the ocean (Vossepoel et al. 2004). Egbert et al. (1994) and Louvel (2001) used the dual approach through minimization in dual observation space of an objective function of form given by Eq. (25). In that approach, each iteration of the minimization process requires first a backward integration of the adjoint model, followed by a forward integration of the tangent linear model. Variational assimilation has also extended to other fields of geophysics and environmental sciences, such as atmospheric chemistry (Fisher and Lary 1995; Errera and Fonteyn 2001; Elbern et al. 2007; Lahoz et al. 2007 – see also chapters in Part IV, *Atmospheric Chemistry*), or surface hydrology (Reichle 2000 – see chapter *Land Surface Data Assimilation*, Houser et al.). Other extensions of the variational methodology, that have largely benefited from the experience in meteorology, have been to terrestrial magnetism (Fournier et al. 2007; Sun et al. 2007) and seismology (Tromp et al. 2005).

## 4 Practical Implementation

If the principle of variational assimilation and of the adjoint method is conceptually perfectly clear and rigorous, practical implementation of variational assimilation raises a number of serious problems. We will discuss below the specific problems

associated with the development and validation of a code for performing the adjoint computations defined by Eq. (35), and are going to consider first a number of purely numerical problems.

## *4.1 The Incremental Approach*

The developments of the previous section seem to require that it is the adjoint of the complete model (Eq. 29b) that has to be used for the computation of the gradient of the objective function. A Numerical Weather Prediction (NWP) model is an extremely complex and lengthy code, and the ensuing "all-or-nothing" choice (take the complete adjoint of the model, or else do nothing) seems particularly impractical. Simplifying the adjoint equation as such, without modification of the direct model nor of the objective function, is not an appropriate solution. That would lead to an approximate gradient of the objective function, and, as has already been said, experience shows that minimization algorithms, especially efficient ones, are very sensitive to even slight misspecification of the gradient. A convenient and versatile solution, known as the *incremental approach* to variational assimilation, has been introduced by Courtier et al. (1994). Several variants of that approach exist. We are going to describe the one that is conceptually the simplest.

The basic idea is to simplify the dynamical model (Eq. 29b) to a form that is both more economical and more manageable, in particular as concerns the adjoint. But that is not done on the model (Eq. 29b) itself, but rather on the tangent linear model (Eq. 31). A *reference solution* $\mathbf{x}_k{}^{(0)}$ of the basic equation (29b) having been determined (emanating for instance from the background $\mathbf{x}_0{}^b = \mathbf{x}_0{}^{(0)}$), the corresponding tangent linear model (Eq. 31) is modified to

$$\delta\mathbf{x}_{k+1} = \mathbf{L}_k \delta\mathbf{x}_k, \quad k = 0, \ldots, K-1, \tag{37}$$

where $\mathbf{L}_k$ is, at any time $k$, an appropriately chosen "simpler" operator than the Jacobian $\mathbf{M}_k$. Consistency then requires to modify the basic model (Eq. 29b) in such a way that the tangent linear equation corresponding to solution $\mathbf{x}_k{}^{(0)}$ is Eq. (37). This is achieved by making the initial condition $\mathbf{x}_0 \equiv \mathbf{x}_0{}^{(0)} + \delta\mathbf{x}_0$ evolve into $\mathbf{x}_k \equiv \mathbf{x}_k{}^{(0)} + \delta\mathbf{x}_k$, where $\delta\mathbf{x}_k$ itself evolves according to Eq. (37). That makes the basic dynamics linear.

As for the objective function (Eq. 29a), several possibilities exist, at least when the observation operators are non-linear. One possibility is to linearize those operators just as the model operator $\mathcal{M}_k$ has been linearized. This leads to replacing the quantity $\mathcal{H}_k(\mathbf{x}_k)$ by $\mathcal{H}_k\left(\mathbf{x}_k^{(0)}\right) + \mathbf{N}_k\delta\mathbf{x}_k$, where $\mathbf{N}_k$ is an appropriate simplified linear operator (possibly, but not necessarily, the Jacobian of $\mathcal{H}_k$ at point $\mathbf{x}_k$). The objective function (Eq. 29a) is then replaced by

$$\begin{aligned} J_1(\delta\mathbf{x}_0) = &\frac{1}{2}\left(\delta\mathbf{x}_0 + \mathbf{x}_0^{(0)} - \mathbf{x}_0^b\right)^T \left[\mathbf{P}_0^b\right]^{-1} \left(\delta\mathbf{x}_0 + \mathbf{x}_0^{(0)} - \mathbf{x}_0^b\right) \\ &+ \frac{1}{2}\sum_{k=0}^{K} (\mathbf{N}_k\delta\mathbf{x}_k - \mathbf{d}_k)^T \mathbf{R}_k^{-1} (\mathbf{N}_k\delta\mathbf{x}_k - \mathbf{d}_k), \end{aligned} \tag{38}$$

where the $\delta\mathbf{x}_k$'s are subject to Eq. (37), and where $\mathbf{d}_k \equiv \mathbf{y}_k - \mathcal{H}_k\left(\mathbf{x}_k^{(0)}\right)$ is the innovation at time $k$.

The function given by Eq. (38) is an exactly quadratic function of the initial perturbation $\delta\mathbf{x}_0$. The minimizing perturbation $\delta\mathbf{x}_{0,m}$ defines a new initial state $\mathbf{x}_0{}^{(1)} \equiv \mathbf{x}_0{}^{(0)} + \delta\mathbf{x}_{0,m}$, from which a new solution $\mathbf{x}_k{}^{(1)}$ of the basic equation (Eq. 29b) is computed. The process is then repeated for solution $\mathbf{x}_k{}^{(1)}$.

This defines a system of two-level nested loops for minimization of the original objective function (Eq. 29a). The fundamental advantage of the incremental approach is that it allows one to define at will the simplified linearized operators $\mathbf{L}_k$ and $\mathbf{N}_k$. Many degrees of freedom are available for ensuring an appropriate tradeoff between practical implementability and meteorological accuracy and usefulness. The simplified dynamics in Eq. (37) can itself be modified in the course of the minimization, by progressively introducing more and more complex dynamics or "physics" in the successive outer loops.

It is the incremental method which, after the adjoint method, makes variational assimilation feasible. It is implemented, either in the form that has just been described or in slightly different variants, in most (if not all) operational NWP systems that use variational assimilation. At ECMWF, it is implemented with two outer loops, the approximations introduced in the linearized dynamics (Eq. 37) consisting first, of a reduced spatial resolution (from triangular spectral truncation T799 to T255 for the second outer loop) and, second, of a simplified "physical" package.

An obvious question is whether the nested-loop process of the incremental process converges and, if it does, to what it converges. In the case where the linearized operators $\mathbf{L}_k$ and $\mathbf{N}_k$ vary from one outer loop to the next, the possible convergence of the process can depend on the way those operators vary. In particular, convergence to the minimum of the original objective function (Eq. 29a) is possible only if the linear operators $\mathbf{L}_k$ and $\mathbf{N}_k$ converge to the corresponding Jacobians $\mathbf{M}_k$ and $\mathbf{H}_k$ at that minimum. The question of the convergence of the incremental process has been studied in some detail by Trémolet (2007) on the ECMWF 4D-Var system. Numerical tests show that the process does not converge asymptotically, at least in the conditions in which it is implemented at ECMWF. The way the incremental approach is implemented, at ECMWF and elsewhere, is largely based on empirical tuning.

## *4.2 First-Guess-At-the-Right-Time 3D-Var*

An extreme case of the incremental approach is what is called *First-Guess-At-the-right-Time 3D-Var*, or *FGAT 3D-Var*. It can be described as a process of form of Eqs. (37) and (38) in which the simplified linear operator $\mathbf{L}_k$ is taken as the identity operator. This process is four-dimensional in that the observations distributed over the assimilation window are compared with their analogues in a time-evolving reference integration of the assimilating model. But it is three-dimensional in that the minimization of the objective function (Eq. 38) does not use any explicit dynamics

other than the trivial dynamics expressed by the unit operator, and that the numerical implementation is in effect three-dimensional. The FGAT 3D-Var approach, which is implemented through a unique minimization (no nested loops), has been shown to improve the quality of the assimilated fields, simply through the fact that is effectively uses a more exact innovation vector than does standard 3D-Var, in which all observations over the assimilation window are compared to the same first-guess field.

## 5 Further Considerations on Variational Assimilation

Independently of its numerical and algorithmic properties, the major advantage of variational assimilation is that it takes into account, through the adjoint equation, the temporal evolution of the uncertainty in the state of the flow, at least over the assimilation window. Although (contrary to the Kalman filter) it does not explicitly compute the evolution of the uncertainty as such (and, in particular, does not produce an explicit estimate of the uncertainty in the estimated fields), it determines an approximation of the minimizing solution of the objective function (Eq. 29), which depends on the dynamics of the flow, and of the temporal evolution of the uncertainty. This was shown in full detail by Thépaut et al. (1993), who compared the impact of individual observations in a 3D-Var process, which ignores the temporal evolution of the uncertainty, and a 4D-Var process. The impact was significantly different, and strongly dependent on the dynamical state of the flow, in the latter case.

Significant impact does not of course mean positive impact. All operational implementations of 4D-Var have been preceded by the development and implementation of a 3D-Var system. This is very convenient in that it allows progressive introduction of the various components of the full 4D-Var system. But it also provides the opportunity for systematic comparison of 3D-Var and 4D-Var. The comparison has always shown the superiority of 4D-Var, in particular in terms of the quality of the ensuing forecasts. Similar comparisons have also been performed, with the same conclusions, on other, non-operational assimilation systems. See also Lorenc and Rawlins (2005) for a detailed discussion of 3D-Var and 4D-Var.

All operational implementations of 4D-Var have so far been of the strong constraint form. In spite of the constant improvement of NWP models, the hypothesis of a perfect model is of course highly disputable. Weak-constraint assimilation, which corresponds to minimization of an objective function of form given by Eq. (23), would certainly be desirable. It however requires a quantitative estimate, in the form of the covariance matrix $\mathbf{Q}_k$, of the model error. A reliable estimate may be difficult to obtain. Derber (1989) has suggested identifying a possible systematic bias in the model by introducing that bias in the control variable. Other authors (Zupanski 1997; Trémolet 2006) have studied algorithms of the general form given by Eq. (23). There is some indication (M. Fisher, personal communication) that weak constraint variational assimilation could be useful over longer assimilation windows (24 h or

more) than used in strong constraint assimilation. That is easily understandable in view of the fact that the perfect model hypothesis becomes less and less valid as the length of the assimilation window increases.

The primal weak-constraint objective function (Eq. 23) becomes singular in the limit of a perfect model ($\mathbf{Q}_k$=0). As already said, the dual approach uses the data error covariance matrices in their direct form, so that the dual objective function (Eq. 25), as defined for weak constraint variational assimilation, is regular for $\mathbf{Q}_k = 0$. This means that the same dual algorithm can be used for both strong- and weak-constraint variational assimilation. This is an attactive feature of the dual approach.

Courtier (1997) has shown that, subject to an appropriate preconditioning of the dual variable $\boldsymbol{\nu}$ in Eq. (25), the numerical conditioning (and therefore the numerical cost) of the dual algorithm is the same as that of the primal approach. In variational assimilation, it is actually the repeated numerical integrations of the direct and adjoint models that takes the major part of the computations, and the numerical cost of strong- and weak-constraint variational assimilation is fundamentally the same. This point is discussed in more detail in Louvel (2001).

The dual approach requires strict linearity of the operator $\mathbf{H}$ in Eq. (25) which, in the case of variational assimilation, means strict linearity of the model and observation operators. Auroux and Blum (2002, 2004) have introduced a double-loop algorithm (which has some similarity with the incremental approach described above) in which successive linear problems of form given by Eq. (25) are solved, each one being based on a linearization about the result of the previous one.

More generally, and independently of the particular numerical algorithm that is used, the validity of the linear approach defined by Eqs. (7) and (10) is questionable in meteorological and oceanographical applications. It has already been said that, from a purely heuristic point of view, the linear approach must be valid if the non-linearities are in a sense small enough. A more accurate description of the real situation that is encountered in meteorology and oceanography is given, rather than by Eqs. (11) and (12), by

$$\mathbf{x}^b = \mathbf{x}^t + \boldsymbol{\varepsilon}^b, \tag{39}$$

$$\mathbf{y} = \mathcal{H}^*(\mathbf{x}^t) + \boldsymbol{\varepsilon}, \tag{40}$$

where $\mathcal{H}^*(\mathcal{H} - \text{star})$ denotes a non-linear observation operator. In the case of 3D-Var, $\mathcal{H}^*$ is the observation operator at estimation time. In the case of 4D-Var, the vector $\mathbf{y}$ denotes the complete temporal sequence of observations, and the operator $\mathcal{H}^*$ includes the (non-linear) dynamical model. The knowledge of the data (Eqs. 39 and 40) is equivalent to the knowledge of Eq. (39) together with what can be called the non-linear innovation vector

$$\mathbf{d} \equiv \mathbf{y} - \mathcal{H}^*(\mathbf{x}^b) = \mathcal{H}^*(\mathbf{x}^t) - \mathcal{H}^*(\mathbf{x}^b) + \boldsymbol{\varepsilon}. \tag{41}$$

If the background $\mathbf{x}^b$ is close enough to the real unknown state $\mathbf{x}^t$, $\mathbf{d}$ can be approximated by

$$\mathbf{d} \approx \mathbf{H}(\mathbf{x}^t - \mathbf{x}^b) + \boldsymbol{\varepsilon}, \tag{42}$$

where $\mathbf{H}$ is here the Jacobian of the full operator $\mathcal{H}^*$ at point $\mathbf{x}^b$. If the so-called *tangent linear approximation* defined by Eq. (42) is valid, Eqs. (39), (40), (41), and (42) define an estimation problem that is linear with respect to the deviation $\mathbf{x}^t$–$\mathbf{x}^b$ of the real state with respect to the background $\mathbf{x}^b$. Equations (15) and (18) are then valid, $\mathbf{H}$ being the Jacobian of $\mathcal{H}^*$. In the case of 4D-Var, this leads to minimization of an objective function of the incremental form given by Eqs. (37) and (38), where the operators $\mathbf{L}_k$ and $\mathbf{N}_k$ replace the exact Jacobians $\mathbf{M}_k$ and $\mathbf{H}_k$ along the (full non-linear) reference model solution.

Both direct (see, e.g., Lacarra and Talagrand 1988) and indirect evidence shows that the tangent linear approximation is valid for large scale geostrophic atmospheric flow (scales larger than 200 km) up to about 24–48 h. This limit, however, rapidly decreases with decreasing spatial scales, to be of the order of a few hours for convective scales. For oceanic geostrophic flow (scales larger than a few tens of kilometres), the limit is a few weeks.

The developments of this chapter are therefore fully valid within those limits. It is to be stressed, however, that in circumstances where the tangent linear approximation is known or hypothesized to be valid, the linearization in Eq. (42) is rarely performed explicitly. Either fully non-linear operators are kept in the objective function to be minimized, or (as is actually the case in the incremental approach described above) approximations that go further than Eq. (42) are implemented. The only case where the linearization given by Eq. (42) seems to have explicitly been implemented is in the above-mentioned works of Auroux and Blum (2002, 2004) relative to the dual approach, which requires exactly linear operators.

But the question arises of what is to be done in circumstances when the tangent linear approximation is not valid. In the context of 4D-Var, there are actually two different questions, depending on the strength of the non-linearities. If the non-linearities are weak, the minimization of an objective function of the general form given by Eq. (29) remains numerically feasible, but may not be justified on the basis of estimation theory. If the non-linearities are strong, even the numerical minimization of the objective function, owing for instance to the presence of distinct minima, can raise difficulties.

These questions have not been discussed so far in much depth. One can mention the work of Pires et al. (1996), who studied variational assimilation for a strongly chaotic non-linear system (specifically, the celebrated three-parameter system of Lorenz 1963). These authors have shown that the objective function given by Eq. (29) possesses an increasing number of local minima with increasing length of the assimilation window. This can be easily understood in view of the repeated folding in state space that is associated with chaos. They have defined a procedure, called *Quasi-Static Variational Assimilation* (*QSVA*), in which the length of the assimilation window, starting from a value for which the objective function

(Eq. 29) possesses a unique minimum, is progressively increased. Each new minimization is started from the result of the previous one. This allows one to keep track of the absolute minimum of the objective function, at least if the temporal density of observations is in a sense high enough. QSVA has been implemented on a quasi-geostrophic atmospheric model by Swanson et al. (1998) who have been able to usefully extend variational assimilation (in the hypothesis of a perfect model) to assimilation windows as long as 5 days. This is largely beyond the limit of validity of the tangent linear approximation. QSVA, or a similar algorithm, could possibly be implemented in operational practice, for instance by using successive overlapping assimilation windows.

Other developments have taken place recently at the research level. Carrassi et al. (2008) have defined a 3D-Var system in which the control variable, instead of consisting of the whole state vector, is restricted to the deviations from the background along the (relatively few) unstable modes of the system. This approach is now being extended to 4D-Var (Trevisan, personal communication). A somewhat similar work has been performed by Liu et al. (2008), who have developed a low-order incremental 4D-Var system. The background error covariance matrix $\mathbf{P}_0{}^b$ (Eq. 38) is defined, not on the basis of an a priori statistical model, but on the basis of the dispersion of an ensemble of background forecasts. As in Carrassi et al. (2008), the control space is not the entire state space, but the state spanned by the background forecasts. Taking advantage of the relatively small dimension of the control space, and of the linearity associated with the incremental character of the procedure, it is not necessary to use an adjoint code for computing the gradient of the objective function. That can be achieved through simple transposition of an appropriate matrix. The results obtained are competitive with a fully-fledged 4D-Var. The "ensemble" feature of those works give them similarity with the Ensemble Kalman filter (see chapter *Ensemble Kalman Filter: Current Status and Potential*, Kalnay).

Both those works suggest that it could be possible to achieve substantial numerical gain, without significant degradation of the final results (and even maybe without the use of an adjoint), by restricting the control variable to an appropriate subspace of the whole state space.

All the algorithms that have been described above are based on the minimization of an objective function of the general form given by Eqs. (10), (18) or (29), which is quadratic in terms of the data-minus-unknown differences, with weights equal to the inverse of the covariance matrices of the corresponding errors. Equations (10) and (18) correspond to least-variance statistical linear estimation, while Eq. (29) corresponds to an extension to weakly non-linear situations. Other forms for the objective function have also been considered. In particular, Fletcher and Zupanski (2006) and Fletcher (2007), following a general Bayesian approach, propose to maximize the conditional probability density function for the state of the flow, given the data. In the case of linear data operators and Gaussian errors, this leads to minimization of an objective function of form given by Eq. (10). Those authors consider the case of lognormal distributions, which are more appropriate for bounded variables such as humidity. This leads to a significantly different form for the objective function.

## 6 More on the Adjoint Method

The adjoint method has been demonstrated above in the particular case of the objective function given by Eq. (29). It is actually very general, and defines a systematic approach for computing the (exact) gradient of a differentiable scalar function with respect to its arguments. Although this may not be obvious from the above developments, the adjoint method consists in a systematic use of the chain rule for differentiation of a compound function. Proceeding backward through the original sequence of computations, it recursively computes the partial derivatives of the scalar function under consideration with respect to the variables in those computations (see, e.g., Talagrand 2003). As such, the adjoint method can be used not only for optimization purposes, as in variational assimilation, but (actually more simply) for determination of gradients as such, and for sensitivity studies.

The advantages and disadvantages of variational assimilation will be further discussed in the Conclusions below (Sect. 7). But its major disadvantage (at least for variational assimilation as it exists at present) is probably the need for developing the adjoint code which performs computations in Eq. (35). Not only must the adjoint code be developed, but it must be carefully validated, since experience shows that even minor errors in the computed gradient can significantly degrade the efficiency of the minimization (if not totally inhibit it). In addition, NWP models are constantly modified, and the corresponding modifications must be made on the adjoint code. Writing the adjoint of a code at the same time as the direct code involves only a rather small amount of additional work (10 or 20%). But developing the adjoint of an already existing code can require a substantial amount of work, and can be a very tedious and time-consuming task. On the other hand, the fact that adjoint computation is in essence a systematic use of the chain rule for differentiation leads to perfectly defined "adjoint" coding rules, which make the development of an adjoint code, if lengthy and tedious, at least totally straightforward. These rules are described in, e.g., Talagrand (1991), Giering and Kaminski (1998) or Kalnay (2002).

Those same rules are at the basis of "adjoint compilers", i.e., software pieces that are designed to automatically develop the adjoint of a given code (see, e.g., http://www.fastopt.de/; Hascoët and Pascual 2004). The adjoint of a particular piece of code is independent of the rest of the code, and automating the derivation of the adjoint instructions for a sequence of coding instructions, which is a purely local operation, is relatively easy. Other aspects, such as the choice and management of non-linear variables to be kept in memory from the direct integration, or to be recomputed in the course of the adjoint integration, require a global view of the code, and are more difficult to automate. For that reason, the use of these software pieces still requires experience of adjoint coding as well as some preparatory work, but they are nevertheless extremely useful, and very substantially reduce the amount of time and work necessary for developing the adjoint of an atmospheric or oceanic circulation model.

The adjoint approach is used in assimilation of meteorological and oceanographical observations for numerically solving, through an iterative minimization process,

an optimization problem. Now, as said above, what the adjoint equations really do is simply compute the gradient of one scalar output of a numerical process with respect to (potentially all) the input parameters of that process. As such, the adjoint approach can be used for sensitivity studies of outputs with respect to inputs, independently of any optimization or minimization. It will be useful to use the adjoint approach when the number of output parameters whose sensitivity is sought is smaller than the number of input parameters with respect to which the sensitivity is sought (in the inverse case, direct perturbation of the input parameters will be more economical).

Actually, the first proponents of the use of the adjoint approach in meteorology and oceanography had primarily sensitivity studies in mind (Marchuk 1974; Hall et al. 1982). Adjoint models have been used to perform sensitivity studies of many different kinds: sensitivity of the atmospheric flow with respect to initial or lateral boundary conditions (Errico and Vukisevic 1992; Rabier et al. 1992; Gustafsson et al. 1998); sensitivity of the global oceanic circulation to parameters (Marotzke et al. 1999); sensitivity of biogeochemical processes (Waelbroeck and Louis 1995); and sensitivity of atmospheric chemical processes (Zhang et al. 1998). See also the special issue of *Meteorologische Zeitschrift* (Ehrendorfer and Errico 2007) devoted to Adjoint Applications in Dynamic Meteorology. Two specific types of applications are worthy of particular mention. The first one has to do with the identification, for a particular situation, of the unstable components of the flow. In its simplest form, this amounts to determining the so-called *singular vectors* of the flow, i.e., the perturbations that amplify most rapidly, over a period of time, in the tangent linear approximation (Lacarra and Talagrand 1988; Farrell 1989; Urban 1993). This has been extended by Mu and colleagues (Mu 2000; Mu et al. 2003) to *Non-Linear Singular Vectors* (*NLSV*s), i.e., perturbations that amplify most rapidly in the full non-linear evolution. A condition must then be imposed on the initial amplitude of the perturbation, which leads to a (technically more difficult to solve) constrained optimization problem. Both linear and non-linear singular vectors allow accurate diagnostic and analysis of instability (Moore and Farrell 1993; Mu and Zhang 2006; Rivière et al. 2008). A related, but more specific, application is the identification of the components of the flow to which a particular feature of the future evolution of the flow (such as, for instance, the deepening of a depression) is most sensitive. This allows one to "target" observations in order to optimize the prediction of the feature under consideration. This has been implemented successfully on the occasion of specific campaigns (see, e.g., Langland et al. 1999; Bergot and Doerenbecher 2002). Observation targeting through adjoint methods is further discussed in Buizza et al. (2007). Another, potentially very promising, application of the adjoint method is the determination of the sensitivity of analysed and predicted fields to observations. It is then the adjoint of the whole assimilation and prediction process, and not only of the assimilating model, that has to be used (Langland and Baker 2004). This has led to very useful diagnostics of the value and usefulness of various types of observations (Langland and Cardinali, personal communication).

## 7 Conclusion

Variational assimilation has now become a basic tool of numerical meteorology and oceanography, and a major component of operational NWP in several major meteorological services. Together with the Ensemble Kalman filter (see chapter *Ensemble Kalman Filter: Current Status and Potential*, Kalnay), it is one of the two most advanced and powerful assimilation methods. The specific advantages of variational assimilation are rather obvious. It is very versatile and flexible, and allows for easy introduction of a new type of observation in an assimilation system. It suffices to specify the corresponding observation operator and the first- and second-order statistical moments of the associated error. It automatically propagates information both forward and backward in time, and makes it easy to take into account temporal correlation between errors (either observation or model errors). To the author's knowledge, this last possibility has been used so far on only one occasion, for taking into account temporally correlated errors in high frequency observations of surface pressure (Järvinen et al. 1999). But it can be extremely useful, especially for the treatment of model error and of the associated temporal correlation (time will presumably come when this will be necessary).

Variational assimilation is costly in that it requires the development, validation and maintenance of the adjoint of the assimilating model, as well as of the various observation operators. This is a time-consuming task. However, owing to the gain in experience and expertise, and to the continuous improvement of adjoint compilers, that task progressively becomes easier and easier. And, as discussed in the previous section, adjoints, once they are available, can be used for many other applications than assimilation, and in particular to powerful diagnostic studies.

Assimilation of meteorological and oceanographical observations may be at a turning point. It seems that the limits of what can be obtained from statistical linear estimation (i.e., from Eq. (7) and its various generalizations to weakly non-linear situations) are being reached. The only exception is likely Quasi-Static Variational Assimilation, discussed in Sect. 5, which is based on minimization of objective functions of form given by Eq. (29), but whose limits have not been identified. Statistical linear estimation is at the basis of variational assimilation and of the "Kalman" component of the Ensemble Kalman filter. It can legitimately be said that the ultimate purpose of assimilation is to achieve Bayesian estimation, i.e., to determine the conditional probability distribution for the state of the atmosphere (or the ocean), given all the relevant available information. In view of the large dimension of the state of the atmosphere, the only possible way to describe the conditional probability distribution seems to be through an ensemble of points in state space, as indeed the Ensemble Kalman filter already does. A basic question is then to determine whether it is possible to develop methods for ensemble variational assimilation, which would produce a Bayesian ensemble, while retaining the specific advantages of variational assimilation, namely easy propagation of information both forward and backward in time, and possibility to easily take error temporal correlations into account. Some results suggest that this should be possible.

## References


Auroux, D. and J. Blum, 2002. A dual data assimilation method for a layered quasi-geostrophic ocean model. *RACSAM*, **96**, 315–320.

Auroux, D. and J. Blum, 2004. Data assimilation methods for an oceanographic problem. In *Multidisciplinary Methods for Analysis, Optimization and Control of Complex Systems, Mathematics in Industry*. Vol. 6, Capasso, V. and J. Periaux (eds.), Springer, Berlin, pp 179–194.

Bergot, T. and A. Doerenbecher, 2002. A study on the optimization of the deployment of targeted observations using adjoint-based methods. *Q. J. R. Meteorol. Soc*., **128**, 1689–1712.

Bonnans, J.-F., J.-C. Gilbert, C. Lemaréchal and C. Sagatizabal, 2003. *Numerical Optimization – Theoretical and Practical Aspects*. Springer-Verlag, Berlin, 485 pp.

Buizza, R., C. Cardinali, G. Kelly and J.-N. Thépaut, 2007. The value of observations. II: The value of observations located in singular-vector-based target areas. *Q. J. R. Meteorol. Soc*., **133**, 1817–1832.

Carrassi, A., A. Trevisan, L. Descamps, O. Talagrand and F. Uboldi, 2008. Controlling instabilities along a 3DVar analysis cycle by assimilating in the unstable subspace: A comparison with the EnKF. *Nonlinear Process. Geophys.*, **15**, 503–521.

Courtier, P., 1997. Dual formulation of four-dimensional variational assimilation. *Q. J. R. Meteorol. Soc*., **123**, 2449–2461.

Courtier, P. and O. Talagrand, 1987. Variational assimilation of meteorological observations with the adjoint vorticity equation. II: Numerical results. *Q. J. R. Meteorol. Soc.*, **113**, 1329–1347.

Courtier, P., J.-N. Thépaut and A. Hollingsworth, 1994. A strategy for operational implementation of 4D-Var, using an incremental approach. *Q. J. R. Meteorol. Soc*., **120**, 1367–1387.

Derber, J., 1989. A variational continuous assimilation technique. *Mon. Weather Rev*., **117**, 2437–2446.

Egbert, G.D., A.F. Bennett and M.G.C. Foreman, 1994. Topex/Poseidon tides estimated using a global inverse model. *J. Geophys. Res*., **99**, 24,821–24,852.

Ehrendorfer, M. and R.M. Errico (eds.), 2007. *Meteorologische Zeitschrift*, **16**(6), 591–818.

Elbern, H., A. Strunk, H. Schmidt and O. Talagrand, 2007. Emission rate and chemical state estimation by 4-dimensional variational inversion. *Atmos. Chem. Phys*., **7**, 3749–3769.

Errera, Q. and D. Fonteyn, 2001. Four-dimensional variational chemical assimilation of CRISTA stratospheric measurements. *J. Geophys. Res*., **106**, 12,253–12,265.

Errico, R.M. and T. Vukisevic, 1992. Sensitivity analysis using an adjoint of the PSU-NCAR mesoscale model. *Mon. Weather Rev.*, **120**, 1644–1660.

Farrell, B.F., 1989. Optimal excitation of baroclinic waves. *J. Atmos. Sci.*, **46**, 1193–1206.

Fisher, M. and D.J. Lary, 1995. Lagrangian four-dimensional variational data assimilation of chemical species. *Q. J. R. Meteorol. Soc*., **121**, 1681–1704.

Fletcher, S.J., 2007. Implications and impacts of transforming lognormal variables into normal variables in VAR. *Meteorologische Zeitschrift*, **16**, 755–765.

Fletcher, S.J. and M. Zupanski, 2006. A hybrid normal and lognormal distribution for data assimilation. *Atmos. Sci. Lett*., **7**, 43–46.

Fournier, A., C. Eymin and T. Alboussière, 2007. A case for variational geomagnetic data assimilation: Insights from a one-dimensional, nonlinear, and sparsely observed MHD system. *Nonlinear Process Geophys*., **14**, 163–180.

Giering, R. and T. Kaminski, 1998. Recipes for adjoint code construction. *Trans. Math. Software*, **24**, 437–474.

Gustafsson, N., E. Källen and S. Thorsteinsson, 1998. Sensitivity of forecast errors to initial and lateral boundary conditions. *Tellus*, **50A**, 167–185.

Hall, M.C.G., D.G. Cacuci and M.E. Schlesinger, 1982. Sensitivity analysis of a radiative-convective model by the adjoint method. *J. Atmos. Sci*., **39**, 2038–2050.

Hascoët, L. and V. Pascual, 2004. *TAPENADE 2.1 user's guide*, available at the address http://www.inria.fr/rrrt/rt-0300.html

Hoffman, R.N., 1986. A four-dimensional analysis exactly satisfying equations of motion. *Mon. Weather Rev*., **114**, 388–397.

Järvinen, H., E. Andersson and F. Bouttier, 1999. Variational assimilation of time sequences of surface observations with serially correlated errors. *Tellus*, **51A**, 469–488.

Kalnay, E., 2002. *Atmospheric Modeling, Data Assimilation and Predictability*, Cambridge University Press, Cambridge, UK, 341 pp.

Klinker, E., F. Rabier, G. Kelly and J.-F. Mahfouf, 2000. The ECMWF operational implementation of four-dimensional variational assimilation. III: Experimental results and diagnostics with operational configuration. *Q. J. R. Meteorol. Soc.*, **126**, 1191–1215.

Lacarra, J.-F. and O. Talagrand, 1988. Short-range evolution of small perturbations in a barotropic model. *Tellus*, **40A**, 81–95.

Lahoz, W.A., A.J. Geer, S. Bekki, N. Bormann, S. Ceccherini, H. Elbern, Q. Errera, H.J. Eskes, D. Fonteyn, D.R. Jackson, B. Khattatov, M. Marchand, S. Massart, V.-H. Peuch, S. Rharmili, M. Ridolfi, A. Segers, O. Talagrand, H.E. Thornton, A.F. Vik and T. von Clarmann, 2007. The assimilation of Envisat data (ASSET) project. *Atmos. Chem. Phys*., **7**, 1773–1796.

Langland, R.H. and N.L. Baker, 2004. Estimation of observation impact using the NRL variational data assimilation adjoint system. *Tellus*, **56A**, 189–201.

Langland, R.H., R. Gelaro, G.D. Rohaly and M.A. Shapiro, 1999. Targeted observations in FASTEX: Adjoint-based targeting procedures and data impact experiments in IOP17 and IOP18. *Q. J. R. Meteorol. Soc*., **125**, 3241–3270.

Le Dimet, F.-X. and O. Talagrand, 1986. Variational algorithms for analysis and assimilation of meteorological observations: Theoretical aspects. *Tellus*, **38A**, 97–110.

Lewis, J.M. and J.C. Derber, 1985. The use of adjoint equations to solve a variational adjustment problem with advective constraints. *Tellus*, **37A**, 309–322.

Lions, J.-L., 1971. *Optimal Control of Systems Governed by Partial Differential Equations* (translated from the French). Springer, Berlin.

Liu, C., Q. Xiao and B. Wang, 2008. An ensemble-based four-dimensional variational data assimilation scheme. Part I: Technical formulation and preliminary test. *Mon. Weather Rev*., **136**, 3363–3373.

Lorenc, A.C. and F. Rawlins, 2005. Why does 4D-Var beat 3D-Var? *Q. J. R. Meteorol. Soc*., **131**, 3247–3257.

Lorenz, E.N., 1963. Deterministic nonperiodic flow. *J. Atmos. Sci*., **20**, 130–141.

Louvel, S., 2001. Implementation of a dual variational algorithm for assimilation of synthetic altimeter data in the oceanic primitive equation model MICOM. *J. Geophys. Res*., **106**, 9199–9212.

Marchuk, G.I., 1974. *Numerical Solution of the Problems of Dynamics of the Atmosphere and the Ocean* (in Russian). Gidrometeoizdat, Leningrad.

Marotzke, J., R. Giering, K.Q. Zhang, D. Stammer, C. Hill and T. Lee, 1999. Construction of the adjoint MIT ocean general circulation model and application to Atlantic heat transport sensitivity. *J. Geophys. Res.*, **104**, 29,529–29,547.

Moore, A.M. and B.F. Farrell, 1993. Rapid perturbation growth on spatially and temporally varying oceanic flows determined using an adjoint method: Application to the gulf stream. *J. Phys. Oceanogr.*, **23**, 1682–1702.

Mu, M., 2000. Nonlinear singular vectors and nonlinear singular values. *Sci. China* (ser. D), **43**, 375–385.

Mu, M., W.S. Duan and B. Wang, 2003. Conditional nonlinear optimal perturbation and its applications. *Nonlinear Process. Geophys*., **10**, 493–501.

Mu, M. and Z. Zhang, 2006. Conditional nonlinear optimal perturbations of a two-dimensional quasigeostrophic model. *J. Atmos. Sci*., **63**, 1587–1604.

Penenko, V.V. and N.N. Obraztsov, 1976. A variational initialization method for the fields of the meteorological elements (English translation). *Soviet Meteorol. Hydrol*., **11**, 1–11.

Pires, C., R. Vautard and O. Talagrand, 1996. On extending the limits of variational assimilation in nonlinear chaotic systems. *Tellus*, **48A**, 96–121.

Rabier, F., P. Courtier and O. Talagrand, 1992. An application of adjoint models to sensitivity analysis. *Beitr. Phys. Atmos*., **65**, 177–192.

Reichle, R.H., 2000. *Variational Assimilation of Remote Sensing Data for land Surface Hydrologic Applications*, Doctoral thesis, Massachusetts Institute of Technology, Cambridge, MA, 192 pp.

Ricci, S., A.T. Weaver, J. Vialard and P. Rogel, 2005. Incorporating state-dependent temperature-salinity constraints in the background error covariance of variational ocean data assimilation. *Mon. Weather Rev*., **133**, 317–338.

Rivière, O., G. Lapeyre and O. Talagrand, 2008. Nonlinear generalization of singular vectors: Behavior in a baroclinic unstable flow. *J. Atmos. Sci*., **65**, 1896–1911.

Sasaki, Y., 1970a. Some basic formalisms in numerical variational analysis. *Mon. Weather Rev*., **98**, 875–883.

Sasaki, Y., 1970b. Numerical variational analysis formulated from the constraints as determined by longwave equations and a low-pass filter. *Mon. Weather Rev*., **98**, 884–898.

Sasaki, Y., 1970c. Numerical variational analysis formulated with weak constraint and application to surface analysis of severe storm gust. *Mon. Weather Rev*., **98**, 899–910.

Sun, Z., A. Tangborn and W. Kuang, 2007. Data assimilation in a sparsely observed one-dimensional modeled MHD system. *Nonlinear Process. Geophys*., **14**, 181–192.

Swanson, K., R. Vautard and C. Pires, 1998. Four-dimensional variational assimilation and predictability in a quasi-geostrophic model. *Tellus*, **50A**, 369–390.

Talagrand, O., 1991. The use of adjoint equations in numerical modeling of the atmospheric circulation. In *Automatic Differentiation of Algorithms: Theory, Implementation, and Application*, Griewank, A. and G.F. Corliss (eds.), Society for Industrial and Applied Mathematics, Philadelphia, PA, pp 169–180.

Talagrand, O., 2003. Variational assimilation. Adjoint Equations. In *Data Assimilation for the Earth System*, NATO Science Series: IV. Earth and Environmental Sciences 26, Swinbank, R., V. Shutyaev and W.A. Lahoz (eds.), Kluwer Academic Publishers, Dordrecht, The Netherlands, pp 37–53, 378pp.

Talagrand, O. and P. Courtier, 1987. Variational assimilation of meteorological observations with the adjoint vorticity equation. I: Theory. *Q. J. R. Meteorol. Soc*., **113**, 1311–1328.

Thacker, W.C. and R.B. Long, 1988. Fitting dynamics to data. *J. Geophys. Res*., **93**, 1227–1240.

Thépaut, J.-N. and P. Courtier, 1991. Four-dimensional variational data assimilation using the adjoint of a multilevel primitive-equation model. *Q. J. R. Meteorol. Soc*., **117**, 1225–1254.

Thépaut, J.-N., R.N. Hoffman and P. Courtier, 1993. Interactions of dynamics and observations in a four-dimensional variational assimilation. *Mon. Weather Rev*., **121**, 3393–3414.

Trémolet, Y., 2006. Accounting for an imperfect model in 4D-Var. *Q. J. R. Meteorol. Soc*., **132**, 2483–2504.

Trémolet, Y., 2007. Incremental 4D-Var convergence study. *Tellus*, **59A**, 706–718.

Tromp, J., C. Tape and Q. Liu, 2005. Seismic tomography, adjoint methods, time reversal, and banana-donut kernels. *Geophys. J. Int*., **160**, 195–216.

Urban, B., 1993. A method to determine the theoretical maximum error growth in atmospheric models. *Tellus*, **45A**, 270–280.

Vialard, J., A.T. Weaver, D.L.T. Anderson and P. Delecluse, 2003. Three- and four-dimensional variational assimilation with a general circulation model of the tropical Pacific Ocean. Part II: Physical validation. *Mon. Weather Rev*., **131**, 1379–1395.

Vossepoel, F., A.T. Weaver, J. Vialard and P. Delecluse, 2004. Adjustment of near-equatorial wind stress with four-dimensional variational data assimilation in a model of the Pacific Ocean. *Mon. Weather Rev.*, **132**, 2070–2083.

Waelbroeck, C. and J.-F. Louis, 1995. Sensitivity analysis of a model of $CO_2$ exchange in tundra ecosystems by the adjoint method. *J. Geophys. Res.*, **100**, 2801–2816.

Weaver, A.T. and D.L.T. Anderson, 1997. Variational assimilation of altimeter data in a multilayer model of the tropical Pacific Ocean. *J. Phys. Oceanogr*., **27**, 664–682.

Zhang, Y., C.H. Bischof, R.C. Easter and P.-T. Wu, 1998. Sensitivity analysis of a mixed-phase chemical mechanism using automatic differentiation. *J. Geophys. Res.*, **103**, 18,953–18,979.

Zupanski, D., 1997. A general weak constraint applicable to operational 4D-VAR data assimilation systems. *Mon. Weather Rev*., **125**, 2274–2292.

# Ensemble Kalman Filter: Current Status and Potential

**Eugenia Kalnay**

## 1 Introduction

In this chapter we give an introduction to different types of Ensemble Kalman filter, describe the Local Ensemble Transform Kalman Filter (LETKF) as a representative prototype of these methods, and several examples of how advanced properties and applications that have been developed and explored for 4D-Var (four-dimensional variational assimilation) can be adapted to the LETKF without requiring an adjoint model. Although the Ensemble Kalman filter is less mature than 4D-Var (Kalnay 2003), its simplicity and its competitive performance with respect to 4D-Var suggest that it may become the method of choice.

The mathematical foundation of data assimilation is reviewed by Nichols (chapter *Mathematical Concepts of Data Assimilation*). Ide et al. (1997) concisely summarized the *sequential* and *variational* approaches in a paper introducing a widely used notation that we follow here, with bold low-case letters and bold capitals representing vectors and matrices, respectively. Non-linear operators are, however, represented in bold Kunster script (as in other chapters in this book). Since variational methods (chapter *Variational Assimilation*, Talagrand) and sequential methods basically solve the same problem (Lorenc 1986; Fisher et al. 2005) but make different approximations in order to become computationally feasible for large atmospheric and oceanic problems, it is particularly interesting to compare them whenever possible.

In this chapter we briefly review the most developed advanced sequential method, the Ensemble Kalman filter (EnKF) and several widely used formulations (Sect. 2). In Sect. 3 we compare the EnKF with the corresponding most advanced variational approach, 4D-Var (see chapter *Variational Assimilation*, Talagrand). Because 4D-Var has a longer history (e.g. Talagrand and Courtier 1987; Courtier and Talagrand 1990; Thépaut and Courtier 1991), and has been implemented in many operational centers (e.g. Rabier et al. 2000), there are many innovative ideas that have been

E. Kalnay (✉)
University of Maryland, College Park, MD 20742-2425, USA
e-mail: ekalnay@atmos.umd.edu

W. Lahoz et al. (eds.), *Data Assimilation*, DOI 10.1007/978-3-540-74703-1_4,

developed and explored in the context of 4D-Var, whereas the EnKF is a newer and less mature approach. We therefore present in Sect. 3 examples of how specific approaches explored in the context of 4D-Var can be simply adapted to the EnKF. These include the 4D-Var smoothing property that leads to a faster spin-up, the outer loop that increases the analysis accuracy in the presence of non-linear observation operators, the adjoint sensitivity of the forecasts to the observations, the use of lower resolution analysis grids, and the treatment of model errors. Section 4 is a summary and discussion.

## 2 Brief Review of Ensemble Kalman Filtering

The Kalman filter equations (Kalman 1960) are discussed by Nichols (chapter *Mathematical Concepts of Data Assimilation*, Sect. 3.1). Here we summarize key points of an alternative derivation of the Kalman filter equations for a linear perfect model due to Hunt et al. (2007) based on a maximum likelihood approach which provides additional insight about the role that the background term plays in the variational cost function (see Nichols, chapter *Mathematical Concepts of Data Assimilation*, Sect. 2; Talagrand, chapter *Variational Assimilation*, Sect. 2).

We start by assuming that the analysis $\bar{\mathbf{x}}^a_{n-1}$ valid at time $t_{n-1}$ has Gaussian errors with covariance $\mathbf{P}^a_{n-1}$ so that the likelihood of the true state $\mathbf{x}^t$ is

$$\rho(\mathbf{x}^t - \bar{\mathbf{x}}^a_{n-1}) \;\propto\; \exp\left\{-\frac{1}{2}(\mathbf{x}^t - \bar{\mathbf{x}}^a_{n-1})^T[\mathbf{P}^a_{n-1}]^{-1}(\mathbf{x}^t - \bar{\mathbf{x}}^a_{n-1})\right\},$$

where the overbar represents the expected value (cf. chapter *Mathematical Concepts of Data Assimilation*, Nichols, Sect. 2.4). The past observations $\mathbf{y}_j$ from time $t_1$ to $t_{n-1}$ (i.e. $j = 1, \ldots, n-1$) are also assumed to have a Gaussian distribution with error covariances $\mathbf{R}_j$, so that the likelihood of a trajectory of states $\{\mathbf{x}(t_j) | \, j = 1, \ldots, n-1\}$ given the past observations is proportional to

$$\prod_{j=1}^{n-1} \exp\left[-\frac{1}{2}(\mathbf{y}_j - \mathbf{H}_j\mathbf{x}(t_j))^T\mathbf{R}_j^{-1}(\mathbf{y}_j - \mathbf{H}_j\mathbf{x}(t_j))\right],$$

where $\mathbf{H}_j$ is the linear observation operator that transforms the model into the corresponding observation. To maximize the likelihood function, it is more convenient, however, to write the likelihood function as a function of the state at a single time rather than for the whole trajectory. Let $\mathbf{M}_{i,j}$ be the linear forecast model that advances a state from $\mathbf{x}(t_i)$ to $\mathbf{x}(t_j)$, we can then express the likelihood function as a function of the state $\mathbf{x}$ at a single time say $t_{n-1}$, as follows

$$\prod_{j=1}^{n-1} \exp\left[-\frac{1}{2}(\mathbf{y}_j - \mathbf{H}_j\mathbf{M}_{n-1,j}\,\mathbf{x}_{n-1})^T\mathbf{R}_j^{-1}(\mathbf{y}_j - \mathbf{H}_j\mathbf{M}_{n-1,j}\,\mathbf{x}_{n-1})\right].$$

Note that in this derivation we allow $t_j$ to be less than $t_{n-1}$, although integrating the model backward in time is problematic, it is used here only to derive the algorithm – in the end the algorithm will not require the backward integration of the model. Such an issue about time integration is found in the derivation of most Kalman smoother algorithms (see for instance Jazwinski 1970).

The analysis $\bar{\mathbf{x}}^a_{n-1}$ and its covariance $\mathbf{P}^a_{n-1}$ are the mean and covariance of a Gaussian probability distribution representing the relative likelihood of a state $\mathbf{x}_{n-1}$ given all previous observations, so that taking logarithms of the likelihoods, for some constant $c$,

$$\begin{aligned}\sum_{j=1}^{n-1}[\mathbf{y}^o_j - \mathbf{H}_j\mathbf{M}_{n-1,j}\,\mathbf{x}_{n-1}]^T\mathbf{R}_j^{-1}[\mathbf{y}^o_j - \mathbf{H}_j\mathbf{M}_{n-1,j}\,\mathbf{x}_{n-1}]\\ = [\mathbf{x}_{n-1} - \bar{\mathbf{x}}^a_{n-1}]^T(\mathbf{P}^a_{n-1})^{-1}[\mathbf{x}_{n-1} - \bar{\mathbf{x}}^a_{n-1}] + c\end{aligned} \tag{1}$$

The Kalman filter determines $\bar{\mathbf{x}}^a_n$ and $\mathbf{P}^a_n$ such that an equation analogous to Eq. (1) holds at time $t_n$. In the *forecast step* of the Kalman filter the analysis $\bar{\mathbf{x}}^a_n$ and its covariance are propagated to time $t_n$ with the linear forecast model $\mathbf{M}_{n-1,n}$ and its adjoint $\mathbf{M}^T_{n-1,n}$ creating the background state and its covariance:

$$\begin{aligned}\bar{\mathbf{x}}^b_n &= \mathbf{M}_{n-1,n}\,\bar{\mathbf{x}}^a_{n-1}\\ \mathbf{P}^b_n &= \mathbf{M}_{n-1,n}\,\mathbf{P}^a_n\,\mathbf{M}^T_{n-1,n}\end{aligned} \tag{2}$$

Propagating Eq. (1), using Eq. (2), we get a relationship valid for states at time $t_n$ (see Hunt et al. 2007 for further details), showing that the background term represents the Gaussian probability distribution of a state, given the past observations up to $t_{n-1}$:

$$\sum_{j=1}^{n-1}[\mathbf{y}^o_j - \mathbf{H}_j\mathbf{M}_{n,j}\,\mathbf{x}_n]^T\mathbf{R}_j^{-1}[\mathbf{y}^o_j - \mathbf{H}_j\mathbf{M}_{n,j}\,\mathbf{x}_n] = [\mathbf{x}_n - \bar{\mathbf{x}}^b_n]^T(\mathbf{P}^b_n)^{-1}[\mathbf{x}_n - \bar{\mathbf{x}}^b_n] + c \tag{3}$$

When the new observations at time $t_n$ are obtained, we use Eq. (3) to obtain an expression equivalent to Eq. (1) valid at time $t_n$, for another constant $c'$:

$$\begin{aligned}[\mathbf{x}_n - \bar{\mathbf{x}}^b_n]^T(\mathbf{P}^b_n)^{-1}[\mathbf{x}_n - \bar{\mathbf{x}}^b_n] \;+\; [\mathbf{y}^o_n - \mathbf{H}_n\mathbf{x}_n]^T\mathbf{R}_n^{-1}[\mathbf{y}^o_n - \mathbf{H}_n\mathbf{x}_n]\\ =\; [\mathbf{x}_n - \bar{\mathbf{x}}^a_n]^T(\mathbf{P}^a_n)^{-1}[\mathbf{x} - \bar{\mathbf{x}}^a_n] + c'\end{aligned} \tag{4}$$

The analysis state that minimizes the variational cost function

$$J(\mathbf{x}_n) = [\mathbf{x}_n - \bar{\mathbf{x}}^b_n]^T(\mathbf{P}^b_n)^{-1}[\mathbf{x}_n - \bar{\mathbf{x}}^b_n] + [\mathbf{y}^o_n - \mathbf{H}_n\mathbf{x}_n]^T\mathbf{R}_n^{-1}[\mathbf{y}^o_n - \mathbf{H}_n\mathbf{x}_n]$$

is the state with maximum likelihood given all the observations (cf. chapter *Mathematical Concepts of Data Assimilation*, Nichols, Sect. 2.4). Equation (3)

shows that in this cost function the background term represents the Gaussian distribution of a state with the maximum likelihood trajectory (history), i.e., $\bar{\mathbf{x}}_n^b$ is the analysis/forecast trajectory that best fits the past data available until $t_{n-1}$.

Equating the terms in Eq. (4) that are quadratic and linear in $\mathbf{x}$, the Kalman filter equations for the *analysis step* are obtained:

$$\mathbf{P}_n^a = \left[\left(\mathbf{P}_n^b\right)^{-1} + \mathbf{H}_n^T \mathbf{R}_n^{-1} \mathbf{H}_n\right]^{-1} = \left[\mathbf{I} + \mathbf{P}_n^b \mathbf{H}_n^T \mathbf{R}_n^{-1} \mathbf{H}_n\right]^{-1} \mathbf{P}_n^b \tag{5}$$

$$\bar{\mathbf{x}}_n^a = \mathbf{P}_n^a \left[\left(\mathbf{P}_n^b\right)^{-1} \bar{\mathbf{x}}_n^b + \mathbf{H}_n^T \mathbf{R}_n^{-1} \mathbf{y}_n^o\right] = \bar{\mathbf{x}}_n^b + \mathbf{P}_n^a \mathbf{H}_n^T \mathbf{R}_n^{-1} \left(\mathbf{y}_n^o - \mathbf{H}_n \bar{\mathbf{x}}_n^b\right) \tag{6}$$

The Remark 1 of Ide et al. (1997) "[In sequential methods] observations are processed whenever available and then discarded" follows from the fact that the background term is the most likely solution given all the past data, i.e., *if the Kalman filter has already spun-up from the initial conditions*, the observations are to be used only once (but see the discussion on spin-up in Sect. 3).

The Kalman gain matrix that multiplies the observational increment $\mathbf{y}_n^o - \mathbf{H}_n \bar{\mathbf{x}}_n^b$ in Eq. (6) can be written as

$$\mathbf{K}_n = \mathbf{P}_n^a \mathbf{H}_n^T \mathbf{R}_n^{-1} = \mathbf{P}_n^b \mathbf{H}_n^T \left(\mathbf{H}_n \mathbf{P}_n^b \mathbf{H}_n^T + \mathbf{R}_n\right)^{-1}.$$

For non-linear models $\mathcal{M}_{n-1,n}$, the *Extended Kalman filter* (EKF) approximation uses the non-linear model in the forecast step to advance the background state, but the covariance is advanced using the model linearized around the trajectory $\bar{\mathbf{x}}_n^b$, and its adjoint (e.g. Ghil and Malanotte-Rizzoli 1991; Nichols, chapter *Mathematical Concepts of Data Assimilation*, Nichols, Sect. 3):

$$\begin{aligned} \bar{\mathbf{x}}_n^b &= \mathcal{M}_{n-1,n}(\bar{\mathbf{x}}_{n-1}^a), \\ \mathbf{P}_n^b &= \mathbf{M}_{n-1,n} \mathbf{P}_{n-1}^a \mathbf{M}_{n-1,n}^T \end{aligned} \tag{7}$$

The cost of advancing the background error covariance with the linear tangent and adjoint models in Eq. (7) makes the EKF computationally unfeasible for any atmospheric model of realistic size without major simplifications.

Evensen (1994) suggested that Eq. (7) could be computed more efficiently with an *Ensemble Kalman filter* (EnKF) for non-linear models. The ensemble is created running $K$ forecasts, where the size of the forecast ensemble is much smaller than $n$, the dimension of the model, $K \ll n$. Then Eq. (7) can be replaced by

$$\begin{aligned} &\bar{\mathbf{x}}_n^b = \mathcal{M}_{n-1,n}(\bar{\mathbf{x}}_{n-1}^a), \bar{\mathbf{x}}_n^b = \frac{1}{K} \sum_{k=1}^{K} \mathbf{x}_{n,k}^b \\ &\mathbf{P}_n^b \approx \tfrac{1}{K-1} \sum_{k=1}^{K-1} (\mathbf{x}_{n,k}^b - \bar{\mathbf{x}}_n^b)(\mathbf{x}_{n,k}^b - \bar{\mathbf{x}}_n^b)^T \end{aligned} \tag{8}$$

where the overbar now represents the *ensemble average*.

Because the background error covariance is estimated from a relatively small ensemble, there are sampling errors at long distances, so that Houtekamer and Mitchell (2001) and Hamill et al. (2001) introduced the idea of *localizing* $\mathbf{P}_n{}^b$, i.e., multiplying each term of the covariance by an approximation of the Gaussian function $\exp(-r_{ij}^2/2L^2)$ (Gaspari and Cohn 1999). Here, $r_{ij}$ is the distance between two grid points $i,j$, and $L$ is the localization scale, so that the effect of localization is that long distance correlations are damped to zero. Mitchell et al. (2002) pointed out that this localization introduces imbalances in the analysis. Hunt (2005) and Miyoshi (2005) used an alternative localization multiplying the inverse of the observation error covariance $\mathbf{R}^{-1}$ by the Gaussian function, thus assuming that long distance observations have larger errors and reducing their impact on the grid point analyses. Because, unlike $\mathbf{P}_n^b$, $\mathbf{R}$ is generally either diagonal or block diagonal, this "observation localization" may be less prone to generate imbalances (Greybush et al. 2009).

There are two basic approaches to the EnKF, perturbed observations and square-root filters. In the *perturbed observations* EnKF, Burgers et al. (1998), Houtekamer and Mitchell (1998), Keppenne (2000), Keppenne and Rienecker (2002), Evensen and van Leeuwen (1996), Houtekamer et al. (2005) and others used ensembles of data assimilation systems with randomly perturbed observations (Evensen 2003). Perturbing the observations assimilated in different ensembles is required in this approach in order to avoid an underestimation of the size of the analysis error covariance, but it may introduce an additional source of sampling errors (Whitaker and Hamill 2002).

An alternative to the perturbed observations (or stochastic) approach are the *ensemble square-root filters* that generate an analysis ensemble mean and covariance satisfying the Kalman filter equations for linear models (Tippett et al. 2003; Bishop et al. 2001; Anderson 2001; Whitaker and Hamill 2002; Ott et al. 2004; Hunt et al. 2007). We will focus in the rest of the chapter on square-root (or deterministic) filters. Houtekamer and Mitchell (2001) pointed out that observations with uncorrelated errors can be assimilated serially (one at a time), with the background for a new observation being the analysis obtained when assimilating the previous observation. Tippett et al. (2003) discuss the differences between several square-root filters that derive computational efficiency by assimilating observations serially. Another Monte Carlo method that avoids using perturbed observations is described in Pham (2001).

Different square-root filters are possible because different analysis ensemble perturbations can have the same analysis error covariance. Of the three schemes discussed in Tippett et al. (2003), the Ensemble Adjustment Kalman Filter (EAKF) of Anderson (2001) has been implemented into the flexible Data Assimilation Research Testbed (DART) infrastructure and has been applied to many geophysical problems (http://www.image.ucar.edu/DAReS/Publications/). The square-root filter of Whitaker and Hamill (2002) results in simple scalar assimilation equations when observations are assimilated serially, and has also been adopted for a number of problems, such as the assimilation of surface observations (Whitaker et al. 2004), and for the regional EnKF of Torn and Hakim (2008) where only non-satellite data are assimilated. We note that the application of EnKF to regional models requires

including appropriate perturbations in the boundary conditions to avoid a reduction in variance in the interior (Nutter et al. 2004). Torn et al. (2006) showed that in the absence of a global EnKF system to provide consistent perturbed boundary conditions, several perturbation methods could give results comparable to those obtained with a global ensemble boundary conditions, thus making regional EnKF practically feasible for many groups without access to global EnKF. The third square-root filter discussed in Tippett et al. (2003) is the Ensemble Transform Kalman Filter, ETKF (Bishop et al. 2001), which introduced the computation of the analysis covariance by a transform method also adopted by Hunt et al. (2007). Zupanski (2005) proposed the Maximum Likelihood Ensemble Filter (MLEF) where a 4D-Var cost function with possibly non-linear observation operators is minimized within the subspace of the ensemble forecasts. In this system, the control forecast is allowed to have higher resolution than the rest of the ensemble. A review of EnKF methods is presented in Evensen (2003), and a comparison of EnKF with 4D-Var results for several models in Kalnay et al. (2007a).

Ott et al. (2002, 2004), and Hunt et al. (2007) developed an alternative type of square-root EnKF without perturbed observations by performing the analyses *locally in space*, as did Keppenne (2000). This is computationally efficient because the analyses at different grid points are independent and thus can be done in parallel. Since observations are assimilated simultaneously, not serially, it is simple to account for observation error correlations.

In this chapter we present results mostly based on the Local Ensemble Transform Kalman Filter (LETKF) as a representative prototype of EnKF. The LETKF algorithm is summarized below (see Hunt et al. 2007, for full details).

## *LETKF Algorithm*

This summary description is written as if all the observations are at the analysis time (i.e., for the 3D-LETKF), but the algorithm is essentially the same for the 4D-LETKF (Hunt et al. 2007). In 4D-LETKF (discussed below) the observations are in a time window that includes the analysis time and the non-linear observation operator $\mathcal{H}$ is evaluated at the observation time. $\mathcal{M}$ is the non-linear model forecast.

(a) LETKF *forecast step* (done globally) for each ensemble member $k$:

$$\mathbf{x}^b_{n,k} = \mathcal{M}_{n-1,n}(\mathbf{x}^a_{n-1,k})\,,\ k = 1,\ldots,K$$

(b) LETKF *analysis step* (at time $t_n$, so the subscript $n$ is dropped):

$$\mathbf{X}^b = \left[\mathbf{x}^b_1 - \bar{\mathbf{x}}^b,\ \ldots\ ,\mathbf{x}^b_K - \bar{\mathbf{x}}^b\right]\ ;\ \mathbf{P}^b = \mathbf{X}^b(\mathbf{X}^b)^T$$
$$\mathbf{y}^b_k = \mathcal{H}(\mathbf{x}^b_k)\ ;\ \mathbf{Y}^b = \left[\mathbf{y}^b_1 - \bar{\mathbf{y}}^b,\ldots,\mathbf{y}^b_K - \bar{\mathbf{y}}^b\right]$$

These computations can be done locally or globally, whichever is more efficient. Here the overbar represents the ensemble average.

*Localization:* choose for each grid point the observations to be used. Compute for each grid point the local analysis error covariance $\hat{\mathbf{P}}^a$ and analysis perturbations $\mathbf{W}^a$ in ensemble space:

$$\hat{\mathbf{P}}^a = \left[(K-1)\mathbf{I} + (\mathbf{Y}^b)^T\mathbf{R}^{-1}\mathbf{Y}^b\right]$$
$$\mathbf{W}^a = \left[(K-1)\hat{\mathbf{P}}^a\right]^{1/2}$$

The square-root required for the matrix of analysis perturbations in ensemble space is computed using the symmetric square root (Wang et al. 2004). This square-root has the advantage of having a zero mean and being closer to the identity than the square-root matrix obtained by Cholesky decomposition. As a result the analysis perturbations (chosen in different ways in different EnKF schemes) are also close to the background perturbations (Ott et al. 2002). Note that $\mathbf{W}^a$ can also be considered a *matrix of weights* since multiplying the forecast ensemble perturbations at each grid point by $\mathbf{W}^a$ gives the grid point analysis ensemble perturbations.

*Local analysis mean increment in ensemble space*:

$$\bar{\mathbf{w}}^a = \hat{\mathbf{P}}^a(\mathbf{Y}^b)^T\mathbf{R}^{-1}(\mathbf{y}^o - \bar{\mathbf{y}}^b)$$

Note that the forecast ensemble at each grid point multiplied by the *vector of weights* $\bar{\mathbf{w}}^a$ gives the grid point analysis $\bar{\mathbf{x}}^a$. The ensemble space analysis $\bar{\mathbf{w}}^a$ is added to each column of $\mathbf{W}^a$ to get the analysis ensemble in ensemble space: $\mathbf{W}^a \leftarrow \mathbf{W}^a \oplus \bar{\mathbf{w}}^a$

The new ensemble analyses are the $K$ columns of

$$\mathbf{X}^a = \mathbf{X}^b\mathbf{W}^a + \bar{\mathbf{x}}^b$$

*Global analysis ensemble:* The analysis ensemble columns for each grid point are gathered together to form the new global analysis ensemble $\mathbf{X}^a_{n,k}$, and the analysis cycle can proceed.

## 3 Adaptation of 4D-Var Techniques into EnKF

4D-Var and EnKF are essentially solving the same problem since they minimize the same cost function in Eq. (2) using different computational methods. These differences lead to several advantages and disadvantages for each of the two methods (see, for example, Lorenc 2003; Table 7 of Kalnay et al. 2007a; discussion of Gustafsson 2007; response of Kalnay et al. 2007b).

A major difference between 4D-Var and the EnKF is the dimension of the subspace of the analysis increments (analysis minus background). 4D-Var corrects the background forecast in a subspace that has the dimension of the linear tangent and the adjoint models used in the minimization algorithm, and this subspace is generally much larger than the local subspace of corrections in the EnKF of dimension

$K-1$ determined by the ensemble size. It would be impractical to try to overcome this apparent EnKF disadvantage by using a very large ensemble size. Fortunately, the localization of the error covariances carried out in the EnKF in order to reduce long distance covariance sampling errors, substantially addresses this problem by greatly increasing the number of degrees of freedom available to fit the data. As a result, experience has been that the quality of the EnKF analyses with localization increases with the number of ensemble members, but that there is little further improvement when the size of the ensemble is increased beyond about 100. The observation that 50–100 ensemble members are sufficient for the EnKF seems to hold for atmospheric problems ranging from the storm-scales and mesoscales to the global-scales (Fuqing Zhang, personal communication).

There are a number of attractive properties of 4D-Var developed over the years. They include the ability to assimilate observations at their right time (Talagrand and Courtier 1987); the fact that within the data assimilation window 4D-Var acts as a smoother (Thépaut and Courtier 1991); the availability of an adjoint model allowing the estimation of the impact of observations on the analysis (Cardinali et al. 2004) and on the forecasts (Langland and Baker 2004); the ability to use long assimilation windows (Pires et al. 1996); the computation of outer loops correcting the background state when computing non-linear observation operators and the ability to use a lower resolution simplified model in the inner loop (see Fig. 3 discussed later); and the possibility of accounting for model errors by using the model as a weak constraint (Trémolet 2007). In the rest of this section we discuss how these advantageous methods that have been developed and implemented for 4D-Var systems can also be adapted and used in the LETKF, a prototype of EnKF.

## *3.1 4D-LETKF and No-Cost Smoother*

Hunt et al. (2004) developed an extension of the Local Ensemble Kalman Filter (LEKF; Ott et al. 2004) to four dimensions (4-D), taking advantage of the fact that the observational increments are expressed as linear combinations (weights) of the forecast ensemble perturbations at the time of the observation. [1] This allows using the same coefficients to "transport" the observational increments either forward or backward in time to the time of the analysis. We note that within this 4-D formulation it is possible to account for observation errors correlated in time, as Järvinen et al. (1999) have done within 4D-Var. Hunt et al. (2007) showed that the 4-D extension is particularly simple within the LETKF framework, requiring the concatenation of observations performed at different times within the assimilation window into the vectors $\mathbf{y}^o, \bar{\mathbf{y}}^b$ and the vertical columns of $\mathbf{Y}^b$ and of a block error

[1] Strictly speaking the combinations are not linear since the weights depend on the forecasts (Nerger et al. 2005).

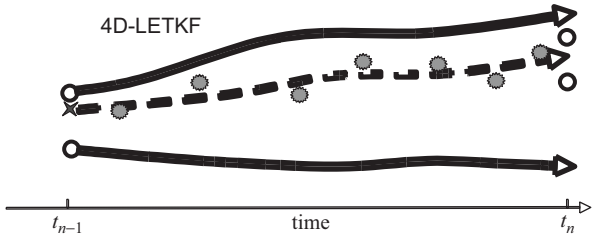



**Fig. 1** Schematic showing that the 4D-LETKF finds the linear combination of the ensemble forecasts at $t_n$ that best fits the observations *throughout* the assimilation window $t_{n-1} - t_n$. The *white circles* represent the ensemble of analyses (whose mean is the analysis $\bar{\mathbf{x}}^a$), the *full lines* represent the ensemble forecasts, the *dashed line* represents the linear combination of the forecasts whose final state is the analysis, and the *grey stars* represent the asynchronous observations. The cross at the initial time of the assimilation window $t_{n-1}$ is a *no-cost Kalman smoother*, i.e., an analysis at $t_{n-1}$ improved using the information of "future" observations within the assimilation window by weighting the ensembles at $t_{n-1}$ with the weights obtained at $t_n$. The smoothed analysis ensemble at $t_{n-1}$ (not shown in the schematic) can also be obtained at no cost using the same linear combination of the ensemble forecasts valid at $t_n$ given by $\mathbf{W}^a$. Adapted from Kalnay et al. (2007b)

covariance $\mathbf{R}$ with blocks corresponding to the same observations. Note that 4D-LETKF *determines the linear combination of ensemble forecasts valid at the end of the assimilation window that best fits the data throughout the assimilation window*.

This property allows creating a "*cost-free*" smoother for the LETKF with analogous smoothing properties as 4D-Var (Fig. 1): the same weighted combination of the forecasts with weights given by the vector $\bar{\mathbf{w}}^a$ is valid at any time of the assimilation interval. It provides a smoothed analysis mean that (as in 4D-Var) is more accurate than the original analysis because it uses the future data available within the assimilation window (Kalnay et al. 2007b; Yang et al. 2009a). As in 4D-Var, the smoothed analysis at the beginning of the assimilation window is an improvement over the filtered analysis computed using only past data. At the end of the assimilation interval only past data is used so that (as in 4D-Var) the smoother coincides with the analysis obtained with the filter.

It should be noted that in the same way we can use the weights $\bar{\mathbf{w}}^a$ to provide a mean smoother solution as a function of time, we can use the matrix $\mathbf{W}^a$ and apply it to the forecast perturbations $\mathbf{X}^b\mathbf{W}^a$ to provide an associated uncertainty evolving with time (Ross Hoffman, personal communication). The updating of the uncertainty is critical for the "Running in Place" method described next, but the uncertainty is not updated in the "outer loop" approach.

## *3.2 Application of the No-Cost Smoother to the Acceleration of the Spin-Up*

4D-Var has been observed to spin up faster than EnKF (e.g. Caya et al. 2005), presumably because of its smoothing properties that allow finding the initial conditions

at the beginning of the assimilation window that will best fit all the observations. The fact that we can compute a no-cost smoother allows the development of a new algorithm, called *Running in Place* by Kalnay and Yang (2008) that should be useful in rapidly evolving situations. For example, at the time radar measurements first detect the development of a severe storm, the available EnKF estimate of the atmospheric state and its uncertainty are no longer very useful. In other words, *while formally the EnKF members and their average are still the most likely state and best estimate of the uncertainty given all the past data, these EnKF estimates are no longer likely at all*. At the start of severe storm convection, the dynamics of the system changes substantially, and the statistics of the processes become non-stationary. In this case, as in the spin-up case in which there are no previous observations available, the running in place algorithm ignores the rule "use the data and then discard it" and repeatedly recycles the new observations.

*Running in place algorithm:* This algorithm is applied to each assimilation window during the spin-up phase. The LETKF is "cold-started" with any initial ensemble mean and perturbations at $t_0$. The "running in place" loop at time $t_n$ (initially $t_0$) is as follows:

1. Integrate the ensemble from $t_n$ to $t_{n+1}$, perform a standard LETKF analysis and obtain the analysis weights for the interval $[t_n, t_{n+1}]$, saving the mean square observations minus forecast (OMF) computed by the LETKF;
2. Apply the no-cost smoother to obtain the smoothed analysis ensemble at $t_n$ by using these weights;
3. Perturb the smoothed analysis ensemble with small zero-mean random Gaussian perturbations, a method similar to additive inflation. Typically, the perturbations have an amplitude equal to a small percentage of the climate variance;
4. Integrate the perturbed smoothed ensemble to $t_{n+1}$. While the forecast fit to the observations continues to improve according to a criterion such as

$$\frac{\mathrm{OMF}^2(\mathrm{iter}) - \mathrm{OMF}^2(\mathrm{iter}+1)}{\mathrm{OMF}^2(\mathrm{iter})} > \varepsilon,$$

   go to step 2 and perform another iteration. If not, replace $t_n$ with $t_{n+1}$ and go to step 1;
5. If no additional iteration beyond the first one is needed, the running in place analysis is the same as the standard EnKF. When the system converges, no additional iterations are needed, so that if several assimilation cycles take place without invoking a second iteration, the running in place algorithm can be switched off and the system returns to a normal EnKF.

The purpose of adding perturbations in step 3 is twofold: it avoids reaching the same analysis as in the previous iteration, and it increases the chances that the

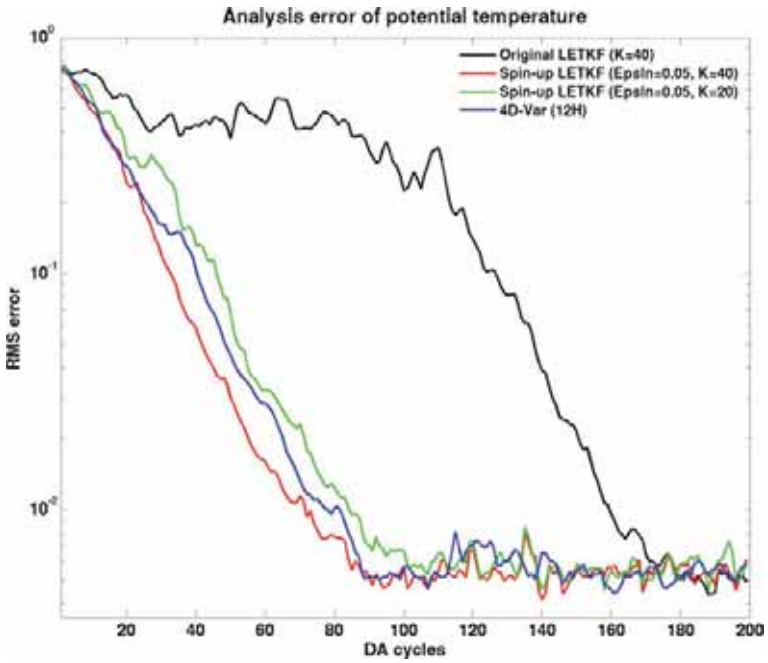



**Fig. 2** Comparison of the spin-up of a quasi-geostrophic model simulated data assimilation when starting from random initial conditions. Observations (simulated radiosondes) are available every 12 h, and the analysis root-mean-square (RMS) errors are computed by comparing with a nature run (see the chapter *Observing System Simulation Experiments*, Masutani et al.). *Black line*: original LETKF with 40 ensemble members, and no prior statistical information, *blue line*: optimized 4D-Var, *red line*: LETKF "running in place" with $\varepsilon = 5\%$ and 40 ensemble members, *green line*: as the *red line* but with 20 ensemble members

ensemble will explore unstable directions of error growth missed by the unperturbed ensemble and not be "trapped" in the "unlikely" subspace of the initial perturbations.

Running in place was tested with the LETKF in the quasi-geostrophic, QG, model of Rotunno and Bao (1996) (Fig. 2 adapted from Kalnay and Yang 2008). When starting from a 3D-Var (three dimensional variational) analysis mean, the LETKF converges quickly (not shown), but from random initial states it takes 120 cycles (60 days) to reach a point in which the ensemble perturbations represent the "errors of the day" (black line in Fig. 2). From then on the ensemble converges quickly, in about 60 more cycles (180 cycles total).

By contrast, the 4D-Var started from the same initial mean state, but using as background error covariance the 3D-Var **B** scaled down with an optimal factor, converges twice as fast, in about 90 cycles (blue line in Fig. 2). The running in place algorithm with $\varepsilon = 5\%$ (red line) converges about as fast as 4D-Var, and it only takes about 2 iterations per cycle (i.e., one additional assimilation for each window). The green line is also for $\varepsilon = 5\%$, but with $K = 20$ ensemble members, not $K = 40$ as

used in the other experiments and also gives good results, but experiments with $K = 10$ failed to spin-up faster with this technique. With $\varepsilon = 1\%$ (not shown) the initial convergence (in real time) is faster, but it requires about 5 times more iterations. It is interesting that when the number of iterations is *fixed* to 10 (not shown), the data are over-fitted so that the system quickly converges to a final level of error about twice as large than when the iterations are chosen adaptively.

### *3.3 "Outer Loop" and Dealing with Non-linear Ensemble Perturbations*

A disadvantage of the EnKF is that the Kalman filter equations used in the analysis assume that the ensemble perturbations are Gaussian, so that when windows are relatively long and perturbations become non-linear, this assumption breaks down and the EnKF is not optimal (Harlim and Hunt 2007a, b). By contrast, 4D-Var is recomputed within an assimilation window until the initial conditions that minimize the cost function for the non-linear model integration in that window are found. In many operational centres (e.g. the National Centers for Environmental Prediction, NCEP, and the European Centre for Medium-Range Weather Forecasts, ECMWF) the minimization of the 3D-Var or 4D-Var cost function is done with a linear "inner loop" that improves the initial conditions minimizing a cost function that is quadratic in the perturbations. In the 4D-Var "outer loop" the non-linear model is integrated from the initial state improved by the inner loop and the linearized observational increments are recomputed for the next inner loop (Fig. 3).

The ability of including an outer loop increases significantly the accuracy of both 3D-Var and 4D-Var analyses (Arlindo da Silva, personal communication), so that it would be important to develop the ability to carry out an equivalent "outer loop" in the LETKF. This can be done by considering the LETKF analysis for a window as an "inner loop" and, using the no-cost smoother, adapting the 4D-Var outer loop algorithm to the EnKF. As in 4D-Var, we introduce into the LETKF the freedom of the inner loop to improve the initial analysis (i.e., the mean of the ensemble) but keep constant the background error covariance, given by the ensemble initial perturbations. This re-centres the initial ensemble forecasts about the value improved by the inner loop, and another "outer loop" with full non-linear integrations can be carried out. [2] Note that this algorithm is identical to "running in place", except that only the mean is updated, not the perturbations about the mean at $t_n$.

This algorithm for an outer loop within the EnKF was tested with the Lorenz (1963) model for which comparisons between LETKF and 4D-Var were made, optimizing simultaneously the background error covariance and the length of the window for 4D-Var (Kalnay et al. 2007a). For short assimilation windows, the 3-member LETKF gives analysis errors similar or smaller than 4D-Var, but with

[2] Takemasa Miyoshi (personal communication) has pointed out that Jazwinski (1970) proposed the same "outer loop" algorithm for Extended Kalman filter (see footnote on page 276).

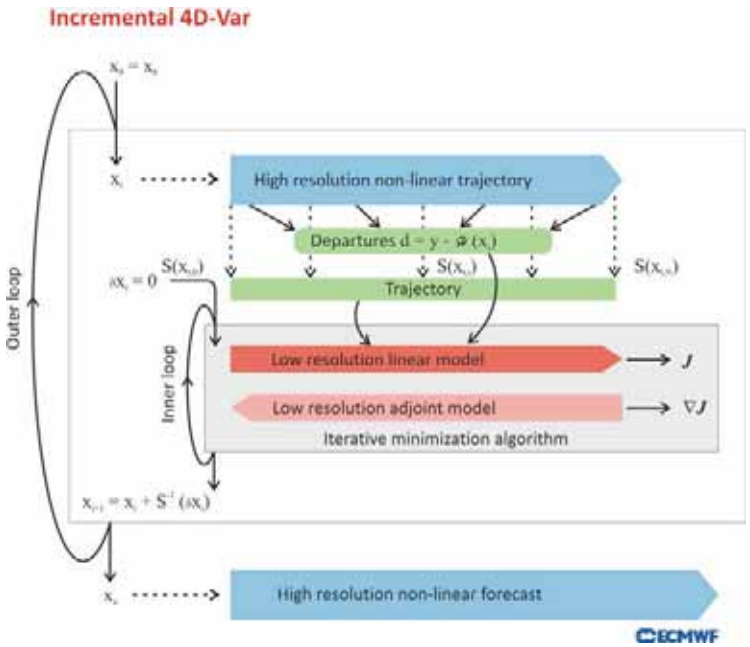



**Fig. 3** Schematic of how the 4D-Var cost function is minimized in the ECMWF system. From Yannick Trémolet, August 2007 class on Incremental 4D-Var at University of Maryland summer Workshop on Applications of Remotely sensed data to Data Assimilation

long assimilation windows of 25 steps, when the perturbations grow non-linearly, Kalnay et al. (2007a) were not able to find an LETKF configuration competitive with 4D-Var. However, as shown in Table 1 below, the LETKF with an outer loop is able to beat 4D-Var. We note that "running in place" (with up to one additional

**Table 1** Comparison of the RMSE (RMS error, non-dimensional units) for 4D-Var and LETKF for the Lorenz (1963) 3-variable model. 4D-Var has been simultaneously optimized for the window length (Kalnay et al. 2007a; Pires et al. 1996) and the background error covariance, and the full non-linear model is used in the minimization. LETKF is performed with 3 ensemble members (no localization is needed for this problem), and inflation is optimized. For the 25 steps case, "running in place" further reduces the error to about 0.39

| Experiment details | 4D-Var | LETKF (inflation factor) | LETKF with less than 3 "outer loop" iterations (inflation factor) |
|---|---|---|---|
| Window = 8 steps (perturbations are approximately linear) | 0.31 | 0.30 (1.05) | 0.27 (1.04) |
| Window = 25 steps (perturbations are non-linear) | 0.53 | 0.66 (1.28) | 0.48 (1.08) |

analysis) can further improve the results for the case of 25 steps, reducing the RMS (root-mean-square) analysis error of 0.48 obtained using the outer loop to about 0.39, with inflation of 1.05. As in the case of the spin-up, this re-use of observations is justified by the fact that for long windows and non-linear perturbations, the background ensemble ceases to be Gaussian, and the assumption of statistical stationarity is no longer viable.

These experiments suggest that it should be possible to deal with non-linearities and obtain results comparable or better than 4D-Var by methods such as an outer loop and running in place.

### *3.4 Adjoint Forecast Sensitivity to Observations Without Adjoint Model*

Langland and Baker (2004) proposed an adjoint-based procedure to assess the observation impact on short-range forecasts without carrying out data-denial experiments. This adjoint-based procedure can evaluate the impact of any or all observations assimilated in the data assimilation and forecast system on a selected measure of short-range forecast error. In addition, it can be used as a diagnostic tool to monitor the quality of observations, showing which observations make the forecast worse, and can also give an estimate of the relative importance of observations from different sources. Following a similar procedure, Zhu and Gelaro (2008) showed that this adjoint-based method provides accurate assessments of the forecast sensitivity with respect to most of the observations assimilated. Unfortunately, this powerful and efficient method to estimate observation impact requires the adjoint of the forecast model which is complicated to develop and not always available, as well as the adjoint of the data assimilation algorithm.

Liu and Kalnay (2008) proposed an ensemble-based sensitivity method able to assess the same forecast sensitivity to observations as in Langland and Baker (2004), but without adjoint. Following Langland and Baker (2004), they define a cost function $\Delta \mathbf{e}_t^2 = (\mathbf{e}_{t|0}^T\, \mathbf{e}_{t|0} - \mathbf{e}_{t|-6}^T\, \mathbf{e}_{t|-6})$ that measures the forecast sensitivity at time $t$ of the observations assimilated at time 0. Here $\mathbf{e}_{t|0} = \bar{\mathbf{x}}_{t|0}^f - \bar{\mathbf{x}}_t^a$ is the perceived error of the forecast started from the analysis at $t = 0$, verified against the analysis valid at time $t$, and $\mathbf{e}_{t|-6} = \bar{\mathbf{x}}_{t|-6}^f - \bar{\mathbf{x}}_t^a$ is the corresponding error of the forecast starting from the previous analysis at $t = -6$ h. The difference between the two forecasts is only due to the observations $\mathbf{y}_0^o$ assimilated at $t = 0$: $\bar{\mathbf{x}}_0^a - \bar{\mathbf{x}}_{0|-6}^b = \mathbf{K}(\mathbf{y}_0^o - \mathcal{H}(\bar{\mathbf{x}}_{0|-6}^b))$, where $\mathbf{K}$ is the gain matrix of the data assimilation system. There is a slight error in Eq. (10) in Liu and Kalnay (2008) so that we re-derive here the forecast sensitivity equation (Hong Li, personal communication):

$$\Delta \mathbf{e}_t^2 = (\mathbf{e}_{t|0}^T \mathbf{e}_{t|0} - \mathbf{e}_{t|-6}^T \mathbf{e}_{t|-6}) = (\mathbf{e}_{t|0}^T - \mathbf{e}_{t|-6}^T)(\mathbf{e}_{t|0} + \mathbf{e}_{t|-6}) = (\bar{\mathbf{x}}_{t|0}^f - \bar{\mathbf{x}}_{t|-6}^f)^T (\mathbf{e}_{t|0} + \mathbf{e}_{t|-6})$$

$$\Delta \mathbf{e}_t^2 \approx \left[\mathbf{M}(\bar{\mathbf{x}}_0^a - \bar{\mathbf{x}}_{0|-6}^b)\right]^T (\mathbf{e}_{t|0} + \mathbf{e}_{t|-6}) = \left[\mathbf{MK}(\mathbf{y} - \mathcal{H}(\mathbf{x}_{0|-6}^b))\right]^T (\mathbf{e}_{t|0} + \mathbf{e}_{t|-6})$$

where $\mathbf{M}$ is the linear tangent forecast model that advances a perturbation from 0-h to time $t$.

Langland and Baker (2004) compute this error sensitivity by using the adjoint of the model and of the data assimilation scheme:

$$\Delta \mathbf{e}^2_{t,LB} = \left[\mathbf{y} - \mathcal{H}(x^b_{0|-6})\right]^T \mathbf{K}^T \mathbf{M}^T (\mathbf{e}_{t|0} + \mathbf{e}_{t|-6})$$

In the EnKF we can take advantage of the fact that the Kalman gain is computed as $\mathbf{K} = \mathbf{P}^a \mathbf{H}^T \mathbf{R}^{-1} = (K-1)^{-1} \mathbf{X}^a (\mathbf{X}^a)^T \mathbf{H}^T \mathbf{R}^{-1}$, so that $\mathbf{MK} = \mathbf{M}\mathbf{X}^a (\mathbf{X}^a)^T \mathbf{H}^T \mathbf{R}^{-1}/(K-1) \approx \mathbf{X}^f_{t|0} (\mathbf{Y}^a)^T \mathbf{R}^{-1}/(K-1)$, with $\mathbf{X}^f_{t|0}$, the forecast differences at time $t$ computed non-linearly rather than with the linear tangent model. As a result, for EnKF the forecast sensitivity is computed as

$$\Delta \mathbf{e}^2_{t,EnKF} = \left[\mathbf{y} - \mathcal{H}(x^b_{0|-6})\right]^T \mathbf{R}^{-1} \mathbf{Y}^a (\mathbf{X}^f_{t|0})^T (\mathbf{e}_{t|0} + \mathbf{e}_{t|-6})/(K-1)$$

Because the forecast perturbation matrix $\mathbf{X}^f_{t|0}$ is computed non-linearly, the forecast sensitivity and the ability to detect bad observations remains valid even for forecasts longer than 24 h, for which the adjoint sensitivity based on the adjoint model $\mathbf{M}^T$ ceases to be accurate. As in Langland and Baker (2004) and Zhu and Gelaro (2008), it is possible to split the vector of observational increments $\mathbf{y} - \mathcal{H}(\mathbf{x}^b_{0|-6})$ into any subset of observations and obtain the corresponding forecast sensitivity.

Figure 4 shows the result of applying this method to the Lorenz (1996) 40-variables model. In this case observations were made at every point every 6 h, created from a "nature" run by adding Gaussian observational errors of mean zero

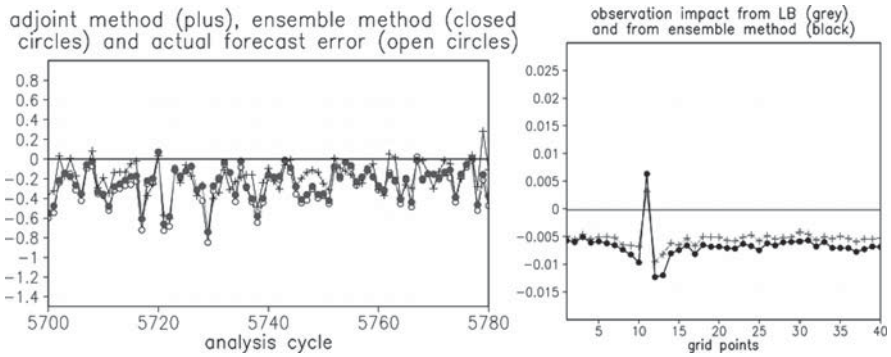



**Fig. 4** *Left*: Domain average variability in the forecast impact estimated by the adjoint method (*plus symbols*), the EnKF sensitivity (*closed circles*) and the actual forecast sensitivity. *Right*: Time average (over the last 7,000 analysis cycles) of the contribution to the reduction of the 1-day forecast errors from each observation location. The observation at the 11th grid point has $\sigma^o = 8$ random errors rather than the specified value of 0.2. Adjoint sensitivity (*grey plus*), EnKF sensitivity (*black*). Adapted from Liu and Kalnay (2008)

and standard deviation 0.2. The left panel shows that both the adjoint and the EnKF sensitivity methods are able to estimate quite accurately the day-to-day variability in the 24-h forecast sensitivity to the observations when all the observations have similar Gaussian errors. A "bad station" was then simulated at grid point 11 by increasing the standard deviation of the errors to 0.8 without "telling" the data assimilation system about the observation problems in this location. The right panel of Fig. 4 shows the time average of the forecast sensitivity for this case, indicating that both the adjoint and the ensemble-based sensitivity are able to identify that the observations at grid point 11 have a deleterious impact on the forecast. They both show that the neighbouring points improved the forecasts more than average by partially correcting the effects of the 11th point observations.

The cost function in this example was based on the Eulerian norm, appropriate for a univariate problem, but the method can be easily extended to an energy norm, allowing the comparison of the impact of winds and temperature observations on the forecasts. Although for short (1-day) forecasts the adjoint and ensemble sensitivities have similar performances (Fig. 4), the (linear) adjoint sensitivity ceases to identify the wrong observations if the forecasts are 2-days or longer. The ensemble sensitivity, which is based on non-linear integrations, continues to identify the observations having a negative impact even on long forecasts (not shown).

We note that Liu et al. (2009) also formulated the sensitivity of the analysis to the observations as in Cardinali et al. (2004) and showed that it provides a good qualitative estimate of the impact of adding or denying observations on the analysis error, without the need to run these costly experiments. Since the Kalman gain matrix is available in ensemble space, complete cross-validations of each observation can be computed exactly within the LETKF without repeating the analysis.

### *3.5 Use of a Lower Resolution Analysis*

The inner/outer loop used in 4D-Var was introduced in Sect. 3.3 above, where we showed that a similar outer loop can be carried out in EnKF. We now point out that it is common practice to compute the inner loop minimization, shown schematically in Fig. 3, using a simplified model (Lorenc 2003), which usually has lower resolution and simpler physics than the full resolution model used for the non-linear outer loop integration. The low-resolution analysis correction computed in the inner loop is interpolated back to the full resolution model (Fig. 3). The use of lower resolution in the minimization algorithm of the inner loop results in substantial savings in computational cost compared with a full resolution minimization, but it also degrades the analysis.

Yang et al. (2009b) took advantage that in the LETKF the analysis ensemble members are a weighted combination of the forecasts, and that the analysis weights $\mathbf{W}^a$ are much smoother (they vary on a much larger scale) than the analysis increments or the analysis fields themselves. They tested the idea of interpolating the weights but using the full resolution forecast model on the same quasi-geostrophic model discussed before. They performed full resolution analyses and compared the

results with a computation of the LETKF analysis (i.e., the weight matrix $\mathbf{W}^a$) on coarser grids, every $3 \times 3$, $5 \times 5$ and $7 \times 7$ grid points, corresponding to an analysis grid coverage of 11, 4 and 2%, respectively, as well as interpolating the analysis increments. They found that interpolating the weights did not degrade the analysis compared with the full resolution, whereas interpolating the analysis increments resulted in a serious degradation (Fig. 5).

The use of a symmetric square-root in the LETKF ensures that the interpolated analysis has the same linear conservation properties as the full resolution analysis. The results suggest that interpolating the analysis weights computed on a coarse grid without degrading the analysis can substantially reduce the computational cost of the LETKF. Although the full resolution ensemble forecasts are still required, they are also needed for ensemble forecasting in operational centres.

We note that the fact that the weights vary on large scales, and that the use of a coarser analyses with weight interpolation actually improves slightly the analysis in data sparse regions, suggest that smoothing the weights is a good approach to filling data gaps such as those that appear in between satellite orbits (Yang et al. 2009b; Lars Nerger, personal communication). Smoothing the weights, both in the horizontal and in the vertical may also reduce sampling errors and increase the accuracy of the EnKF analyses.

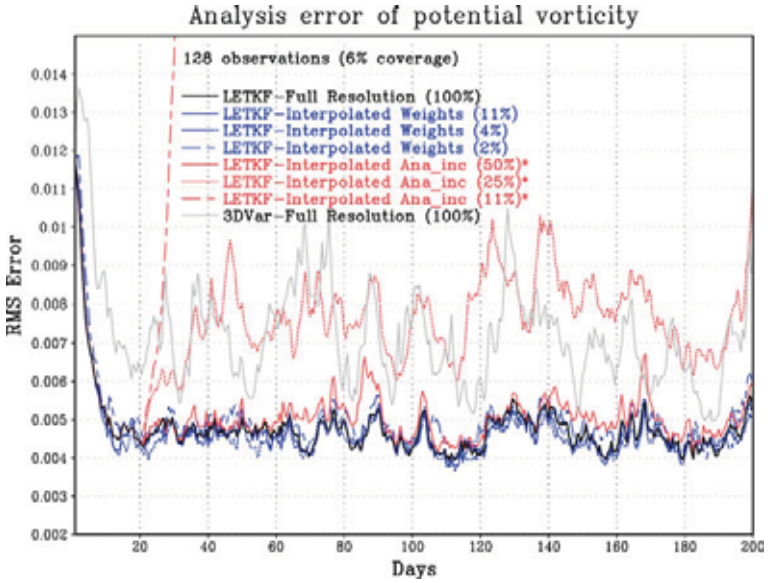



**Fig. 5** The time series of the RMS analysis error in terms of the potential vorticity from different data assimilation experiments. The LETKF analysis from the full-resolution is denoted as the *black line* and the 3D-Var derived at the same resolution is denoted as the *grey line*. The LETKF analyses derived from weight-interpolation with different analysis coverage are indicated with *blue lines*. The LETKF analyses derived after the first 20 days from increment-interpolation with different analysis coverage are indicated with the *red lines*. Adapted from Yang et al. (2009b)

### *3.6 Model and Observational Error*

Model error can seriously affect the EnKF because, among other reasons, the presence of model biases cannot be detected by the EnKF original formulation, and the ensemble spread is the same with or without model bias (Li 2007). For this reason, the most widely used method for imperfect models is to increase the multiplicative or additive inflation (e.g. Whitaker et al. 2008). Model biases can also be taken into account by estimating the bias as in Dee and da Silva (1998) or its simplified approximation (Radakovich et al. 2001) – see also chapter *Bias Estimation* (Ménard). More recently, Baek et al. (2006) pointed out that model bias could be estimated accurately augmenting the model state with the bias, and allowing the error covariance to eventually correct the bias. Because in this study the bias was assumed to be a full resolution field, this required doubling the number of ensemble members in order to reach convergence.

In the standard 4D-Var, the impact of model bias cannot be neglected within longer windows because the model (assumed to be perfect) is used as a strong constraint in the minimization (e.g. Andersson et al. 2005). Trémolet (2007) has developed several techniques allowing for the model to be a weak constraint in order to estimate and correct model errors. Although the results are promising, the methodology for the weak constraint is complex, and still under development.

Li (2007) and Li et al. (2009a) compared several methods to deal with model bias (Fig. 6), including a "Low-dimensional" method based on an independent estimation of the bias from averages of 6-h estimated forecast errors started from a reanalysis (or any other available good quality analysis). This method was applied to the SPEEDY (Simplified Parameterizations primitivE-Equation DYnamics) model (Molteni 2003) assimilating simulated observations from the NCEP-NCAR (National Centers for Environmental Prediction-National Center for Atmospheric Research) reanalysis, and it was found to be able not only to estimate the bias, but also the errors in the diurnal cycle and the model forecast errors linearly dependent on the state of the model (Danforth et al. 2006; Danforth and Kalnay 2008).

The results obtained by Li (2009a) accounting for model errors within the LETKF, presented in Fig. 6, indicate that: (a) additive inflation is slightly better than multiplicative inflation; and (b) methods to estimate and correct model bias (e.g. Dee and da Silva 1998; Danforth et al. 2006) should be combined with inflation, which is more appropriate in correcting random model errors. The combination of the low-dimensional method with additive inflation gave the best results, and was substantially better than the results obtained assuming a perfect model (Fig. 6).

We note that the approach of Baek et al. (2006) of correcting model bias by augmenting the state vector with the bias can be used to determine other parameters, such as surface fluxes, observational bias, nudging coefficients, etc. It is similar to increasing the control vector in the variational approach, and is only limited by the number of degrees of freedom that are added to the control vector and sampling errors in the augmented background error covariance.

With respect to observation error estimations, Desroziers et al. (2005) derived statistical relationships between products of observations minus background,

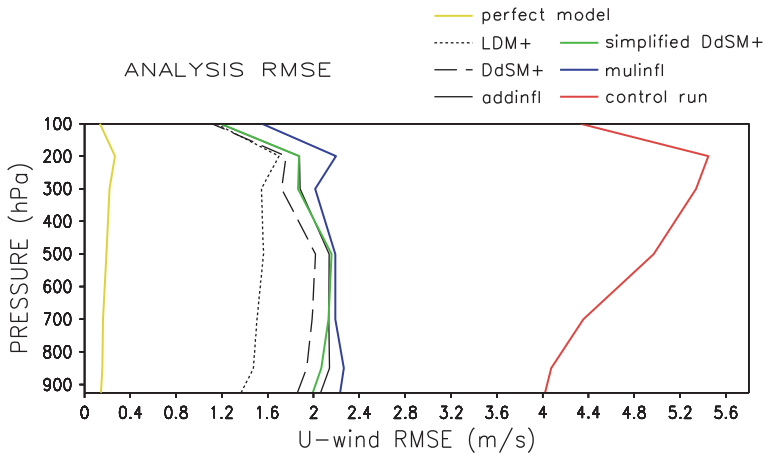



**Fig. 6** Comparison of the analysis error averaged over 2 months for the zonal velocity in the SPEEDY model for several simulations with the radiosonde observations available every other point. The *yellow line* corresponds to a perfect model simulation with the observations extracted from a SPEEDY model "nature run" (see chapter *Observing System Simulation Experiments*, Masutani et al.). The *red* is the control run, in which the observations were extracted from the NCEP-NCAR reanalysis, but the same multiplicative inflation was used as in the perfect model case. The *blue line* and the *black solid lines* correspond to the application of optimized multiplicative and additive inflation, respectively. The *long-dashed line* was obtained correcting the bias with the Dee and da Silva (1998) method, and combining it with additive inflation. The *short-dashed* is as the *long-dashed* but using the Danforth et al. (2006) low-dimensional method to correct the bias, and the *green line* is as the *long-dashed line* but using the simplified Dee and da Silva method. Adapted from Li (2007)

observations minus analysis, and analysis minus forecasts and the background and observational error covariances. Li (2007) took advantage of these relationships to develop a method to adaptively estimate both the observation errors variance and the optimal inflation of the background error covariance. This method has been successfully tested in several models (Li et al. 2009a; Reichle et al. 2008).

## 4 Summary and Discussion

4D-Var and the EnKF are the most advanced methods for data assimilation. 4D-Var has been widely adopted in operational centres, with excellent results and much accumulated experience. EnKF is less mature, and has the disadvantage that the corrections introduced by observations are done in spaces of lower dimension that depend on the ensemble size, although this problem is ameliorated by the use of localization. The main advantages of the EnKF are that it provides an estimate of the forecast and analysis error covariances, and that it is much simpler to implement than 4D-Var. A recent WWRP/THORPEX Workshop in Buenos Aires,

10–13 November 2008, was dedicated to 4D-Var and Ensemble Kalman Filter Intercomparisons with many papers and discussions (http://4dvarenkf.cima.fcen.uba.ar/). Buehner et al. (2008) presented "clean" comparisons between the operational 4D-Var and EnKF systems in Environment Canada, using the same model resolution and observations, showing that their forecasts had essentially identical scores in the Northern Hemisphere, whereas a hybrid system based on 4D-Var but using a background error covariance based on the EnKF gave a 10-h improvement in the 5-day forecasts in the Southern Hemisphere. This supports the statement that the best approach should be a hybrid that combines "the best characteristics" of both EnKF and 4D-Var (e.g. Lorenc 2003; Barker 2008). This would also bring the main disadvantage of 4D-Var to the hybrid system, i.e., the need to develop and maintain an adjoint model. This makes the hybrid approach attractive to operational centres that already have appropriate linear tangent and adjoint models, but less so to other centres.

In this review we have focused on the idea that the advantages and new techniques developed over the years for 4D-Var, can be adapted and implemented within the EnKF framework, without requiring an adjoint model. The LETKF (Hunt et al. 2007) was used as a prototype of the EnKF. It belongs to the square-root or deterministic class of the EnKF (e.g. Whitaker and Hamill 2002) but simultaneously assimilates observations locally in space, and uses the ensemble transform approach of Bishop et al. (2001) to obtain the analysis ensemble as a linear combination of the background forecasts.

We showed how the LETKF could be modified to include some of the most important 4D-Var advantages. In particular, the 3D-LETKF or 4D-LETKF can be used as a smoother that is cost-free beyond the computation of the filter and storing the weights. This allows a faster spin-up in the "running in place" method, so that the LETKF spins up as fast as 4D-Var. This is important in situations such as the forecast of severe storms, which cannot wait for a slow ensemble spin-up. Long assimilation windows and the consequent non-linearity of the perturbations typically result in non-Gaussianity of the ensemble perturbations and, as a result, a poorer performance of LETKF compared to 4D-Var. The no-cost smoothing method can be used to perform the equivalent of the 4D-Var "outer loop" and help deal with the problem of non-linearity. One of the most powerful applications of the adjoint model is the ability to estimate the impact of a class of observations on the short range forecast (Langland and Baker 2004). Liu and Kalnay (2008) have shown how to perform the same "adjoint sensitivity" within the LETKF without an adjoint model. Yang et al. (2009b) showed that the analysis weights created by the LETKF vary smoothly on horizontal scales much larger than the analyses or the analysis increments, so that the analyses can be performed on a very coarse grid and the weights interpolated to the full resolution grid. Because these weights are applied to the full resolution model, Yang et al. (2009b) found that the weight interpolation from a coarse resolution grid did not degrade the analysis, suggesting that the weights vary on large scales, and that smoothing the weights can increase the accuracy of the analysis. Li et al. (2009a) compared several methods used to correct model errors and showed that it is advantageous to combine methods that correct the bias, such as that of

Dee and da Silva (1998) and the low-dimensional method of Danforth et al. (2006), with methods like inflation that are more appropriate to account for random model errors. This is an alternative to the weak constraint method (Trémolet 2007) to deal with model errors in 4D-Var, and involves the addition of a relatively small number of degrees of freedom to the control vector. Li et al. (2009b) also showed how observation errors and background error inflation can be estimated adaptively within EnKF.

In summary, we have emphasized that the EnKF can profit from the methods and improvements that have been developed in the wide research and operational experience acquired with 4D-Var. Given that operational tests comparing 4D-Var and the LETKF indicate that the performance of these two methods is already very close (e.g. Miyoshi and Yamane 2007; Buehner et al. 2008), and that the LETKF and other EnKF methods are simpler to implement, their future looks bright. For centres that have access to the model adjoint, hybrid 4D-Var-EnKF may be optimal.

**Acknowledgments** I want to thank the members of the Chaos-Weather group at the University of Maryland, and in particular to Profs. Shu-Chih Yang, Brian Hunt, Kayo Ide, Eric Kostelich, Ed Ott, Istvan Szunyogh, and Jim Yorke. My deepest gratitude is to my former students at the University of Maryland, Drs. Matteo Corazza, Chris Danforth, Hong Li, Junjie Liu, Takemasa Miyoshi, Malaquías Peña, Shu-Chih Yang, Ji-Sun Kang, Matt Hoffman, and present students Steve Penny, Steve Greybush, Tamara Singleton, Javier Amezcua and others, whose creative research allowed us to learn together. Interactions with the thriving Ensemble Kalman Filter community members, especially Ross Hoffman, Jeff Whitaker, Craig Bishop, Kayo Ide, Joaquim Ballabrera, Jidong Gao, Zoltan Toth, Milija Zupanski, Tom Hamill, Herschel Mitchell, Peter Houtekamer, Chris Snyder, Fuqing Zhang and others, as well as with Michael Ghil, Arlindo da Silva, Jim Carton, Dick Dee, and Wayman Baker, have been a source of inspiration. Richard Ménard, Ross Hoffman, Kayo Ide, Lars Nerger and William Lahoz made important suggestions that improved the review and my own understanding of the subject.

## References

Anderson, J.L., 2001. An ensemble adjustment Kalman filter for data assimilation. *Mon. Weather Rev.*, **129**, 2884–2903.

Andersson, E., M. Fisher, E. Hólm, L. Isaksen, G. Radnoti and Y. Trémolet, 2005. Will the 4D-Var approach be defeated by nonlinearity? *ECMWF Technical Memorandum,* No. 479.

Baek, S.-J., B.R. Hunt, E. Kalnay, E. Ott and I. Szunyogh, 2006. Local ensemble Kalman filtering in the presence of model bias. *Tellus A*, **58**, 293–306.

Barker, D.M., 2008. How 4DVar can benefit from or contribute to EnKF (a 4DVar perspective). Available from http://4dvarenkf.cima.fcen.uba.ar/Download /Session_8/4DVar_EnKF_Barker.pdf.

Bishop, C.H., B.J. Etherton and S.J. Majumdar, 2001. Adaptive sampling with ensemble transform Kalman filter. Part I: Theoretical aspects. *Mon. Weather Rev.*, **129**, 420–436.

Buehner, M., C. Charente, B. He, et al., 2008. Intercomparison of 4-D Var and EnKF systems for operational deterministic NWP. Available from http://4dvarenkf.cima.fcen.uba.ar/Download /Session_7/Intercomparison_4D-Var_EnKF_Buehner.pdf.

Burgers, G., P.J. van Leeuwen and G. Evensen, 1998. On the analysis scheme in the ensemble Kalman filter. *Mon. Weather Rev.*, **126**, 1719–1724.

Cardinali, C., S. Pezzulli and E. Andersson, 2004. Influence-matrix diagnostic of a data assimilation system. *Q. J. R. Meteorol. Soc.*, **130**, 2767–2786.

Caya, A., J. Sun and C. Snyder, 2005. A comparison between the 4D-VAR and the ensemble Kalman filter techniques for radar data assimilation. *Mon. Weather Rev.*, **133**, 3081–3094.

Courtier, P. and O. Talagrand, 1990. Variational assimilation of meteorological observations with the direct and adjoint shallow water equations. *Tellus*, **42A**, 531–549.

Danforth, C.M. and E. Kalnay, 2008. Using singular value decomposition to parameterize state dependent model errors. *J. Atmos. Sci.*, **65**, 1467–1478.

Danforth, C.M., E. Kalnay and T. Miyoshi, 2006. Estimating and correcting global weather model error. *Mon. Weather Rev.*, **134**, 281–299.

Dee, D.P. and A.M. da Silva, 1998. Data assimilation in the presence of forecast bias. *Q. J. R. Meteorol. Soc.*, **124**, 269–295.

Desroziers, G., L. Berre, B. Chapnik and P. Poli, 2005. Diagnosis of observation, background and analysis-error statistics in observation space. *Q. J. R. Meteorol. Soc.*, **131**, 3385–3396.

Evensen, G., 1994. Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *J. Geophys. Res.*, **99**, 10,143–10,162.

Evensen, G., 2003. The ensemble Kalman filter: Theoretical formulation and practical implementation. *Ocean Dyn.*, **53**, 343–367.

Evensen, G. and P.J. van Leeuwen, 1996. Assimilation of Geosat altimeter data for the Agulhas current using the ensemble Kalman filter with a quasi-geostrophic model. *Mon. Weather Rev.*, **124**, 85–96.

Fisher, M., M. Leutbecher and G. Kelly, 2005. On the equivalence between Kalman smoothing and weak-constraint four-dimensional variational data assimilation. *Q. J. R. Meteorol. Soc.*, **131**, 3235–3246.

Gaspari, G. and S.E. Cohn, 1999. Construction of correlation functions in two and three dimensions. *Q. J. R. Meteorol. Soc.*, **125**, 723–757.

Ghil, M. and P. Malanotte-Rizzoli, 1991. Data assimilation in meteorology and oceanography. *Adv. Geophys.*, **33**, 141–266.

Greybush, S., E. Kalnay, T. Miyoshi, K. Ide and B. Hunt, 2009. EnKF localization techniques and balance. *Presented at the WMO 5th International Symposium on Data Assimilation*. Melbourne, Australia, 6–9 October 2009, submitted to *Mon. Weather Rev*.,Available at http://www.weatherchaos.umd.edu/papers/Greybush_Melbourne2009.ppt.

Gustafsson, N., 2007. Response to the discussion on "4-D-Var or EnKF?". *Tellus A*, **59**, 778–780.

Hamill, T.M., J.S. Whitaker and C. Snyder, 2001. Distance-dependent filtering of background error covariance estimates in an ensemble Kalman filter. *Mon. Weather Rev.*, **129**, 2776–2790.

Harlim, J. and B.R. Hunt, 2007a. A non-Gaussian ensemble filter for assimilating infrequent noisy observations. *Tellus A*, **59**, 225–237.

Harlim, J. and B.R. Hunt, 2007b. Four-dimensional local ensemble transform Kalman filter: Variational formulation and numerical experiments with a global circulation model. *Tellus A*, **59**, 731–748.

Houtekamer, P.L. and H.L. Mitchell, 1998. Data assimilation using an ensemble Kalman filter technique. *Mon. Weather Rev.*, **126**, 796–811.

Houtekamer, P.L. and H.L. Mitchell, 2001. A sequential ensemble Kalman filter for atmospheric data assimilation. *Mon. Weather Rev.*, **129**, 123–137.

Houtekamer, P.L., H.L. Mitchell, G. Pellerin, M. Buehner, M. Charron, L. Spacek and B. Hansen, 2005. Atmospheric data assimilation with an ensemble Kalman filter: Results with real observations. *Mon. Weather Rev.*, **133**, 604–620.

Hunt, B.R., 2005. An efficient implementation of the local ensemble Kalman filter. Available at http://arxiv.org./abs/physics/0511236.

Hunt, B.R., E. Kalnay, E.J. Kostelich, et al., 2004. Four-dimensional ensemble Kalman filtering. *Tellus*, **56A**, 273–277.

Hunt, B.R., E.J. Kostelich and I. Szunyogh, 2007. Efficient data assimilation for spatiotemporal chaos: A local ensemble transform Kalman filter. *Physica D*, **230**, 112–126.

Ide, K., P. Courtier, M. Ghil and A. Lorenc, 1997. Unified notation for data assimilation: Operational, sequential and variational. *J. Meteorol. Soc. Jpn.*, **75**, 181–189.

Järvinen, H., E. Andersson and F. Bouttier, 1999. Variational assimilation of time sequences of surface observations with serially correlated errors. *Tellus*, **51A**, 469–488.

Jazwinski, A.H., 1970. *Stochastic Processes and Filtering Theory*. Academic Press, NY, 376 pp.

Kalman, R.E., 1960. A new approach to linear filtering and prediction problems. *J. Basic Eng.*, **82**, 35–45.

Kalnay, E., 2003. *Atmospheric Modeling, Data Assimilation and Predictability.* Cambridge University Press, Cambridge, UK, 341 pp.

Kalnay, E., H. Li, T. Miyoshi, S.-C. Yang and J. Ballabrera-Poy, 2007a. 4D-Var or ensemble Kalman filter? *Tellus A*, **59**, 758–773.

Kalnay, E., H. Li, T. Miyoshi, S.-C. Yang and J. Ballabrera-Poy, 2007b. Response to the discussion on "4D-Var or EnKF?" by Nils Gustaffson. *Tellus A*, **59**, 778–780.

Kalnay, E. and S.-C. Yang, 2008. Accelerating the spin-up in EnKF. *Arxiv: physics:Nonlinear/0.806.0180v1*.

Keppenne, C.L., 2000. Data assimilation into a primitive-equation model with a parallel ensemble Kalman filter. *Mon. Weather Rev.*, **128**, 1971–1981.

Keppenne, C. and H. Rienecker, 2002. Initial testing of a massively parallel ensemble Kalman filter with the Poseidon isopycnal ocean general circulation model. *Mon. Weather Rev.,* **130**, 2951–2965.

Langland, R.H. and N.L. Baker, 2004. Estimation of observation impact using the NRL atmospheric variational data assimilation adjoint system. *Tellus*, **56A**, 189–201.

Li, H., 2007. Local ensemble transform Kalman filter with realistic observations. Ph. D. thesis. Available at http://hdl.handle.net/1903/7317.

Li, H., E. Kalnay and T. Miyoshi, 2009a. Simultaneous estimation of covariance inflation and observation errors within an ensemble Kalman filter. *Q. J. R. Meteorol. Soc.*, **135**, 523–533.

Li, H., E. Kalnay, T. Miyoshi and C.M. Danforth, 2009b. Accounting for model errors in ensemble data assimilation. *Mon. Weather Rev.*, **137**, 3407–3419.

Liu, J. and E. Kalnay, 2008. Estimating observation impact without adjoint model in an ensemble Kalman filter. *Q. J. R. Meteorol. Soc.*, **134**, 1327–1335.

Liu, J., E. Kalnay, T. Miyoshi and C. Cardinali, 2009. Analysis sensitivity calculation in an ensemble Kalman filter. *Q. J. R. Meteorol. Soc.*, **135**, 523–533.

Lorenc, A.C., 1986. Analysis methods for numerical weather prediction. *Q. J .R. Meteorol. Soc.*, **112**, 1177–1194.

Lorenc, A.C., 2003. The potential of the ensemble Kalman filter for NWP – a comparison with 4D-Var. *Q. J. R. Meteorol. Soc.*, **129**, 3183–3203.

Lorenz, E., 1963. Deterministic non-periodic flow. *J. Atmos. Sci.*, **20**, 130–141.

Mitchell, H.L., P.L. Houtekamer and G. Pellerin, 2002. Ensemble size, balance, and model-error representation in an ensemble Kalman filter. *Mon. Weather Rev.*, **130**, 2791–2808.

Miyoshi, T., 2005. *Ensemble Kalman Filter Experiments with a Primitive-Equation Global Model*. Doctoral dissertation, University of Maryland, College Park, 197 pp. Available at https://drum.umd.edu/dspace/handle/1903/3046.

Miyoshi, T. and S. Yamane, 2007. Local ensemble transform Kalman filtering with an AGCM at a T159/L48 resolution. *Mon. Weather Rev.*, **135**, 3841–3861.

Molteni, F., 2003. Atmospheric simulations using a GCM with simplified physical parameterizations. I: Model climatology and variability in multi-decadal experiments. *Clim. Dyn.*, **20**, 175–191.

Nerger, L., W. Hiller and J. Scroeter, 2005. A comparison of error subspace Kalman filters. *Tellus*, **57A**, 715–735.

Nutter, P., M. Xue and D. Stensrud, 2004. Application of lateral boundary condition perturbations to help restore dispersion in limited-area ensemble forecasts. *Mon. Weather Rev.*, **132**, 2378–2390.

Ott, E., B.R. Hunt, I. Szunyogh, A.V. Zimin, E.J. Kostelich, M. Corazza, E. Kalnay, D.J. Patil and J.A. Yorke, 2004. A local ensemble Kalman filter for atmospheric data assimilation. *Tellus*, **56A**, 415–428.

Pham, D.T., 2001. Stochastic methods for sequential data assimilation in strongly nonlinear systems. *Mon. Weather Rev.*, **129**, 1194–1207.

Pires, C., R. Vautard and O. Talagrand, 1996. On extending the limits of variational assimilation in chaotic systems. *Tellus*, **48A**, 96–121.

Rabier, F., H. Järvinen, E. Klinker, J.-F. Mahfouf and A. Simmons, 2000. The ECMWF operational implementation of four-dimensional variational physics. *Q. J. R. Meteorol. Soc.*, **126**, 1143–1170.

Radakovich, J.D., P.R. Houser, A.M. da Silva and M.G. Bosilovich, 2001. Results from global land-surface data assimilation methods. *Proceeding of the 5th Symposium on Integrated Observing Systems*, 14–19 January 2001, Albuquerque, NM, pp 132–134.

Reichle, R.H., W.T. Crow and C.L. Keppenne, 2008. An adaptive ensemble Kalman filter for soil moisture data assimilation. *Water Resour. Res.*, **44**, W03423, doi:10.1029/2007WR006357.

Rotunno, R. and J.W. Bao, 1996. A case study of cyclogenesis using a model hierarchy. *Mon. Weather Rev.*, **124**, 1051–1066.

Szunyogh, I., E. Kostelich, G. Gyarmati, E. Kalnay, B.R. Hunt, E. Ott, E. Satterfield and J.A. Yorke, 2008. A local ensemble transform Kalman filter data assimilation system for the NCEP global model. *Tellus*, **60A**, 113–130.

Talagrand, O. and P. Courtier, 1987. Variational assimilation of meteorological observations with the adjoint vorticity equation I: theory. *Q. J. R. Meteorol. Soc.*, **113**, 1311–1328.

Thépaut, J.-N. and P. Courtier, 1991. Four-dimensional data assimilation using the adjoint of a multilevel primitive equation model. *Q. J. R. Meteorol. Soc.*, **117**, 1225–1254.

Tippett, M.K., J.L. Anderson, C.H. Bishop, T.M. Hamill and J.S. Whitaker, 2003. Ensemble square root filters. *Mon. Weather Rev.*, **131**, 1485–1490.

Torn, R.D. and G.J. Hakim, 2008. Performance characteristics of a Pseudo-Operational Ensemble Kalman Filter. *Mon. Weather Review*, **136**, 3947–3963.

Torn, R.D., G.J. Hakim and C. Snyder, 2006. Boundary conditions for limited-area ensemble Kalman filters. *Mon. Weather Rev.*, **134**, 2490–2502.

Trémolet, Y., 2007. Model-error estimation in 4D-Var. *Q. J. R. Meteorol. Soc.*, **133**, 1267–1280.

Wang, X., C.H. Bishop and S.J. Julier, 2004. Which is better, an ensemble of positive-negative pairs or a centered spherical simplex ensemble? *Mon. Weather Rev.*, **132**, 1590–1605.

Whitaker, J.S., G.P. Compo, X. Wei and T.M. Hamill, 2004. Reanalysis without radiosondes using ensemble data assimilation. *Mon. Weather Rev.*, **132**, 1190–1200.

Whitaker, J.S. and T.M. Hamill, 2002. Ensemble data assimilation without perturbed observations. *Mon. Weather Rev.*, **130**, 1913–1924.

Whitaker, J.S., T.M. Hamill, X. Wei, Y. Song and Z. Toth, 2008. Ensemble data assimilation with the NCEP global forecast system. *Mon. Weather Rev.*, **136**, 463–482.

Yang, S.-C., M. Corazza, A. Carrassi, E. Kalnay and T. Miyoshi, 2009a. Comparison of ensemble-based and variational-based data assimilation schemes in a quasi-geostrophic model. *Mon. Weather Rev*., **137**, 693–709.

Yang, S.-C., E. Kalnay, B. Hunt and N. Bowler, 2009b. Weight interpolation for efficient data assimilation with the local ensemble transform Kalman filter. *Q. J. R. Meteorol. Soc.*, **135**, 251–262.

Zhu, Y. and R. Gelaro, 2008. Observation sensitivity calculations using the adjoint of the gridpoint statistical interpolation (GSI) analysis system. *Mon. Weather Rev.*, **136**, 335–351.

Zupanski, M., 2005. Maximum likelihood ensemble filter: Theoretical aspects. *Mon. Weather Rev.*, **133**, 1710–1726.

Zupanski, M., S.J. Fletcher, I.M. Navon, et al., 2006. Initiation of ensemble data assimilation. *Tellus*, **58A**, 159–170.

# Error Statistics in Data Assimilation: Estimation and Modelling

**Mark Buehner**

## 1 Introduction

As already discussed in previous chapters in Part I, *Theory*, the purpose of data assimilation is to use observations to compute an "optimal" correction to a background state by using estimates of the uncertainty associated with the background state and the observations. The uncertainty is typically characterized by covariance matrices for the error in the background state and the observations. These covariance matrices determine the level of influence each observation has on the analysis and how this influence is distributed spatially, temporally and among the different types of analysis variables. The optimality of any assimilation approach based on linear estimation theory depends on the validity of a set of assumptions, including that the errors in the background state and observations are Gaussian with both zero bias and precisely known covariances. The estimation of these covariances is a difficult problem, partly due to a lack of knowledge of the statistical properties of background and observation error. As pointed out by Dee (1995), the number of available observations of the atmosphere or ocean is generally many orders of magnitude less than that required to estimate the full error covariances. In addition, especially for the case of background errors, the computational challenge of estimating the full covariance matrix of a random vector containing at least $O(10^6)$ elements also poses a significant challenge. This chapter outlines the theory and some practical approaches used to estimate and model background and observation error statistics. Because most data assimilation approaches currently used for realistic atmospheric and oceanographic applications rely on the assumption of Gaussian error distributions, our focus here is restricted to the estimation of error covariances and not the higher-order statistical moments or the full probability distributions.

M. Buehner (✉)
Meteorological Research Division, Data Assimilation and Satellite Meteorology Section, Environment Canada, Canada
e-mail: mark.buehner@ec.gc.ca

W. Lahoz et al. (eds.), *Data Assimilation*, DOI 10.1007/978-3-540-74703-1_5,

## 1.1 Source of Statistical Information

Errors in the background state and the observations are defined with respect to the true state of the system (e.g. atmosphere or ocean), respectively, as

$$\begin{aligned} \boldsymbol{\varepsilon}^b &= \mathbf{x}^b - \mathbf{x}^t \,, \\ \boldsymbol{\varepsilon}^o &= \mathbf{y} - \mathcal{H}(\mathbf{x}^t) \,, \end{aligned} \tag{1}$$

where $\mathbf{y}$ is a vector containing the observations, $\mathcal{H}$ is the observation operator, $\mathbf{x}^t$ is the true state and $\mathbf{x}^b$ is the background state. Since we do not know the true state of the system, it is impossible to directly compute the background and observation errors. Assuming for the moment that we do precisely know the observation operator $\mathcal{H}$, the only useful quantity from which we can compute the required error statistics is the innovation vector, defined as

$$\mathbf{d} = \mathbf{y} - \mathcal{H}\left(\mathbf{x}^b\right) = \boldsymbol{\varepsilon}^o + \mathcal{H}(\mathbf{x}^t) - \mathcal{H}(\mathbf{x}^t + \boldsymbol{\varepsilon}^b) \approx \boldsymbol{\varepsilon}^o - \mathbf{H}\,\boldsymbol{\varepsilon}^b \tag{2}$$

where $\mathbf{H}$ is the linearized version of $\mathcal{H}$. Clearly, the observation and background errors cannot both be obtained from this single equation. By making the common assumption that the background and observation errors are uncorrelated with each other, the innovation covariance matrix is given by

$$\mathbf{S} = \mathbf{R} + \mathbf{H}\mathbf{B}\mathbf{H}^T \,, \tag{3}$$

where $\mathbf{R}$ is the observation error covariance matrix, $\mathbf{B}$ is the background error covariance matrix and the superscript $T$ represents matrix transposition. Practically, the covariance matrix $\mathbf{S}$ can be estimated by averaging over time, if the observing network remains relatively fixed and the error covariances are assumed stationary in time. However, again it is impossible to obtain both terms on the right-hand side from this single equation. This represents a fundamental problem in estimating the error probability distribution functions (PDFs) for data assimilation. Only by relying on additional assumptions about the background and observation error PDFs, can the two components that contribute to $\mathbf{S}$ be separated. Furthermore, these additional assumptions cannot be directly validated, but must be based on information independent from the actual values of the background state and observations (see, e.g., Dee 1995; Talagrand 1999).

## 1.2 Importance of Background and Observation Error Statistics in Data Assimilation

The importance of the background error covariances can be seen by examining the linear analysis equation (see, e.g., Gelb 1974)

$$\Delta\mathbf{x} = \mathbf{B}\mathbf{H}^T \left(\mathbf{H}\mathbf{B}\mathbf{H}^T + \mathbf{R}\right)^{-1} \mathbf{d} \,, \tag{4}$$

where $\Delta\mathbf{x}$ is the analysis increment. If we take the case where only a single observation is assimilated, then the quantities in parentheses in Eq. (4) are scalars

and the resulting analysis increment is proportional to $\mathbf{BH}^T$, where $\mathbf{H}^T$ is a column vector. Consequently, for a given type of observation the spatial and multivariate structure of the analysis increment depends strongly on the background error covariances.

Moreover, as seen in the chapter *Mathematical Concepts of Data Assimilation* (Nichols), the weight given to an observation is governed by the relative amplitude of background and observation error variances. In general, a larger background error variance results in a larger correction to the background state and more weight given to the observation. Similarly, a larger observation error variance results in a smaller correction to the background state. The effect of the background and observation error correlations is to determine how the relative importance of the background state and observations varies as a function of the spatial or temporal scale. For example, broad monotonically decreasing spatial correlations have the effect of concentrating more of the error variance in the large-scale component of the error and less in the small-scale component (Daley 1991, Sect. 4.8). As a result, the relative weight given to the observations can vary as a function of scale (either in space or time) if the background and observation errors have different spatial or temporal correlations. For example, when the specified background error correlations are broader than the observation error correlations, the large-scale component of the background state is considered to be less accurate than the small-scale component relative to the observations. Consequently, the analysis increment tends to be smooth because more weight is given to the large-scale component of the observations.

## 2 Estimation of Background and Observation Error Statistics

Background and observation error statistics are typically estimated from either a statistical study of the innovations (Hollingsworth and Lönnberg 1986) or an ad hoc method such as the NMC (National Meteorological Center) method (Parrish and Derber 1992). Another approach is to use Monte Carlo simulations to approximate the effect of observation and model errors (assuming their PDFs are known) in the forecast-analysis cycle to obtain random realizations of background error. Using an ensemble of such error realizations, the background error PDF can be estimated. An example of this is the Ensemble Kalman filter (EnKF) (see chapter *Ensemble Kalman Filter: Current Status and Potential*, Kalnay; Evensen 1994). In this section several approaches for estimating the background and observation error covariances are described.

### *2.1 Estimation of Background and Observation Error Statistics from Innovations*

As already mentioned, the innovations represent the only direct source of information for estimating the background and observation error statistics. A frequently employed approach for using innovations to estimate the error statistics was

developed for use with conventional radiosonde observations (Hollingsworth and Lönnberg 1986). The principle assumption on which the method is based is that the errors in the observations originating from distinct balloons are uncorrelated. In addition, the horizontal spacing of the observations must be sufficiently small to resolve the background error correlations and the background and observation error statistics are assumed to be horizontally homogeneous. The innovation covariances are then estimated for a particular pressure level as a function of the horizontal separation distance. It is usually necessary to assume the error statistics are stationary in time so that data over a time period of several weeks can be used to obtain a sufficiently large sample size. The covariances from distinct balloons (non-zero separation distance) are then extrapolated to zero separation distance. The extrapolated value represents an estimate of the background error variance and by subtracting this from the innovation variance an estimate of the observation error variance is obtained. The horizontal background error correlations can also be estimated from the horizontal innovation covariances, with the estimated observation error variance removed at zero separation distance. Due to the limited and heterogeneous distribution of radiosonde observations over the globe, the approach can only provide information on the background error covariances over particular regions and for a limited range of spatial scales. Also, the application of this approach to satellite observations that are more uniformly distributed may be difficult due to the possibility of horizontally correlated observation errors and the limited horizontal and vertical resolution of the observations.

Another type of approach uses an existing variational data assimilation system or components of that system to estimate elements of the error covariances from the innovations. With this approach, the lack of consistency of the covariance matrices specified in the data assimilation system and the innovations is used to tune a small set of covariance parameters. For example, it may be assumed that the specified background and observation error correlations are correct and that only the variances need be scaled by a set of horizontally constant factors. The level of consistency between the specified covariances and the innovations is measured by comparing the value of a component of the cost function with its expected value computed using a randomization method (Desroziers and Ivanov 2001; Chapnik et al. 2004, 2006). Alternatively, a likelihood function can be constructed using the innovations and the covariance matrix $\mathbf{S}$ and an iterative scheme used to find the covariance parameters that maximize the likelihood (Dee 1995). In the case where only the variances are tuned, the accuracy of the estimated values depends strongly on the assumption of accurate error correlations. For example, Chapnik et al. (2004) showed that if the observation error is assumed to be uncorrelated, but the real error is correlated, then the approach will underestimate the observation error variance (sometimes giving a value as small as zero) and overestimate the background error variance.

More recently, Desroziers et al. (2005) demonstrated how covariance parameters could be estimated by simply computing particular statistics from the routine output of a data assimilation system. For example, the expected value for the observation error covariance matrix should be equal to $\mathrm{cov}(\mathbf{y} - \mathbf{H}\mathbf{x}^a, \mathbf{y} - \mathbf{H}\mathbf{x}^b)$, where $\mathbf{x}^a$ is the analysed state. A similar relation is also easily computed for the expected

value of the background error covariances after they are projected into observation space. An inconsistency between the expected covariance and the covariance specified in the assimilation system, for example, a particular observation error variance, would mean that this error variance should be adjusted. Desroziers et al. (2005) showed, in an idealized setting, that an iterative scheme involving the data assimilation system itself can be constructed that converges towards the true error variance. The advantage of this approach over the others mentioned above is that it requires very minimal changes to an existing data assimilation system. However, as with the other approaches, any incorrect assumption regarding the structure of the error covariances (e.g. spatially uncorrelated observation errors) will likely result in convergence towards incorrect variance estimates.

## *2.2 Estimation of Background Error Covariances with the Lagged-Forecast (NMC) Method*

Several Numerical Weather Prediction (NWP) centres employ variational assimilation systems with stationary background error covariances estimated using the "NMC method" or lagged-forecast difference method (Parrish and Derber 1992; Gauthier et al. 1998; Rabier et al. 1998; Derber and Bouttier 1999). Following this method, the differences between pairs of forecasts valid at the same time, but having different lead times, are taken to be representative of background error. Such forecast differences can easily be computed for a past period using the archived output of an operational forecasting system. For example, at the Canadian Meteorological Centre, the differences between 48- and 24-h forecasts taken over a period of 2–3 months are used (Gauthier et al. 1998). However, a lack of correspondence between these lagged forecast differences and 6-h forecast error necessitates modification of the computed covariances, especially the variances. The variances may be tuned using a method based on the innovations as outlined in the previous section.

## *2.3 Estimation of Background Error Covariances with Monte Carlo Approaches*

Methods based on Monte Carlo simulation have been developed to address the problem of how errors in the inputs to a data assimilation system lead to errors in the background (and analysed) state. If the PDFs of both the observation and model error are known, then these approaches, such as the EnKF, provide an estimate of the PDF (and therefore the covariances) of the resulting background error. An ensemble of analysis-forecast experiments are run, each using a set of observations and short-term model integrations perturbed with an independent realization of errors drawn from their known observation and model error PDFs. If the error PDF remains close to Gaussian, the resulting ensemble of background states is representative

of a random sample drawn from the background error PDF (Burgers et al. 1998).

In the EnKF, the analysis step for each ensemble member is performed using background error covariances estimated from the ensemble spread of forecasts valid for that specific analysis time. Typically at least $O(100)$ ensemble members are used to obtain a sufficiently accurate estimate of the error statistics. A simpler approach used to estimate the stationary background error covariances is similar to that described by Houtekamer et al. (1996) and was recently used in place of the NMC method at several NWP centres (Fisher and Andersson 2001; Buehner 2005; Buehner et al. 2005; Belo Pereira and Berre 2006). In that approach, the analysis step for each ensemble member employs a previous, usually static, estimate of the background error covariances. Unlike the EnKF, such approaches have been employed with only a small number of parallel analysis-forecast experiments, but where the background error realizations are pooled over a sufficiently long time period to obtain an estimate of the stationary, or slowly varying, component of the error statistics. However, like with the NMC method, this approach can only be used to estimate the background error covariances after an initial assimilation system with its own background error covariances is available. A major challenge for all approaches based on Monte Carlo simulation is the specification of the model error PDF. Approaches have been examined to adaptively tune a parametrized form of the model error covariances with some success in idealized settings (e.g. Mitchell and Houtekamer 2000).

## *2.4 Other Approaches for the Estimation of Background Error Covariances*

Some other approaches have been examined for obtaining low-rank estimates of the true flow-dependent background error covariances such as would be obtained, in a linear context, with a Kalman filter.

An approximate reduced-rank Kalman filter was developed and tested at the European Centre for Medium-Range Weather Forecasts (ECMWF) with the goal of providing an improved background error covariance matrix for the operational 4D-Var (four dimensional variational) assimilation system (Fisher 1998; Fisher and Andersson 2001; see chapter *Variational Assimilation*, Talagrand). The approach uses partially evolved singular vectors to define the background error covariances in a low-dimensional subspace. In the orthogonal subspace that spans the remainder of the analysis space, the standard stationary background error covariances are used. The singular vectors are computed with a 48-h optimization time and an initial-time norm defined using the inverse of an approximation to the analysis error covariance matrix at the previous analysis time. The result is a set of vectors that eventually evolve into the leading eigenvectors of the 48-h forecast error covariance matrix, under the assumption that the error growth can be described by linearized dynamics and that the contribution from model error is negligible. After extensive testing in

a realistic NWP context, it was found that the reduced-rank Kalman filter did not lead to consistent improvements to forecast quality. Possible explanations given for the lack of positive impact include that the estimate of the analysis error covariance matrix may not have been sufficiently accurate.

A related approach uses the gradient, with respect to the initial conditions, of some specified scalar function of the future state of the system (Hello and Bouttier 2001). Like the reduced-rank Kalman filter described above, the gradient vector is used to specify the background error covariances in a low-dimensional subspace (in this case just a single direction); the standard background error covariances are used for the remaining orthogonal subspace. By employing the sensitivity to initial conditions of a series of 48-h forecasts of cyclones, Hello and Bouttier (2001) were able to improve the forecasts of these cyclones as compared with the standard 3D-Var (three-dimensional variational; see chapter *Variational Assimilation*, Talagrand) approach. This is despite the fact that, unlike the singular vectors employed by Fisher and Andersson (2001), the calculation of the gradient vector involves solely the dynamics and does not include any statistical information concerning the errors.

## *2.5 Estimation of Observation-Error Correlations*

Compared with background errors, approaches for estimating the correlations of observation error have not been examined as extensively. However, to make optimal use of the ever increasing volume of available satellite data, it is becoming necessary to obtain accurate estimates of both their spatial and inter-channel error correlations. When employing a four-dimensional assimilation approach it may even be necessary to account for temporal error correlations for both satellite and conventional data.

To date, a common approach for dealing with correlated errors is to simply thin the data either temporally, spatially, or with respect to the radiance frequency channel. The thinned data is then assimilated under the hypothesis that the observation errors are uncorrelated. While this approach effectively reduces the error correlations among the data that survive the thinning procedure, it also may eliminate a significant amount of useful information on the small-scale structure of the atmospheric or oceanic state. For example, let us assume that a particular observation type has errors that are positively correlated in the horizontal direction. Properly accounting for these correlations (instead of assuming uncorrelated errors) in the data assimilation procedure would increase the weight given to the small-scale component while reducing the weight given to the large-scale component. Horizontally thinning the data also results in decreased weight given to the large-scale component, but it does so at the expense of reducing the information content of the data at the small scales. In fact, data with smooth positive error correlations are more accurate with respect to the small-scale component than data with uncorrelated errors, assuming the spatial resolution and error variance are equal (Daley 1991, Sect. 4.8).

The problem of horizontally correlated errors associated with atmospheric motion vector (AMV) data derived from satellite observations of cloud or humidity field motion has been studied by Bormann et al. (2003). The AMV data are co-located with radiosonde wind observations and the covariances of their difference estimated as a function of horizontal separation distance. The covariances for non-zero separation distance are then extrapolated to zero separation, like in the approaches described in Sect. 2.1 above for separating innovation covariances into the contributions from observation and background errors. Under the assumption that the radiosonde observation errors are horizontally uncorrelated, the AMV error correlations are estimated from the horizontally correlated component of the covariances.

A similar approach, used by Garand et al. (2007), allows the inter-channel error correlations to be estimated for AIRS (Atmospheric InfraRed Sounder) data. First, the inter-channel covariances for all possible pair-wise channel combinations of the innovation vector are computed as a function of horizontal separation distance. For each channel pair, the covariances for non-zero horizontal separation distance are extrapolated to zero separation distance. Then, the assumption is made that the covariances for non-zero separation distance are dominated by the background error. Consequently, the difference between the covariances computed at zero separation distance and the values obtained by extrapolation represent an estimate of the observation error variances for each channel and the inter-channel error covariances for all pair-wise channel combinations. The inter-channel error correlations are then obtained by normalizing the inter-channel covariances by the product of the corresponding error standard deviations. Even though the approach relies on the assumption that errors associated with AIRS data are horizontally uncorrelated, which has yet to be independently verified, the results appear physically realistic. Error correlations are generally high among the water vapour sensing channels and among surface sensitive channels. In contrast, they are negligible for channels within the main $CO_2$ absorption band.

With the increasing use of four-dimensional assimilation schemes for both atmospheric and oceanic state estimation, accounting for temporal correlations of observation error is becoming increasingly important. Properly incorporating estimates of temporal error correlations when assimilating time series of data would increase the weight given to the time tendency (high frequency component) and less weight to the time mean (low frequency component) of the data. This was demonstrated by Järvinen et al. (1999) for the case of time series of surface pressure data in a 4D-Var assimilation system. In that study the temporal error correlations were assumed to have a particular functional form and associated decorrelation time-scale, since objective approaches for estimating temporal error correlations had not yet been demonstrated.

In summary, the issue of accurately estimating spatial and temporal observation error correlations is becoming increasingly important. They will be necessary to make optimal use of the growing volume of both satellite and conventional data to

extract information on the small spatial and temporal scales at which the errors are often significantly correlated.

## 3 Modelling Error Covariances

For realistic NWP or oceanographic applications, a series of simplifying assumptions must be employed to obtain useful estimates of the background and observation error covariances. This is necessitated by both a lack of precise information on the background and observation errors (as addressed in the previous section) and the computational expense of utilizing the full covariance matrices in data assimilation systems. A typical NWP system, for example, could have background and observation error covariance matrices with $O(10^{14})$ elements. To be practical, any approach for modelling the covariances must significantly decrease the number of parameters required to define the covariances and also decrease the computational expense of employing the covariances in a data assimilation system. In addition, any approach must still capture the most important aspects of the true covariance structure. The main challenge to date has been to model the spatial correlations of the background errors, whereas observation errors have typically been assumed to be uncorrelated in most assimilation systems. Consequently, only approaches for modelling background error correlations are briefly discussed in this section.

### *3.1 Spectral Representation: Homogeneous and Isotropic Error Correlations*

A very efficient approach for modelling the background error correlations is to employ a spectral representation together with the assumption of homogeneity and isotropy for the horizontal correlations. Under these assumptions, the correlation matrix for a sphere in spectral space has a simple diagonal structure with elements that depend only on the total wavenumber (Courtier et al. 1998). Consequently, a full-rank matrix with reasonably smooth and robust correlations can be estimated from relatively few error samples. This representation for the horizontal correlations is often combined with vertical correlations in a way that does not require separability between the horizontal and vertical correlations. The resulting three-dimensional correlation matrix has a block diagonal structure given by

$$\hat{C}\,(n,\,k_1,\,k_2) = \left[\hat{C}_{\mathrm{h}}\,(n,\,k_1)\,\hat{C}_{\mathrm{h}}\,(n,\,k_2)\right]^{1/2} C_{\mathrm{v}}\,(k_1,\,k_2,\,n)\ , \tag{5}$$

where $\hat{C}_{\mathrm{h}}$ is the spectral horizontal correlations, $C_{\mathrm{v}}$ is the vertical correlation matrix for each horizontal total wavenumber ($n$), and $k$ represents the vertical level index. The non-separability of the correlations results in the dependence of the vertical correlations on the horizontal scale. Consequently, the horizontally small-scale

contribution to the errors typically has sharper vertical correlations than components at larger scales. This dependence is necessary to simultaneously obtain the correct correlations for wind and mass fields (Phillips 1986).

The original analysis variables can be transformed into a set of variables for which the assumptions of homogeneity and isotropy are more valid. For example, modelling wind correlations in terms of vorticity and divergence (or streamfunction and velocity potential) is often more accurate than using velocity components for which the correlations can be significantly anisotropic.

## *3.2 Physical-Space Representation*

Alternatively, the error covariances can be estimated without constraining the correlations to be horizontally homogeneous and isotropic. However, if the correlations are estimated directly from a small sample of background-error realizations without imposing any additional constraints, the rank of the resulting correlation matrix cannot exceed the sample size. In addition, the correlations often have a noisy structure and do not approach zero at long separation distances, even with unrealistically large correlations on the opposite side of the globe. To overcome these problems, a procedure for spatially localizing the correlations was proposed by Gaspari and Cohn (1999) and examined in the context of an EnKF by Houtekamer and Mitchell (2001) and Hamill et al. (2001). The technique for efficiently employing a spatially localized ensemble representation of the background error correlations in a variational assimilation framework was described by Lorenc (2003) and Buehner (2005). While reducing or eliminating distant correlations, spatial localization also increases the rank of the correlation matrix estimated from a given sample size of error realizations.

Another approach for reducing the problem of estimating the full correlation matrix from a small sample is to compute the weighted mean of such a correlation matrix with another matrix for which the assumptions of homogeneity and isotropy are imposed. This hybrid approach was used in the context of an EnKF by Hamill and Snyder (2000). In the variational context, a convenient approach is to combine two correlation matrices by augmenting the control vector used by minimization algorithm as described by Buehner (2005). Alternatively, a Householder transformation can be used to separate the analysis increment into the part that projects onto the subspace spanned by the sample of error realizations and the complementary subspace (Fisher 1998).

Efficient approaches for modelling spatial background-error correlations with various classes of functional forms in physical space have also been examined. Weaver and Courtier (2001) showed how the application of a diffusion operator can be used to efficiently implement spatial correlations that are generally Gaussian-shaped. Similarly, recursive filters have been used to model correlations efficiently, while partially relaxing the assumptions of homogeneity and isotropy (Derber and Rosati 1989; Wu et al. 2002; Purser et al. 2003).

### 3.3 Spectral/Physical-Space Representation

As described above, imposing the constraint that the correlations are to some extent local, either in the spatial or spectral domain is often necessary to obtain a useful covariance estimate. However, spectral localization has typically only been employed in the limiting case of the correlations being diagonal. Since diagonal correlations in spectral space correspond to homogeneous correlations in physical space, it is natural to consider what a more moderate amount of spectral localization would produce in physical space. This has been studied in the context of using a set of wavelets to define a space in which the correlations are assumed to be diagonal (Fisher and Andersson 2001; Deckmyn and Berre 2005; Pannekoucke et al. 2007) and also through explicit localization of correlations in spectral space (Buehner and Charron 2007). In both cases, the moderate localization of correlations in spectral space is shown to allow a certain amount of inhomogeneity of the correlations in physical space while still having a smoothing effect on the correlations. In effect, increasing amounts of spectral localization is equivalent with spatially averaging the local correlation functions in physical space over increasingly large areas (Buehner and Charron 2007). Of course, the limiting case where the spectral correlations become diagonal corresponds with an averaging of the local correlation functions over the entire domain. There likely exists an optimal level of combined spectral and spatial localization that depends on several factors, including the size of the sample of error realizations and the level of spatial inhomogeneity and typical spatial length scale of the true correlations. Some examples of the effect of spectral and spatial localization are shown in Sect. 4.

### 3.4 Theoretically-Based Correlation Modelling

The approaches discussed so far are mostly empirical approaches that rely on assumptions about the statistical properties of the background errors. In this section, examples of approaches for modelling error correlations are described that rely on theoretical assumptions concerning the dynamical properties of the errors. Due to their being based on dynamical relationships, the resulting correlations may be flow-dependent. Such approaches are often used to transform a set of intermediate variables that are assumed to have simpler (possibly stationary, homogeneous and isotropic) correlations into the actual analysis variables.

A common example of such theoretically-based correlation models involves a so-called balance operator to construct between-variable correlations. For example, in several operational NWP data assimilation systems the temperature and surface pressure increments are constructed by adding the increments of unbalanced and balanced components of these variables, where the latter is computed from the streamfunction (or vorticity) increment via a balance operator for geostrophy (e.g. Gauthier et al. 1998; Derber and Bouttier 1999). For example, the three-dimensional temperature increment field is computed using

$$\Delta \mathbf{T} = \Delta \mathbf{T}_{\mathrm{u}} + \mathbf{G}_{\mathrm{T}} \Delta \boldsymbol{\psi} \,, \tag{6}$$

where $\Delta \mathbf{T}_{\mathrm{u}}$ is the analysis increment of the unbalanced component of temperature and $\mathbf{G}_{\mathrm{T}}$ is the balance operator for obtaining the geostrophically balanced component of temperature from the streamfunction. This implies correlations between the wind and mass field increments that are consistent with geostrophy and the hydrostatic relationship. Additional balance operators are also usually employed to create correlations between the rotational and divergent components of the wind field near the surface (cf. Ekman balance). The result is that the background error covariance matrix is represented as a series of separate matrices or operators:

$$\mathbf{B} = \mathbf{G}\mathbf{V}^{1/2}\mathbf{C}_{\mathrm{u}}(\mathbf{V}^{1/2})^{T}\mathbf{G}^{T} \,, \tag{7}$$

where $\mathbf{C}_{\mathrm{u}}$ is the correlation matrix for the set of independent variables with horizontally homogeneous and isotropic correlations, $\mathbf{V}$ is a diagonal matrix containing the error variances, and $\mathbf{G}$, which includes $\mathbf{G}_{\mathrm{T}}$, transforms the unbalanced variables into the full quantities for temperature, surface pressure, and velocity potential (or divergence) using the balance operators. Consequently, the effective correlations in $\mathbf{B}$ are neither horizontally homogeneous nor isotropic due to the spatial dependence of the balance operators. The temperature correlations at the Equator are mostly determined by the correlations of the unbalanced temperature, whereas in the extra-tropics they are a combination of the unbalanced and balanced temperature correlations. In turn, the balanced temperature correlations are derived from the streamfunction (or vorticity) correlations. Figure 1 shows the fraction of temperature variance explained by a simple linear balance with streamfunction when either the NMC method or a Monte Carlo approach applied to a 3D-Var assimilation system is used to generate the error sample. Note that the temperature and wind fields are more strongly in balance when using the NMC method than when using the 6-h spread of background states from a Monte Carlo simulation. The results were obtained using the Canadian 3D-Var system described by Chouinard et al. (2001).

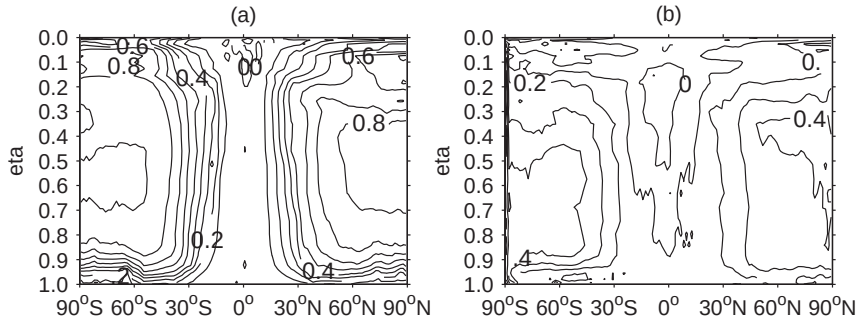



**Fig. 1** The zonally averaged ratio of balanced temperature variance normalized by the full temperature variance from the background error covariances estimated using (**a**) the NMC method and (**b**) a Monte Carlo approach applied to a 3D-Var assimilation system

If a linear balance is used to compute the balanced component of temperature, then the operator **G** does not depend on the flow. However, the use of the non-linear balance

$$\nabla^2 P_\mathrm{b} = -\nabla \cdot (\mathbf{v}_\mathrm{r} \cdot \nabla \mathbf{v}_\mathrm{r} + f\mathbf{k} \times \mathbf{v}_\mathrm{r}) \ , \tag{8}$$

where $P_\mathrm{b}$ is the balanced pressure variable and $\mathbf{v}_\mathrm{r}$ is the rotational wind vector, results in a balance operator that, when linearized with respect to the background state, is flow-dependent (Fisher 2003). Therefore, **G** depends on the background state itself and, therefore, so does the correlation structure in **B** of temperature in the extra-tropics. Similarly, to compute a more realistic balanced component of divergence, Fisher (2003) evaluated using the quasi-geostrophic omega equation, and Byrom and Roulstone (2003) examined using Richardson's equation.

Another approach for introducing a theoretically-based correlation model that is flow-dependent relies on the coordinate transformation described by Dee and Gaspari (1996). The idea is to take advantage of the efficiency of using a homogeneous and isotropic correlation model, but to apply it in a space with transformed spatial coordinates. This transformation can be chosen so that when transformed back into the original coordinate system, the resulting correlations are heterogeneous, anisotropic and possibly flow-dependent. They used a simple coordinate transform to obtain a latitudinal dependence of the horizontal correlations, that is, with a larger length-scale in the tropics than in the extra-tropics. Desroziers (1997) used a similar approach and a coordinate transformation based on semi-geostrophic theory to obtain more realistic correlations in the vicinity of frontal structures.

## 4 Illustrative Examples

In this section, several examples are shown to illustrate some of the approaches for estimating and modelling error statistics discussed previously. All have been taken from Canadian operational or experimental atmospheric data assimilation systems used for NWP.

### *4.1 Estimated Error Variances*

Figure 2 shows the estimated background error standard deviation (stdev) of temperature obtained from using the NMC method and the EnKF of Houtekamer et al. (2005). Note that the stdev field obtained with the NMC method is zonally invariant, because the original estimates have been zonally averaged. Without this averaging, the estimated stdev fields tend to have unrealistic spatial variations, with larger values downstream of well-observed areas and lower values near data sparse regions. Even though the EnKF can be used to estimate flow-dependent background error covariances for each analysis time, here the variances obtained from the ensemble

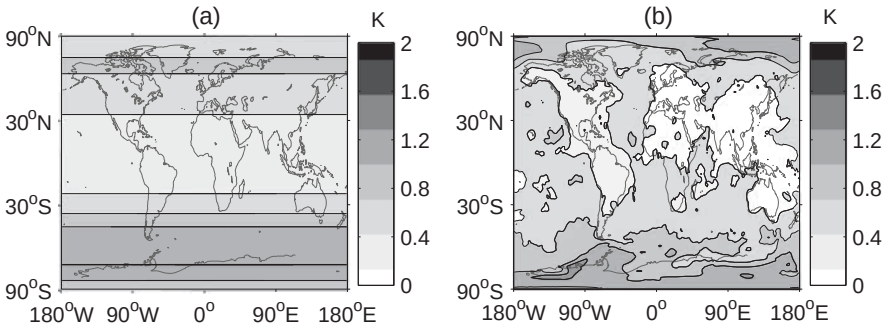



**Fig. 2** Estimated background error stdev of temperature near 500 hPa computed using: (**a**) the NMC method and (**b**) a temporal average of the background ensemble spread variances from the EnKF. Figure adapted from Buehner (2005)

spread of background states are temporally averaged over 1 month. When compared with the NMC method, the EnKF produces more realistic spatial variations with higher values over the oceans than over the continents.

## *4.2 Single Observation Experiments*

The analysis increment resulting from the assimilation of a single observation provides a partial view of the background error covariances by showing how information from the observation is distributed both spatially and among the different analysis variables. From the linear analysis equation, the analysis increment from assimilating a single observation is proportional to $\mathbf{BH}^T$, where $\mathbf{H}$ is reduced to a row vector. For observation types closely related to one of the variables represented in the background error covariances, the analysis increment is simply proportional to a column of $\mathbf{B}$. This is especially convenient when the background error covariances are modelled using a series of operators and therefore cannot be easily computed explicitly.

Non-stationary features such as strong horizontal gradients and regions of instability can significantly influence the background error statistics. As already discussed, the EnKF is able to capture this flow dependence. To demonstrate this, a single temperature observation 1 K greater than the background temperature near 900 hPa was assimilated within a strong near-surface temperature front that appeared over the North Pacific on 27 May 2002 at 1200 UTC (adapted from Buehner 2005). All analyses were performed with a variational analysis system using either the homogeneous and isotropic background error correlations estimated with the NMC method or the spatially localized ensemble correlations estimated from the output of the EnKF. The analysis increment produced using the background error covariances from the 3D-Var (Fig. 3a) is clearly unaffected by the local meteorological conditions (the background temperature is shown in the dark

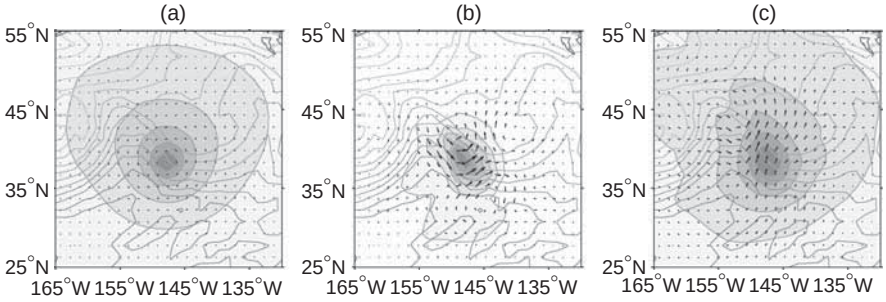



**Fig. 3** Analysis increment of temperature (*shaded contours*, contour increment of 0.15 K) and wind (vectors) from a single temperature observation near 900 hPa located in a strong near-surface temperature front at 1200 UTC, 27 May 2002. The experiment was performed using: (**a**) the homogeneous and isotropic background error correlations estimated with the NMC method in 3D-Var, (**b**) the spatially localized background error correlations estimated with the EnKF, and (**c**) the same background error covariances as the *first panel*, except in a 4D-Var system with the observation occurring 6 h after the beginning of the assimilation window. The background temperature field is shown as *black unshaded contours* with a contour interval 10 times larger than for the temperature increment

contours). The temperature increment decays in a nearly isotropic fashion away from the observation location and the wind increment is nearly zero at the location of the temperature observation. In contrast, when using covariances estimated from the EnKF ensemble of background states valid for the same analysis time (Fig. 3b), the temperature increment is slightly elongated along the front and the wind increment is larger with vectors oriented parallel with the background temperature gradient at the observation location.

Finally, a 4D-Var analysis was performed with the beginning of the assimilation window occurring 6 h before the time of the temperature observation. The same background error covariances are used as in 3D-Var, but in 4D-Var they are implicitly propagated throughout the assimilation window with the linearized version of the atmospheric forecast model. The result is an analysis increment (Fig. 3c) that is slightly modified relative to the result with 3D-Var. However, the change in the wind increment demonstrates that the covariance propagation has introduced qualitatively similar local correlations between temperature and wind as in the EnKF covariances such that the winds are again parallel to the background temperature gradient near the observation location.

The next series of examples demonstrate different approaches for modelling background error correlations (adapted from Buehner and Charron 2007).

Figure 4 shows the zonal cross-section of the meridional wind analysis increment from using 3D-Var to assimilate a single zonal wind observation located over the southern Pacific ocean at 60°S, 180°E and 300 hPa. The background error covariances were estimated using a Monte Carlo simulation approach applied to a 3D-Var assimilation system. The error sample was obtained from differences

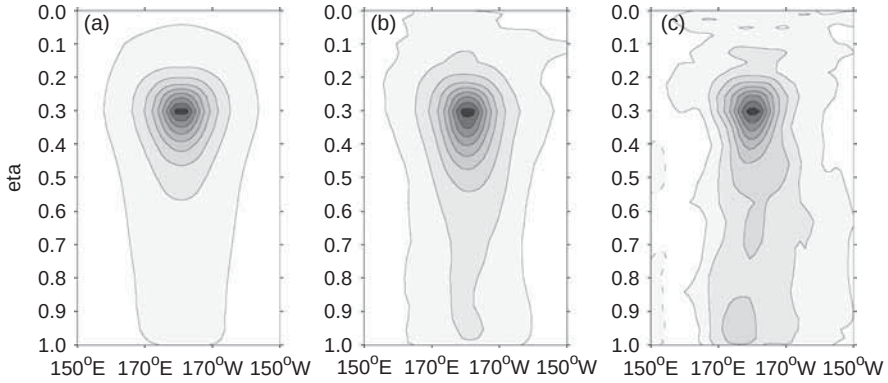



**Fig. 4** Zonal cross-sections of the zonal wind covariances with respect to 60° S, 180° E and 300 hPa with: (**a**) homogeneous and isotropic correlations, (**b**) correlations with spectral localization, and (**c**) correlations with no localization. The covariances are normalized to have a maximum value of one, the contour interval is 0.1 and *dashed contours* denote negative values

between the background states from an assimilation experiment that assimilated perturbed observations and an experiment employing unperturbed observations. The error realizations were pooled over a 1-month period to estimate the covariances.

For the first result, the correlations were modelled as being horizontally homogeneous and isotropic, that is, diagonal in spectral space (Fig. 4a). In the second result, the same correlations were used after applying only a moderate amount of spectral localization (Fig. 4b). Finally, the sample estimate of the correlations with no localization was used to produce the third result (Fig. 4c). Note how the use of horizontally homogeneous and isotropic correlations produces a spatially smooth covariance structure. Conversely, when no localization is applied, the covariances are quite noisy. When spectral localization is applied with a localization radius that sets to zero correlations with a difference in total wavenumber greater than 10, the covariance structure is slightly more noisy than with the diagonal spectral correlations, but significantly smoother than when no localization is applied. With no localization, the correlation structure is sharper in the zonal direction and broader in the vertical direction relative to the homogeneous correlations. The spectrally localized correlations appear to also exhibit this difference with the homogeneous correlation, although to a lesser degree.

Figure 5 shows the same type of result as Fig. 4, except the location is at the Equator. Again, the spectrally localized correlations result in a spatially smoother covariance structure (Fig. 5b) than the raw sample estimate (Fig. 5c), but slightly noisier than when employing horizontally homogeneous and isotropic correlations (Fig. 5a). For this location the homogeneous correlations again differ from the previous results. Now the correlations with spectral localization or no localization both exhibit a correlation structure that is broader in the zonal direction and sharper in the vertical direction. This is in the opposite sense compared to the

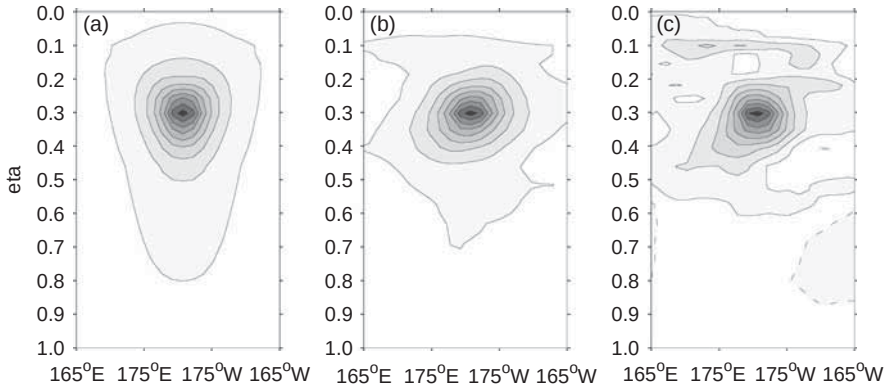



**Fig. 5** Same as Fig. 4, but for a location at the Equator

covariances at 60°S and, presumably, is a robust result related to differences in the extra-tropical versus tropical atmospheric dynamics. Similar latitudinal variations were also demonstrated by Ingleby (2001).

## 5 Summary

This chapter has provided an overview of the relevant issues and common approaches used for estimating and modelling the error covariances required for data assimilation, with an emphasis on approaches used for NWP. The discussion of error covariance estimation highlighted the fundamental theoretical limitation encountered when trying to estimate the covariances of both the observations and the background state from a single quantity, the innovation (that is, the difference between the observations and the background state projected into observation space). This limitation necessitates the introduction of external assumptions and the different approaches described vary with respect to the assumptions adopted. Examples of these include assuming the observation errors are spatially uncorrelated (allowing the variances and background error correlations to be estimated) or assuming the observation and model error statistics are known (allowing the background error covariances to be estimated with a Monte Carlo technique). The chapter *Evaluation of Assimilation Algorithms* (Talagrand) provides further details.

Approaches for modelling the error covariances, especially of the background error, must be computationally feasible, in terms of both memory and time limitations. Due to the high dimensionality of the problem and a lack of sufficient observation data to explicitly estimate and use the complete covariances, assumptions also must be employed regarding the structure of the error covariances. The most common assumptions are that the spatial correlations are either partially or completely horizontally homogeneous (possibly for a set of transformed analysis

variables) and/or that they are to some extent spatially local such that the correlations become zero at a specified distance. By employing one or both of these assumptions, a robust estimate of the covariances can usually be obtained with a relatively small sample of error realizations and used within a data assimilation system. However, the quality of the resulting analysis depends on how well the imposed assumptions decrease the sampling error in the covariance estimate while preserving the essential aspects of the covariances.

Current research on data assimilation for application to NWP is generally focused on two approaches: the variational approach, namely 3D-Var and 4D-Var, and variations of the EnKF. Typically, applications of these two approaches employ very different assumptions regarding the estimation and modelling of background error covariances. Variational approaches commonly use temporally static covariances with horizontally homogeneous and isotropic correlations (for a specific set of transformed analysis variables) with theoretically-based balance relationships and estimated with an ad hoc method. Applications of the EnKF use time-dependent covariances estimated from an ensemble of model states where usually the only assumption is that the spatial correlations are to some extent local. It is interesting to note that despite the large differences in the resulting background error covariances employed by each approach, both can produce analyses of comparable quality (as of yet unpublished results presented at "WMO-sponsored workshop on 4D-Var and EnKF inter-comparisons": http://4dvarenkf.cima.fcen.uba.ar). This suggests that an in-depth comparison of the way background error covariances are estimated and modelled in applications of the two approaches may help identify which aspects of each are most beneficial with respect to analysis quality. By combining aspects of each approach, it is possible that new approaches for estimating and modelling background error covariances may be obtained that result in better analyses than those produced by either of the original two approaches.

## References

Belo Pereira, M. and L. Berre, 2006. The use of an ensemble approach to study the background error covariances in a global NWP model. *Mon. Weather Rev.*, **134**, 2466–2489.

Bormann, N., S. Saarinen, G. Kelly and J.-N. Thépaut, 2003. The spatial structure of observation errors in atmospheric motion vectors from geostationary satellite data. *Mon. Weather Rev*., **131**, 706–718.

Buehner, M., 2005. Ensemble-derived stationary and flow-dependent background error covariances: Evaluation in a quasi-operational NWP setting. *Q. J. R. Meteorol. Soc.*, **131**, 1013–1044.

Buehner, M. and M. Charron, 2007. Spectral and spatial localization of background-error correlations for data assimilation. *Q. J. R. Meteorol. Soc.*, **133**, 615–630.

Buehner, M., P. Gauthier and Z. Liu, 2005. Evaluation of new estimates of background and observation error covariances for variational assimilation. *Q. J. R. Meteorol. Soc.*, **131**, 3373–3383.

Burgers, G., P.J. Van Leeuwen and G. Evensen, 1998. Analysis scheme in the ensemble Kalman filter. *Mon. Weather Rev.*, **126**, 1719–1724.

Byrom, M. and I. Roulstone, 2003. Calculating vertical motion using Richardson's equation. In *Proceedings of the ECMWF/GEWEX Workshop on Humidity Analysis,* Reading, UK, 8–11 July 2002, pp 49–58.

Chapnik, B., G. Desroziers, F. Rabier and O. Talagrand, 2004. Properties and first application of an error-statistics tuning method in variational assimilation. *Q. J. R. Meteorol. Soc.*, **130**, 2253–2275.

Chapnik, B., G. Desroziers, F. Rabier and O. Talagrand, 2006. Diagnosis and tuning of observation error statistics in a quasi operational data assimilation setting. *Q. J. R. Meteorol. Soc.*, **132**, 543–565.

Chouinard, C., C. Charette, J. Hallé, P. Gauthier, J. Morneau and R. Sarrazin, 2001.The Canadian 3D-var analysis scheme on model vertical coordinate. In *Proceedings of the 18th AMS Conference on Weather Analysis and Forecasting,* Fort Lauderdale, USA, 30 July–2 August, pp 14–18.

Courtier, P., E. Andersson, W. Heckley, J. Pailleux, D. Vasiljević, M. Hamrud, A. Hollingsworth, F. Rabier and M. Fisher, 1998. The ECMWF implementation of three-dimensional variational assimilation (3D-Var). I: Formulation. *Q. J. R. Meteorol. Soc.*, **124**, 1783–1807.

Daley, R., 1991. *Atmospheric Data Analysis*, Cambridge University Press, Cambridge, UK, 457 pp.

Deckmyn, A. and L. Berre, 2005. A wavelet approach to representing background error covariances in a limited-area model. *Mon. Weather Rev.*, **133**, 1279–1294.

Dee, D.P., 1995. On-line estimation of error covariance parameters for atmospheric data assimilation. *Mon. Weather Rev.*, **123**, 1128–1145.

Dee, D.P. and G. Gaspari, 1996. Development of anisotropic correlation models for atmospheric data assimilation. Preprints, *11th Conference on Numerical Weather Prediction*. Norfolk, VA, American Meteor Society., pp. 249–251.

Derber, J. and F. Bouttier, 1999. A reformulation of the background error covariance in the ECMWF global data assimilation system. *Tellus*, **51A**, 195–221.

Derber, J. and A. Rosati, 1989. A global oceanic data assimilation system. *J. Phys. Oceanogr.*, **19**, 1333–1347.

Desroziers, G., 1997. A coordinate change for data assimilation in spherical geometry of frontal structures. *Mon. Weather Rev.*, **125**, 3030–3038.

Desroziers, G., L. Berre, B. Chapnik and P. Poli, 2005. Diagnosis of observation, background and analysis-error statistics in observation space. *Q. J. R. Meteorol. Soc.*, **131**, 3385–3396.

Desroziers, G. and S. Ivanov, 2001. Diagnosis and adaptive tuning of observation-error parameters in a variational assimilation. *Q. J. R. Meteorol. Soc.*, **127**, 1433–1452.

Evensen, G., 1994. Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *J. Geophys. Res.*, **99**, 10143–10162.

Fisher, M., 1998. Development of a Simplified Kalman Filter, *ECMWF Research Department Technical Memorandum*, **260**. ECMWF, Shinfield Park, Reading .

Fisher, M., 2003. Background error covariance modelling. In *Proceedings of the ECMWF Seminar on Recent Developments in Data Assimilation for Atmosphere and Ocean*, Reading, UK, 8–12 September, 2003, pp 45–63.

Fisher, M. and E. Andersson, 2001. Developments in 4D-Var and Kalman filtering, *ECMWF Research Department Technical Memorandum*, **347**. Reading, UK.

Garand, L., S. Heilliette and M. Buehner, 2007. Inter-channel error correlation associated with AIRS radiance observations: Inference and impact in data assimilation. *J. Appl. Meteor. Climat.*, **46**, 714–725.

Gaspari, G. and S. Cohn, 1999. Construction of correlation functions in two and three dimensions. *Q. J. R. Meteorol. Soc.*, **125**, 723–757.

Gauthier, P., M. Buehner and L. Fillion, 1998. Background-error statistics modelling in a 3D variational data assimilation scheme. In *Proceedings of the ECMWF Workshop on Diagnosis of Data Assimilation Systems*, November 2–4, 1998. Reading, UK, pp 131–145.

Gelb, A. (ed.), 1974. *Applied Optimal Estimation*, MIT Press, Cambridge, USA, 382 pp.

Hamill, T.M. and C. Snyder, 2000. A hybrid ensemble Kalman filter – 3D variational analysis scheme. *Mon. Weather Rev*., **128**, 2905–2919.

Hamill, T.M., J.S. Whitaker and C. Snyder, 2001. Distance-dependent filtering of background error covariance estimates in an ensemble Kalman filter. *Mon. Weather Rev.*, **129**, 2776–2790.

Hello, G. and F. Bouttier, 2001. Using adjoint sensitivity as a local structure function in variational data assimilation. *Nonlinear Process. Geophys.*, **8**, 347–355.

Hollingsworth, A. and P. Lönnberg, 1986. The statistical structure of short-range forecast errors as determined from radiosonde data. Part I: The wind field. *Tellus*, **38A**, 111–136.

Houtekamer, P.L., L. Lefaivre, J. Derome, H. Ritchie and H.L. Mitchell, 1996. A system simulation approach to ensemble prediction. *Mon. Weather Rev.*, **124**, 1225–1242.

Houtekamer, P.L. and H.L. Mitchell, 2001. A sequential ensemble Kalman filter for atmospheric data assimilation. *Mon. Weather Rev*., **129**, 123–137.

Houtekamer, P.L., H.L. Mitchell, G. Pellerin, M. Buehner, M. Charron, L. Spacek and B. Hansen, 2005. Atmospheric data assimilation with an ensemble Kalman filter: Results with real observations. *Mon. Weather Rev.*, **133**, 604–620.

Ingleby, B.N., 2001. The statistical structure of forecast errors and its representation in the Met. Office global 3-D variational data assimilation scheme. *Q. J. R. Meteorol. Soc.*, **127**, 209–231.

Järviven, H., E. Andersson and F. Bouttier, 1999. Variational assimilation of time sequences of surface observations with serially correlated errors. *Tellus*, **51A**, 469–488.

Lorenc, A.C., 2003. The potential of the ensemble Kalman filter for NWP – a comparison with 4D-Var. *Q. J. R. Meteorol. Soc.*, **129**, 3183–3203.

Mitchell, H.L. and P.L. Houtekamer, 2000. An adaptive ensemble Kalman filter. *Mon. Weather Rev*., **128**, 416–433.

Pannekoucke, O., L. Berre and G. Desroziers, 2007. Filtering properties of wavelets for the local background error correlations. *Q. J. R. Meteorol. Soc*., **133**, 363–379.

Parrish, D.F. and J.C. Derber, 1992. The national meteorological center's spectral statistical interpolation analysis system. *Mon. Weather Rev*., **120**, 1747–1763.

Phillips, N.A., 1986. The spatial statistics of random geostrophic modes and first-guess errors. *Tellus*, **38A**, 314–332.

Purser, R.J., W.S. Wu, D.F. Parrish and N.M. Roberts, 2003. Numerical aspects of the application of recursive filters to variational statistical analysis; Part II: Spatially inhomogeneous and anisotropic general covariances. *Mon. Weather Rev.*, **191**, 1536–1548.

Rabier, F., A. McNally, E. Andersson, P. Courtier, P. Undén, J. Eyre, A. Hollingsworth and F. Bouttier, 1998. The ECMWF implementation of three dimensional variational assimilation (3D-Var). Part II: Structure functions. *Q. J. R. Meteorol. Soc*., **124**, 1809–1830.

Talagrand, O., 1999. A posteriori evaluation and verification of analysis and assimilation algorithms. In *Proceedings of the Workshop on Diagnosis of Data Assimilation Systems*, Reading, UK, November 2–4, 1999. pp 17–28.

Weaver, A.T. and P. Courtier, 2001. Correlation modelling on the sphere using a generalized diffusion equation. *Q. J. R. Meteorol. Soc*., **127**, 1815–1846.

Wu, W.-S., R.J. Purser and D.F. Parrish, 2002. Three-dimensional variational analysis with spatially inhomogeneous covariances. *Mon. Weather Rev*., **130**, 2905–2916.

# Bias Estimation

**Richard Ménard**

## 1 Introduction

One of the standard assumptions in data assimilation is that observation and model errors are purely random, i.e., they do not contain systematic errors (see chapter *Mathematical Concepts of Data Assimilation*, Nichols). In reality, the distinction between random errors and systematic errors is somewhat academic. Because of non-linearities and the complexity in the different processes involved, model errors and observation errors arise both as random and systematic. Model errors originate from parametrizations, unrepresented model physics, inaccurate boundary forcing, and resolution, among others sources. With satellite observations, the forward model often gives rise to large systematic errors. Conventional observations can also be contaminated by missing or inadequate representation of physical processes. Removing systematic errors from observations or models requires considerable effort and is made, basically, by improving the representation of physical processes involved. As such, it is never complete.

Data assimilation schemes built on the standard assumptions that the errors are purely random, cannot produce analyses with no bias if either observations or the model have systematic errors – no matter how the error variances are specified (Dee and da Silva 1998). The problem of dealing with biases is thus unavoidable.

The detection of bias is made by comparing models or observations with independent data that are trusted as accurate and unbiased. This comparison is best made when spatial and temporal co-location is used. Then, from those residuals and with appropriate modelling assumptions, a model representation of the bias can be obtained and bias correction can be applied.

In this chapter we are going one step further by considering bias estimation and correction as an integral part of data assimilation. From the point of view of estimation theory, combining bias and state estimation is performed by using an augmented system where bias parameters are added to the state vector. Although this may sound

R. Ménard (✉)
Air Quality Research Division, Environment Canada, Dorval, Canada
e-mail: richard.menard@ec.gc.ca

W. Lahoz et al. (eds.), *Data Assimilation*, DOI 10.1007/978-3-540-74703-1_6,

simple, the proper mathematical formulation of this problem has been obtained only recently. In earlier attempts the problems of bias in observations and bias in models were dealt separately. With recent advances in the theory, it is now possible to formulate the complete problem of estimation in presence of bias. This theory will be presented in Sect. 3. The application of these schemes have also made encouraging progress: First, by implementing such schemes in an operational data assimilation system, the real issues of implementation and the determination of the input information is now being investigated; Second, as the theory has progressed, a better understanding of these schemes is obtained which, in turn, has provided further insights for its application.

Before we turn to the main development of this chapter, it may be useful to give some clarification about the terminology. Bias in data assimilation broadly refers to the presence of systematic errors, while in statistics it is a property of an estimator. Specifically, in statistics we say that $\hat{x}$ is an unbiased estimator of $x$, if $\mathcal{E}[\hat{x}\,|x] = x$ when $x$ is deterministic or $\mathcal{E}[\hat{x}] = \mathcal{E}[x]$ when $x$ is stochastic. We can reconcile the statistical definition with its usage in data assimilation by considering that observations, model forecasts and analyses all aim at determining the true state, and in that sense and broadly speaking, they can be considered as estimators of the true state. Biases in that context refer to the mean differences between the estimator and true state, i.e., the systematic error.

## 2 Detection of Bias

Although biases are usually diagnosed by comparison with independent and trusted unbiased data sets, and for models by forecast drift, the question arises "How do we detect biases in a data assimilation cycle?" This section will address specifically this issue. The detection of bias has been discussed at length and presented in several applications by Dee (2005), for which we owe much of the discussion presented here.

### *2.1 Bias Detection Using Innovations*

Statistics of observed-minus-background residuals (also called innovations) provide information on systematic errors in model and observations. Routine monitoring of observations-minus-background residuals in operational assimilation centres provides a wealth of information on the biases and performance of the assimilation system. Non-zero-mean residuals (see chapter *Mathematical Concepts of Data Assimilation*, Nichols, for notation):

$$\left\langle \mathbf{y} - \mathbf{H}\mathbf{x}^f \right\rangle = \left\langle \boldsymbol{\varepsilon}^o \right\rangle - \mathbf{H}\left\langle \boldsymbol{\varepsilon}^f \right\rangle = \mathbf{b}^o - \mathbf{H}\mathbf{b}^f, \tag{1}$$

indicate the presence of bias in either model forecast or observations (or both) but cannot identify the source. However, a closer analysis of the residuals can reveal the source of bias. For example, an abrupt change in a particular channel or observation is indicative of instrument malfunction. An objective method to detect artificial and local changes in an observation network is the standard normal homogeneity test (SNHT; Alexandersson and Moberg 1997). The idea behind this method is that natural changes are similar in time series at different stations, whereas artificial irregularities are site-specific. This method was applied to diagnose problems with the radiosonde network that occurred in the ERA-40 reanalysis (Haimberger 2007). Time series of the analysis of observed-minus-background residuals are also useful as they can reveal sources of bias. Such an analysis applied to the radiosonde temperatures in the NCEP (National Centers for Environmental Prediction) global assimilation system revealed excessive power in periods longer than 10 days, as well as a strong peak in the diurnal cycle which pointed to model underestimation of mean surface diurnal temperature variations (Dee 2005). The inspection of observed-minus-background residuals is also useful for revealing biases in radiative transfer models. Saunders (2005) investigated the origin of systematic errors by looking at biases of observed-minus-background residuals in radiation spectral space for the AIRS (Atmospheric InfraRed Sounder – see *Appendix* for a list of acronyms) instrument. As different bands and wavelengths are associated with different gases, different aspects of the spectroscopy and its modelling, insights on problems with the radiative transfer modelling can thus be obtained.

## *2.2 Bias Detection Using Analysis Increments*

Analysis-minus-forecast residuals, called analysis increments, also provide information on systematic errors. Using a *BLUE* (Best Linear Unbiased Estimate), the average analysis increment is

$$\left\langle \mathbf{x}^a - \mathbf{x}^f \right\rangle = \left\langle \mathbf{K}(\boldsymbol{\varepsilon}^o - \mathbf{H}\boldsymbol{\varepsilon}^f) \right\rangle \tag{2}$$

in fact, closely related to mean observed-minus-background residuals or mean innovations. It can be argued that if the averaging procedure (e.g. zonal time-mean) used to obtain the observation and background error statistics is the same as that used to compute the analysis increments, and if the observation network is fairly uniform in the averaging sense (e.g. an observation network that is zonally uniform and regular in time, and is represented by zonal time-mean statistics), then the gain matrix $\mathbf{K}$ can be factored out

$$\left\langle \mathbf{x}^a - \mathbf{x}^f \right\rangle \approx \mathbf{K}\left\langle \boldsymbol{\varepsilon}^o - \mathbf{H}\boldsymbol{\varepsilon}^f \right\rangle = \mathbf{K}(\mathbf{b}^o - \mathbf{H}\mathbf{b}^f). \tag{3}$$

The average analysis increment gives, however, the false impression that it provides bias information on the model space, whereas in essence it only contains bias

information in observation space. Moreover, wrong conclusions about the bias can occur if the Kalman gain is somehow erroneous, by projecting erroneous increments away from the observation locations (see Polavarapu et al. 2005 for further discussion). Nevertheless, the average analysis increments provide a useful tool to collectively assemble the biases (obtained in observation space) onto the model space. We remark also that the analysis increment is one of the two components of the *Data-minus-Analysis* (DmA) vector (see chapter *Evaluation of Assimilation Algorithms*, Talagrand). As noted in this chapter by Talagrand there is a one-to-one correspondence between *DmA* and *OmF* (the innovations), so that basically these quantities are equivalent. The diagnostic based on analysis increments is thus chosen as a matter of convenience.

Figure 1 shows the zonal monthly mean analysis increments of temperature from two assimilation systems. The left panel shows the result produced from the ERA-40 reanalysis for the month of August 2002, and the right panel from the Canadian GEM-BACH model (Ménard et al. 2007) for a similar time period (September 2003) but with no observation bias correction on the AMSU-A stratospheric channels 11–14. Strong biases of slightly over 1 K and of alternating signs are noted in the stratospheric polar and tropical regions.

Although the mean analysis increments indicate the presence of large biases in the stratosphere, their origin is unclear. In free running mode, models are known to have large systematic errors in the stratosphere. The main source of stratospheric data in the ERA-40 reanalysis is TOVS/ATOVS (left panel, Fig. 1). The assimilated

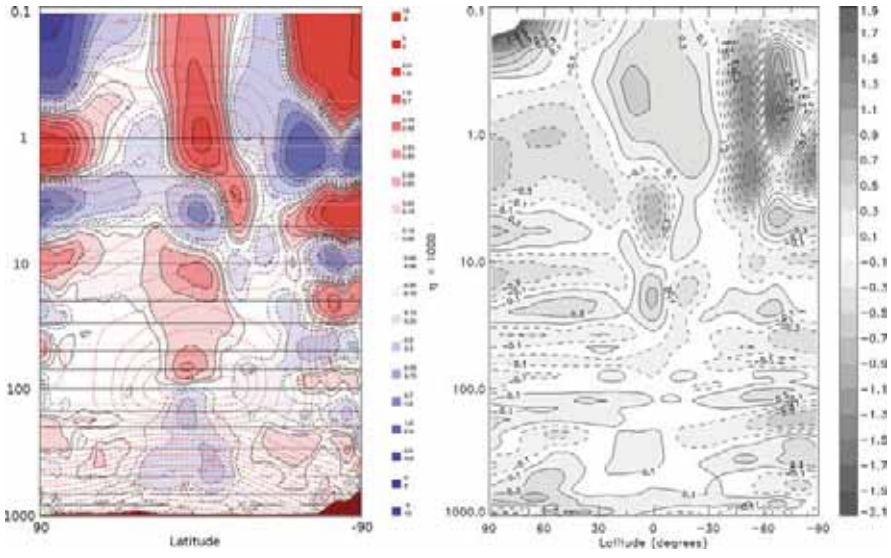



**Fig. 1** Zonal mean time-averaged temperature increments. *Left panel*, from ERA-40 reanalysis for August 2002. *Right panel*, from the Canadian model GEM-BACH without AMSU bias correction on channels 11–14 (Reproduced from Dee 2005; © Royal Meteorological Society)

radiances have been corrected for biases by correcting for scan angle systematic errors and air-mass dependence using an off-line procedure (see Sect. 4). To account for state-dependent systematic errors in radiative transfer calculations, a regression of the residuals with model layer-mean temperatures is made. Lacking unbiased observations in the stratosphere to anchor the stratospheric analyses, it is possible, therefore, that model biases and observations become interdependent. The striking similarity of the bias patterns with those of the right panel in Fig. 1, where a different model was used and no radiance correction was applied on the stratospheric channels (compare the left and right panels in Fig. 1), suggests that analysis increments originate primarily from observation bias and that the observation bias correction and the model temperature bias become interdependent in the stratosphere.

## 3 Bias Analysis

We have seen that the basic information from which biases can be estimated arises from innovations. In this section we derive the analysis equation following Lea et al. (2008) where for the first time both observation and model have biases. As in Lea et al. (2008) the derivation contains both variational and sequential formulations, but to simplify the development we do not address the issue of representativeness error. The derivation uses simple and clear assumptions in a Bayesian formulation. To apply this method requires, however, some knowledge about the model and observation bias characteristics as well as knowledge of the bias error covariances – which, in the current state of knowledge, we are severely lacking. The ability to distinguish the model bias and observation bias from the innovation information needs to be developed. We hope, nevertheless, that having the problem well posed to begin with, will help make further steps in this important problem for data assimilation.

To set the stage, let us introduce the equations for the state, the measurement, the model bias, and the observation bias,

$$\begin{aligned} \mathbf{x}^f &= \mathbf{x}^t + \mathbf{e}^t + \boldsymbol{\varepsilon}^f \\ \mathbf{y} &= \mathcal{H}(\mathbf{x}^t) + \mathbf{b}^t + \boldsymbol{\varepsilon}^o \\ \mathbf{e}^f &= \mathbf{e}^t + \boldsymbol{\varepsilon}^q \\ \mathbf{b}^f &= \mathbf{b}^t + \boldsymbol{\varepsilon}^b. \end{aligned} \tag{4}$$

The parameters on the left hand side of the equations, the forecast $\mathbf{x}^f$, the observation $\mathbf{y}$, the model bias forecast or model bias prior $\mathbf{e}^f$, and the observation bias forecast or observation bias prior $\mathbf{b}^f$ are known. $\mathcal{H}()$ is the non-linear observation operator. The true state $\mathbf{x}^t$, the (true) model bias $\mathbf{e}^t$ and (true) observation bias $\mathbf{b}^t$ are to be estimated, and the epsilon ($\boldsymbol{\varepsilon}$) variables represent zero-mean normally-distributed errors associated with each variable: $\boldsymbol{\varepsilon}^f$ is the forecast (random) error with covariance $\mathbf{P}^f$; $\boldsymbol{\varepsilon}^o$ is the observation (random) error with covariance $\mathbf{R}$; $\boldsymbol{\varepsilon}^q$ is

the model (random) error with covariance **Q**; and $\boldsymbol{\varepsilon}^b$ is the random error of the observation bias with covariance **S**.

## *3.1 Variational Formulation*

Following Lea et al. (2008) we make three fundamental assumptions, which permit us to well-pose the analysis equation in presence of both observation and model error biases. For convenience, we have dropped the superscript $t$ to denote the truth in this subsection:

1. The observation **y** is independent of the model bias **e**. If $p$ denotes the (conditional) probability density function, then

$$p(\mathbf{y}|\mathbf{x},\mathbf{b},\mathbf{e}) = p(\mathbf{y}|\mathbf{x},\mathbf{b}); \tag{5}$$

2. The model state **x** is independent of the observation bias **b**, i.e.,

$$p(\mathbf{x}|\mathbf{b},\mathbf{e}) = p(\mathbf{x}|\mathbf{e}); \tag{6}$$

3. The model bias is independent of the observation bias, that is

$$p(\mathbf{b},\mathbf{e}) = p(\mathbf{b}|\mathbf{e})\,p(\mathbf{e}) = p(\mathbf{e}|\mathbf{b}\,)\,p(\mathbf{b}) = p(\mathbf{b})\,p(\mathbf{e}). \tag{7}$$

In this general context, the analysis consists in finding the maximum a posteriori estimate of the state, the observation bias and the model bias, given the observations and any prior knowledge of the state and biases. The starting point is the calculation of the conditional probability density function $p(\mathbf{x},\mathbf{b},\mathbf{e}|\mathbf{y})$. Using Bayes' theorem, we have

$$p(\mathbf{x},\mathbf{b},\mathbf{e}|\mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{x},\mathbf{b},\mathbf{e})\,p(\mathbf{x},\mathbf{b},\mathbf{e})}{p(\mathbf{y})} \tag{8}$$

According to assumption 1 (Eq. 5), the first factor in the numerator simplifies to

$$p(\mathbf{y}|\mathbf{x},\mathbf{b},\mathbf{e}) = p(\mathbf{y}|\mathbf{x},\mathbf{b}). \tag{9}$$

Using again Bayes' theorem, the second factor in the numerator can be re-written as

$$p(\mathbf{x},\mathbf{b},\mathbf{e}) = p(\mathbf{x}|\mathbf{b},\mathbf{e})\,p(\mathbf{b}|\mathbf{e})\,p(\mathbf{e}), \tag{10}$$

and using assumption 2 (Eq. 6), and assumption 3 (Eq. 7), this simplifies to

$$p(\mathbf{x},\mathbf{b},\mathbf{e}) = p(\mathbf{x}|\mathbf{e})\,p(\mathbf{b})\,p(\mathbf{e}), \tag{11}$$

so that the posterior probability density function is

$$p(\mathbf{x},\mathbf{b},\mathbf{e}|\mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{x},\mathbf{b})\,p(\mathbf{x}|\mathbf{e})\;p(\mathbf{b})\,p(\mathbf{e})}{p(\mathbf{y})}. \tag{12}$$

Assuming a normal distribution, the probability densities are expressed as:

$$p(\mathbf{y}\,|\mathbf{x},\mathbf{b}\,) = \mathcal{N}(\mathcal{H}(\mathbf{x})+\mathbf{b},\mathbf{R}) = \alpha_1\,\exp\left[-\frac{1}{2}(\mathbf{y}-\mathcal{H}(\mathbf{x})-\mathbf{b})^T\mathbf{R}^{-1}(\mathbf{y}-\mathcal{H}(\mathbf{x})-\mathbf{b})\right], \tag{13}$$

where $\alpha_1$ is a constant that does not depend on any of the estimated variables, **x**, **b**, or **e**, and similarly

$$p(\mathbf{x}\,|\mathbf{e}\,) = \mathcal{N}(\mathbf{x}^f-\mathbf{e},\mathbf{P}^f) = \alpha_2\,\exp\left[-\frac{1}{2}(\mathbf{x}-\mathbf{x}^f+\mathbf{e})^T[\mathbf{P}^f]^{-1}(\mathbf{x}-\mathbf{x}^f+\mathbf{e})\right], \tag{14}$$

$$p(\mathbf{b}) = \mathcal{N}(\mathbf{b}^f,\mathbf{S}) = \alpha_3\exp\left[-\frac{1}{2}(\mathbf{b}-\mathbf{b}^f)^T\,\mathbf{S}^{-1}(\mathbf{b}-\mathbf{b}^f)\right], \tag{15}$$

$$p(\mathbf{e}) = \mathcal{N}(\mathbf{e}^f,\mathbf{Q}) = \alpha_4\,\exp\left[-\frac{1}{2}(\mathbf{e}-\mathbf{e}^f)^T\mathbf{Q}^{-1}(\mathbf{e}-\mathbf{e}^f)\right]. \tag{16}$$

The probability density $p(\mathbf{y})$ does not depend on any of the estimated parameters **x**, **b** or **e**, but only on their priors. Maximizing the a posteriori probability is equivalent to minimizing the following cost function (this is quadratic if the observation operator is linear):

$$\begin{aligned} J(\mathbf{x},\mathbf{b},\mathbf{e}) \;=\; & \frac{1}{2}(\mathbf{y}-\mathcal{H}(\mathbf{x})-\mathbf{b})^T\mathbf{R}^{-1}(\mathbf{y}-\mathcal{H}(\mathbf{x})-\mathbf{b}) \\ & + \frac{1}{2}(\mathbf{x}-\mathbf{x}^f+\mathbf{e})^T[\mathbf{P}^f]^{-1}(\mathbf{x}-\mathbf{x}^f+\mathbf{e}) \\ & + \frac{1}{2}(\mathbf{b}-\mathbf{b}^f)^T\,\mathbf{S}^{-1}(\mathbf{b}-\mathbf{b}^f) \\ & + \frac{1}{2}(\mathbf{e}-\mathbf{e}^f)^T\,\mathbf{Q}^{-1}(\mathbf{e}-\mathbf{e}^f). \end{aligned} \tag{17}$$

A remark is worth making with regard to Eq. (14) and the resulting cost function given by Eq. (17). In a dynamically evolving system, the forecast is not independent of the model bias since it depends on the model bias in the previous time step. A cross covariance between **x** and **e** should be introduced accordingly. A better approach is to account for the time dependence in the estimation problem and introduce the model bias as a tendency on the state. For the purpose of this derivation, we will neglect the cross-covariance term. This issue will be treated later in this chapter. We should also remark that assumption (6) should not be confused with the fact that **b** usually depends on **x**. Assumption (6) only says that the true state does not depend on the observation bias. The dependence of **b** on **x** can introduce a cross-covariance term in the cost function. It is possible however, to avoid such a term by making **b** depend on $\mathbf{x}^f$ rather than $\mathbf{x}^t$ which for all practical purposes should be sufficient (one should consult the Appendix in Ménard et al. 2000 for an example).

### 3.2 Sequential Formulation

The equations for the sequential formulation are found by setting to zero the partial derivatives of $J$ with respect to $\mathbf{x}$, $\mathbf{b}$, and $\mathbf{e}$. When $\mathcal{H}$ is linear we obtain a set of coupled linear equations,

$$[(\mathbf{P}^f)^{-1} + \mathbf{H}^T\mathbf{R}^{-1}\mathbf{H}]\,\hat{\mathbf{x}} + \mathbf{H}^T\mathbf{R}^{-1}\hat{\mathbf{b}} + (\mathbf{P}^f)^{-1}\,\hat{\mathbf{e}} = (\mathbf{P}^f)^{-1}\mathbf{x}^f + \mathbf{H}^T\mathbf{R}^{-1}\mathbf{y} \tag{18}$$

$$\mathbf{R}^{-1}\mathbf{H}\,\hat{\mathbf{x}} \quad + (\mathbf{R}^{-1} + \mathbf{S}^{-1})\,\hat{\mathbf{b}} \quad = \mathbf{R}^{-1}\mathbf{y} + \mathbf{S}^{-1}\mathbf{b}^f \tag{19}$$

$$(\mathbf{P}^f)^{-1}\hat{\mathbf{x}} \quad + [(\mathbf{P}^f)^{-1} + \ \mathbf{Q}^{-1}]\,\hat{\mathbf{e}} = (\mathbf{P}^f)^{-1}\,\mathbf{x}^f + \mathbf{Q}^{-1}\mathbf{e}^f \tag{20}$$

where $\hat{\mathbf{x}}$, $\hat{\mathbf{b}}$, $\hat{\mathbf{e}}$ are the estimated or analysis values, and $\mathbf{H}$ is the Jacobian of the observation operator (see chapter *Mathematical Concepts of Data Assimilation*, Nichols); $\mathbf{H}^T$ is the transpose of $\mathbf{H}$. Multiplying Eq. (19) by $\mathbf{H}^T$ and using Eqs. (18) and (20) we can eliminate $\hat{\mathbf{x}}$ from the system and we get

$$\mathbf{H}^T\mathbf{S}^{-1}(\hat{\mathbf{b}} - \mathbf{b}^f) + \mathbf{Q}^{-1}(\hat{\mathbf{e}} - \mathbf{e}^f) = 0. \tag{21}$$

Using the matrix inversion lemma or the Sherman-Morrison-Woodbury formula (see, for instance, Lewis et al. 2006), Eq. (18) can be rewritten as

$$\hat{\mathbf{x}} = \mathbf{x}^f - \hat{\mathbf{e}} + \mathbf{K}[\mathbf{y} - \hat{\mathbf{b}} - \mathbf{H}(\mathbf{x}^f - \hat{\mathbf{e}})] \tag{22}$$

$$\mathbf{K} = \mathbf{P}^f\mathbf{H}^T(\mathbf{H}\mathbf{P}^f\mathbf{H}^T + \mathbf{R})^{-1}, \tag{23}$$

which requires knowledge of the model bias estimate $\hat{\mathbf{e}}$ and the observation bias estimate $\hat{\mathbf{b}}$. The model bias estimate $\hat{\mathbf{e}}$ can be obtained by eliminating $\hat{\mathbf{x}}$ from Eqs. (22) and (20),

$$\hat{\mathbf{e}} = \mathbf{e}^f - \mathbf{L}[\mathbf{y} - \hat{\mathbf{b}} - \mathbf{H}(\mathbf{x}^f - \mathbf{e}^f)] \tag{24}$$

$$\mathbf{L} = \mathbf{Q}\mathbf{H}^T(\mathbf{H}\mathbf{P}^f\mathbf{H}^T + \mathbf{H}\mathbf{Q}\mathbf{H}^T + \mathbf{R})^{-1}, \tag{25}$$

but then it depends on the knowledge of the observation bias estimate. Finally, the observation bias estimate can be obtained from Eqs. (24) and (21) by eliminating $\hat{\mathbf{e}}$, and we get an expression that depends only on forecast (or prior) values,

$$\hat{\mathbf{b}} = \mathbf{b}^f + \mathbf{M}[\mathbf{y} - \mathbf{b}^f - \mathbf{H}(\mathbf{x}^f - \mathbf{e}^f)], \tag{26}$$

$$\mathbf{M} = \mathbf{S}(\mathbf{H}\mathbf{P}^f\mathbf{H}^T + \mathbf{H}\mathbf{Q}\mathbf{H}^T + \mathbf{R} + \mathbf{S})^{-1}. \tag{27}$$

In this semi-coupled solution, the system is first solved by estimating the observation bias, then the model bias, and then the state. We note, however, that it requires the inversion of three different error covariances.

An entirely uncoupled system solution and much more practical form can be obtained by substituting Eqs. (26) and (27) into Eqs. (24) and (25) and using the following identity $(\mathbf{S}-\mathbf{X})^{-1}(\mathbf{S}\mathbf{X}^{-1}-\mathbf{I})=\mathbf{X}^{-1}$ to obtain,

$$\hat{\mathbf{e}} = \mathbf{e}^f - \mathbf{L}^*[\mathbf{y} - \mathbf{b}^f - \mathbf{H}(\mathbf{x}^f - \mathbf{e}^f)] \tag{28}$$

$$\mathbf{L}^* = \mathbf{Q}\mathbf{H}^T(\mathbf{H}\mathbf{P}^f\mathbf{H}^T + \mathbf{H}\mathbf{Q}\mathbf{H}^T + \mathbf{R} + \mathbf{S})^{-1}, \tag{29}$$

and, similarly, by substituting Eqs. (26), (27), and (29) into Eqs. (22) and (23) to get

$$\hat{\mathbf{x}} = \mathbf{x}^f - \mathbf{e}^f + \mathbf{K}^*[\mathbf{y} - \mathbf{b}^f - \mathbf{H}(\mathbf{x}^f - \mathbf{e}^f)] \tag{30}$$

$$\mathbf{K}^* = (\mathbf{P}^f + \mathbf{Q})\mathbf{H}^T(\mathbf{H}\mathbf{P}^f\mathbf{H}^T + \mathbf{H}\mathbf{Q}\mathbf{H}^T + \mathbf{R} + \mathbf{S})^{-1}. \tag{31}$$

The presence of $\mathbf{Q}$ in the first term in parentheses in Eq. (31) comes from the use of model forecast bias rather than model analysis bias as in Eq. (22).

In this new formulation (Eqs. 26, 27, 28, 29, 30, and 31), the same observation residual $\mathbf{d} = \mathbf{y} - \mathbf{b}^f - \mathbf{H}(\mathbf{x}^f - \mathbf{e}^f)$ is used in all equations. Also, only one matrix, i.e., $\mathbf{X} = \mathbf{H}\mathbf{P}^f\mathbf{H}^T + \mathbf{H}\mathbf{Q}\mathbf{H}^T + \mathbf{R} + \mathbf{S}$, needs to be inverted. The appearance of the observation in the analysis equations for the state, model bias and observation bias does not mean that the information content of the observation is used three times, as was noted in Dee and Todling (2000); Eqs. (26), (27), (28), (29), (30), and (31) can in fact be rewritten in the following form

$$\begin{aligned}
\hat{\mathbf{b}} &= \mathbf{b}^f + \mathbf{S}\mathbf{X}^{-1}\mathbf{d} \\
\hat{\mathbf{e}} &= \mathbf{e}^f - \mathbf{Q}\mathbf{H}^T\mathbf{S}^{-1}(\hat{\mathbf{b}} - \mathbf{b}^f) \\
\hat{\mathbf{x}} &= \mathbf{x}^f - \mathbf{e}^f - (\mathbf{P}^f + \mathbf{Q})\mathbf{Q}^{-1}(\hat{\mathbf{e}} - \mathbf{e}^f),
\end{aligned} \tag{32}$$

which shows clearly that the observation information, $\mathbf{d}$, the innovation vector, is used only once.

Schemes where only the model is biased or only the observations are biased are easily derived from this general formulation. The form given by Eq. (32) also shows clearly that the bias can be estimated separately from the state estimate, the so-called *bias-separation* property (Dee and da Silva 1998). It is important to note, however, that the bias-separation property found by Friedland (1969) actually referred to the separation of the propagation of error covariances in a Kalman filter, which only occurs for a constant model bias with no stochastic forcing. The bias-separation property in the state-bias variables in form given by Eq. (32) (see Dee and da Silva 1998) seems to occur in any optimal linear system, and is just a reflection of the fact that the observation information is only used once.

## 4 Observation Bias Correction Schemes

An important, and actually one of the first, application of bias estimation has been for correcting satellite observations. Radiance observations from satellites usually have large systematic errors. It is essential to remove these biases in the measurement to properly extract the information content for data assimilation.

Consider the typical problem of temperature remote sensing. Systematic errors in measurements and radiative transfer are typically much larger than the model's short-term forecast (e.g. 6 h) bias. The mean observation residual is then a good approximation of the observation bias

$$\langle \mathbf{y} - \mathcal{H}(\mathbf{x}) \rangle \approx \left\langle \boldsymbol{\varepsilon}^o \right\rangle . \tag{33}$$

In the troposphere, the model short-term forecast error is constrained and, to a certain extent, negligible due in part to the fact that other accurate observations such as radiosondes are used in the assimilation. In the middle and upper stratosphere and for other components of the Earth system that are not so well sampled by accurate observations, the property (2) (Eq. 6) may not be valid.

Following Eyre (1992), a parametric form is used to represent the observation bias as a scan angle bias $\beta_0$ and an air mass correction represented as regression of $N$ atmospheric predictors, and which is introduced to account for the fact that radiative transfer systematic errors are state-dependent,

$$\mathbf{b} = \beta_0 + \sum_{i=1}^{N} \beta_i p_i(\mathbf{x}), \tag{34}$$

Typically, only a few predictors are chosen in order to avoid overfitting. It is usual to have as predictors:

- geopotential thickness of the layer 1,000–300 hPa;
- geopotential thickness of the layer 200–50 hPa;
- geopotential thickness of the layer 50–5 hPa;
- geopotential thickness of the layer 10–1 hPa.

Different approaches have been proposed to estimate the parameters $\beta_i$: a static scheme; an adaptive off-line scheme; and an adaptive on-line or variational bias correction scheme.

### *4.1 Static Bias Correction Scheme*

In the static scheme, the optimal values of the parameters are calculated from a set of observations and background $\mathbf{x}^b$ from a control assimilation over a period

of typically a month. The parameters are then fixed and applied to all subsequent analyses.

This scheme is implemented by minimizing the cost function,

$$J(\boldsymbol{\beta}) = \sum_k \frac{1}{2} \left[ \mathbf{y}_k - \mathcal{H}_k \left( \mathbf{x}_k^b \right) - \mathbf{b}(\boldsymbol{\beta}) \right]^T \mathbf{R}^{-1} \left[ \mathbf{y}_k - \mathcal{H}_k \left( \mathbf{x}_k^b \right) - \mathbf{b}(\boldsymbol{\beta}) \right]. \tag{35}$$

Figure 2 illustrates the effect of this scheme for AMSU-B of NOAA-16 (Garand et al. 2006). Panel (a) shows the raw radiances without any bias correction for channels 2–5 as a function of scan angle. Panel (b) shows the radiances after the scan position bias correction. Panel (c) displays the radiances after scan angle bias correction and air-mass correction. Panel (d) shows the standard deviation.

Note how the curvature of the mean observation residual line has been eliminated after the scan angle correction. We also observe that the air-mass correction reduces the observation bias by almost an order of magnitude. Finally, we note also that, in

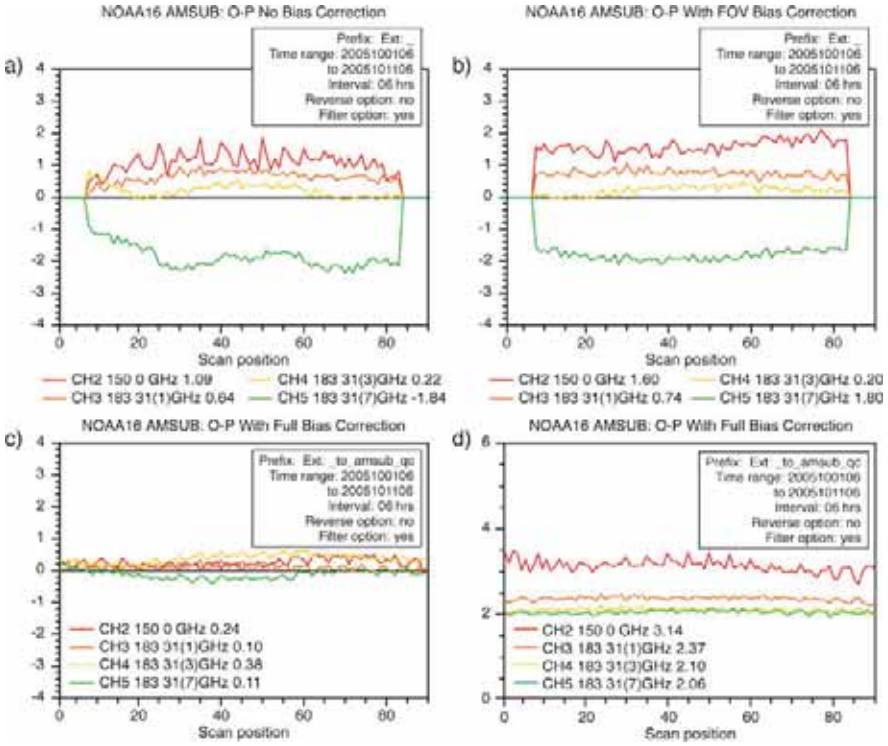



**Fig. 2** Mean AMSU-B observation residuals (O-P, observation minus forecast) versus scan position. (**a**) Raw radiance data; (**b**) after scan angle bias correction; (**c**) after bias correction and air-mass correction. Corresponding standard deviations are given by (**d**). First and last 7 scans not used

practice, the bias correction is made for each channel individually so, in principle, the value of $\mathbf{R}$ is irrelevant in the minimization.

## 4.2 Adaptive Off-Line Bias Correction Scheme

Changes in the nature of the bias such as during contamination, instrument problems or changes in data processing, cannot be accounted for properly by the static bias correction scheme. An adaptive bias correction scheme is an off-line bias correction scheme similar to the static scheme but where the parameters $\beta i$ are updated continuously (but not the predictors nor the scan-angle correction). Typically, the updates are made at each analysis cycle and the correction is made prior to each new analysis, by minimizing the cost function,

$$J(\boldsymbol{\beta}_k) = \frac{1}{2}\left[\mathbf{y}_{k-1} - \mathcal{H}_{k-1}\left(\mathbf{x}^a_{k-1}\right) - \mathbf{b}(\boldsymbol{\beta}_k)\right]^T \mathbf{R}^{-1}\left[\mathbf{y}_{k-1} - \mathcal{H}_{k-1}\left(\mathbf{x}^a_{k-1}\right) - \mathbf{b}(\boldsymbol{\beta}_k)\right] + \frac{1}{2}(\boldsymbol{\beta}_k - \boldsymbol{\beta}_{k-1})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta}_k - \boldsymbol{\beta}_{k-1}); \tag{36}$$

$k$ increases by one after each analysis cycle. $\boldsymbol{\Sigma}$ is equivalent to $\mathbf{S}$, but for the parameter value space, and controls the quasi-stationarity of the bias parameters. The second term in Eq. (36) acts like an inertia constraint (i.e., $\boldsymbol{\beta}_k$ does not change easily with time) but the value of $\boldsymbol{\Sigma}$ is somewhat arbitrary.

The off-line adaptive scheme cannot distinguish observation bias from model bias. As shown by Auligné et al. (2007), if a model bias is present, the information that pulls away the model from its biased solution is gradually removed by the bias correction as it gets contaminated by the model bias, and the scheme converges eventually to the model biased solution.

## 4.3 Adaptive On-Line Bias Correction Scheme or Variational Correction Scheme

A better approach is to update the bias inside the assimilation system by finding corrections that minimize the radiance departure while simultaneously improving the fit to other observed data inside the analysis cycle. This is achieved by including the bias parameters in the control state vector of the variational analysis problem. The cost function to minimize is of the form,

$$\begin{aligned} J(\mathbf{x}, \boldsymbol{\beta}) = {} & \frac{1}{2}[\mathbf{y} - \mathcal{H}(\mathbf{x}) - \mathbf{b}(\boldsymbol{\beta})]^T \mathbf{R}^{-1}[\mathbf{y} - \mathcal{H}(\mathbf{x}) - \mathbf{b}(\boldsymbol{\beta})] \\ & + \frac{1}{2}(\mathbf{x} - \mathbf{x}^f)^T [\mathbf{P}^f]^{-1}(\mathbf{x} - \mathbf{x}^f) \\ & + \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}^f)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta} - \boldsymbol{\beta}^f). \end{aligned} \tag{37}$$

This approach was first developed at NCEP (Derber and Wu 1998) and then implemented at the European Centre for Medium-Range Weather Forecasts, ECMWF (Dee 2004). The adjustment balances the uncertainty of the state forecast, the observations and the inertia constraint on the bias parameters. It accounts naturally for all observations simultaneously inside the analysis. This approach shows some robustness in presence of model bias (Auligné et al. 2007) but, most importantly, goes a long way towards the automation of satellite bias corrections, which is becoming critical in numerous weather prediction centres as more and more satellite observations are assimilated.

Finally, we add that other bias models other than the scan angle-air mass factor correction are in use. The so-called gamma-delta method introduced by Watts and McNally (2004) is based on the assumption that the main bias comes from systematic errors in the radiative transfer model that can be modelled by a multiplier of the total optical depth.

## 5 Model Bias Correction Schemes

Model systematic errors can be estimated and incorporated in the state estimation using data assimilation. Model error is generally represented by an added term to the model forecast and is either a deterministic or stochastic term. Several methods have been proposed; they fall into two main categories, *static estimation* and *dynamical estimation* schemes, depending on whether the bias evolution of errors (either implicit as in 4D-Var or explicit as in a Kalman filter scheme) is accounted for (*dynamical*) or not (*static*) in the optimization scheme.

### *5.1 Static Schemes*

Static schemes have constant error covariances. The background (or forecast) error covariance and the bias error covariances are not propagated in time nor updated as a result of observations, but the model bias is allowed to evolve in time. Static schemes were first developed by Dee and da Silva (1998), where it was assumed that the observation errors have no biases. It is a special case of the more general bias analysis derived in Sect. 3 above. As in the general case, these schemes can be formulated either as a sequential or parallel scheme, which takes the following form in the case of no observation bias:

- In a *sequential form*, the bias estimate is computed first

$$\hat{\mathbf{e}} = \mathbf{e}^f - \mathbf{L}[\mathbf{y} - \mathbf{H}(\mathbf{x}^f - \mathbf{e}^f)] \tag{38}$$

$$\mathbf{L} = \mathbf{Q}\mathbf{H}^T(\mathbf{H}\mathbf{P}^f\mathbf{H}^T + \mathbf{H}\mathbf{Q}\mathbf{H}^T + \mathbf{R})^{-1}, \tag{39}$$

It uses a bias prior or bias forecast $\mathbf{e}^f$ which can be estimated from the previous analysis in the case of a constant bias, or a forecast bias following an evolution law. Once $\hat{\mathbf{e}}$ is computed then the state estimate can be computed as

$$\hat{\mathbf{x}} = \mathbf{x}^f - \hat{\mathbf{e}} + \mathbf{K}[\mathbf{y} - \mathbf{H}(\mathbf{x}^f - \hat{\mathbf{e}})] \tag{40}$$

$$\mathbf{K} = \mathbf{P}^f\mathbf{H}^T(\mathbf{H}\mathbf{P}^f\mathbf{H}^T + \mathbf{R})^{-1}, \tag{41}$$

This scheme has the advantage of using the same gain matrix as the usual unbiased estimation problems;

- In a *parallel form*, both state and bias estimates can be computed independently of each other. It uses Eqs. (38) and (39) for the model bias estimate, and

$$\hat{\mathbf{x}} = \mathbf{x}^f - \mathbf{e}^f + \mathbf{K}^*[\mathbf{y} - \mathbf{H}(\mathbf{x}^f - \mathbf{e}^f)] \tag{42}$$

$$\mathbf{K}^* = (\mathbf{P}^f + \mathbf{Q})\mathbf{H}^T(\mathbf{H}\mathbf{P}^f\mathbf{H}^T + \mathbf{H}\mathbf{Q}\mathbf{H}^T + \mathbf{R})^{-1}, \tag{43}$$

for the state estimate. Contrary to the sequential scheme, the observation-minus-model residuals are the same in both bias and state analysis equations. Also, the matrix to be inverted is the same in both cases. In a PSAS (Physical Space Assimilation System; see chapter *Variational Assimilation*, Talagrand) algorithm the conjugate gradient step need only be solved once.

The equivalent 3D-Var scheme derives directly from Eq. (17) letting $\mathbf{b} = 0$ and has no $\mathbf{S}$ penalty term. In practice, $\mathbf{Q}$ is unknown, but Dee and da Silva (1998) suggested making the assumption that the model bias correlation scales are roughly the same as those of the random components of the error covariance and, thus, an approximation of the form

$$\mathbf{Q} = \gamma\mathbf{P}^f, \tag{44}$$

can be used. Furthermore, if we assume that the model systematic error is small compared to the random forecast errors, i.e., $\gamma \ll 1$, the bias gain matrix $\mathbf{L}$ can then be approximated as,

$$\mathbf{L} = \gamma\mathbf{P}^f\mathbf{H}^T[(1+\gamma)\mathbf{H}\mathbf{P}^f\mathbf{H}^T + \mathbf{R}]^{-1} \approx \gamma\mathbf{K}. \tag{45}$$

This can reduce the computational cost since only one gain matrix needs to be computed. Note that an optimum interpolation type of analysis solver would benefit from this latter approximation, but a conjugate gradient solver in PSAS would not.

A successful implementation of the static scheme was performed by Dee and Todling (2000) for the moisture analysis, and using a constant bias. The parameter was tuned to reduce the energy of the long-wave portion of the spectrum of bias corrected observed-minus-forecast residuals so as to become as flat as possible. It is interesting to note that the bias-corrected observation-minus-forecast residuals were fairly white in the mid troposphere but showed degradation near the surface and higher up near the tropopause.

While some model errors are persistent, others are cyclical, and others, although not cyclical, are predictable. In an ocean data assimilation problem, Chepurin et al. (2005) used an EOF (Empirical Orthogonal Function) analysis of observed-minus-forecast statistics to model the model bias. Leading EOFs that evolved in time were used as the bias evolution model. To untangle the random component from the systematic component of the forecast errors, they assumed that the spatial scales of the systematic errors were basin-wide while random forecast errors had shorter length scales in comparison. This assumption is completely different from that of Eq. (44), but the results showed an improvement in the analyses.

The effectiveness of the bias correction strongly depends on the actual form of the bias model used. In a land surface model a skin temperature bias correction that accounts for diurnal variation was shown to be very effective (Radakovich et al. 2004). In the context of model ozone bias, Dee (2005) was able to construct a predictive bias model using analysis increments and a fit to a lag-6 autoregressive moving average model. In the simple context examined by Dee (2005) the bias correction indicated an improvement in the RMS (root-mean-square) analysis error, but, unfortunately, the scheme was never implemented. The question of using an additive bias model was also revisited with simplified models in the context that the truth model and forecast model may have different attractors (Baek et al. 2006).

## *5.2 Dynamical Schemes*

Dynamical schemes account for the evolution of the state and model bias in the optimization problem. Kalman filtering, the Ensemble Kalman filter (EnKF), and 4D-Var (strong and weak constraint) algorithms have been developed to address this problem (see chapter *Mathematical Concepts of Data Assimilation*, Nichols, for details of these assimilation schemes). To set the stage let us assume that the evolution of the state can be described by

$$\mathbf{x}_k^t = \mathcal{M}_{k-1}\left(\mathbf{x}_{k-1}^t\right) + \mathbf{T}_{k-1}\mathbf{e}_{k-1}^t, \tag{46}$$

and that of the model bias by

$$\mathbf{e}_k^t = \mathcal{G}_{k-1}\left(\mathbf{e}_{k-1}^t, \mathbf{x}_{k-1}^t\right). \tag{47}$$

$\mathbf{T}_k$ represents the transformation of the bias parameter space to the model state space. The transformation $\mathbf{T}_k$ is used when a limited number of bias parameters are estimated (in association with an appropriate bias evolution equation), otherwise we assume that $\mathbf{T}_k = \mathbf{I}$ when the bias parameters are identical to the model variables. A zero-mean white noise can also be added to either one or both of these equations. Such a term in the state equation (Eq. 46) represents the standard model error in Kalman filtering. A random noise added to the bias evolution (Eq. 47), reflects the fact that the bias evolution equation is not perfect. In current

state-of-the-art bias evolution models this is quite a valid assumption. The error covariance of this random noise was introduced earlier in Sect. 3, as the covariance matrix **Q**.

The modelling of bias evolution is fairly recent, and only simple forms have been investigated so far – typically a product of a spatial function with a temporal function, which in the simplest case is just flat and thus represents a constant bias, although with spatial dependence. Strong and weak constraint variational formulations have been investigated with atmospheric models of diverse complexity ranging from simple models to complex operational models. A limited number of studies with generally simple models have also been conducted using a Kalman filtering approach.

The feasibility of estimating model error bias using a strong constraint variational method was first examined by Derber (1989). Using a low resolution limited area quasi-geostrophic model and controlling only the model bias (and not the initial conditions) with a bias model

$$\mathbf{e}_k^t = f(t_k)\boldsymbol{\phi}(x, y, z), \tag{48}$$

that has a prescribed time evolution, Derber (1989) was able to get consistently a better fit to the control analyses and a superior forecast compared to a variational assimilation problem controlled by the initial conditions only. The model was somewhat crude and, as expected, showed large biases in comparison with errors due to initial conditions. With more sophisticated atmospheric models, it is expected that the effect of the initial error will become more important. It is then necessary to estimate both initial conditions and the model bias. Zupanski (1993) generalized the variational assimilation problem of Derber (1989) to include a control over the initial conditions, and the study was conducted with an operational weather prediction model. Interestingly, in this approach the gradient of the cost function with respect to the initial conditions depends on the adjoint variable at the initial time as in the standard 4D-Var framework, and the gradient of the cost function with respect to the model bias is evaluated in the same fashion as in Derber (1989). In the experiments of Zupanski (1993), optimum interpolation analyses were used in place of observations, thus introducing model information in the data. The results were somewhat disappointing, showing that better results were obtained when a 4D-Var estimation of the initial conditions was conducted first, rather than doing a simultaneous model bias and initial condition estimation. Griffith and Nichols (1996) and Nichols (2003) explored further this approach by investigating other simple bias models and touched upon the weak constraint problem, although only for a special case.

A simple derivation of the adjoint equations when biases are considered can be obtained using the Lagrange multiplier method. Consider for instance the cost function,

$$
\begin{aligned}
J(\mathbf{x}_0, \mathbf{e}_0) \;=\; & \frac{1}{2} \sum_{k=0}^{N} [\mathbf{y}_k - \mathcal{H}_k(\mathbf{x}_k)]^T \mathbf{R}_k^{-1} [\mathbf{y}_k - \mathcal{H}_k(\mathbf{x}_k)] \\
& + \frac{1}{2} \left( \mathbf{x}_0 - \mathbf{x}_0^f + \mathbf{e}_0 \right)^T \mathbf{B}^{-1} \left( \mathbf{x}_0 - \mathbf{x}_0^f + \mathbf{e}_0 \right) \\
& + \frac{1}{2} \left( \mathbf{e}_0 - \mathbf{e}_0^f \right)^T \mathbf{Q}^{-1} \left( \mathbf{e}_0 - \mathbf{e}_0^f \right)
\end{aligned} \tag{49}
$$

subject to the strong constraint,

$$
\mathbf{x}_k^t = \mathcal{M}_{k-1} \left( \mathbf{x}_{k-1}^t \right) + \mathbf{e}_{k-1}^t, \tag{50}
$$

$$
\mathbf{e}_k^t = \mathbf{e}_{k-1}^t. \tag{51}
$$

The constrained optimization problem can be turned into an unconstrained problem by introducing $2N$ Lagrange multipliers $\lambda_i (i = 1, \ldots, N)$, $\mu_i (i = 1, \ldots, N)$, and optimizing the new cost function

$$
\begin{aligned}
L(\mathbf{x}_0, \mathbf{e}_0, \mathbf{x}_k, \mathbf{e}_k, \boldsymbol{\lambda}_k, \boldsymbol{\mu}_k) \;=\; & J(\mathbf{x_0}, \mathbf{e_0}) \\
& + \sum_{k=1}^{N} \boldsymbol{\lambda}_k^T [\mathbf{x}_k - \mathcal{M}_{k-1}(\mathbf{x}_{k-1}) + \mathbf{e}_{k-1}] \\
& + \sum_{k=1}^{N} \boldsymbol{\mu}_k^T (\mathbf{e}_k - \mathbf{e}_{k-1}).
\end{aligned} \tag{52}
$$

The gradient of $L$ with respect to each variable is

$$
\begin{aligned}
\frac{\partial L}{\partial \mathbf{x}_0} &= \mathbf{B}^{-1} \left( \mathbf{x}_0 - \mathbf{x}_0^f - \mathbf{e}_0 \right) - \mathbf{H}_0^T \mathbf{R}_0^{-1} [\mathbf{y}_0 - \mathcal{H}_0(\mathbf{x}_0)] - \mathbf{M}_0^T \boldsymbol{\lambda}_1 \\
\frac{\partial L}{\partial \mathbf{e}_0} &= -\mathbf{B}^{-1} \left( \mathbf{x}_0 - \mathbf{x}_0^f - \mathbf{e}_0 \right) + \mathbf{Q}^{-1} \left( \mathbf{e}_0 - \mathbf{e}_0^f \right) - \boldsymbol{\lambda}_1 - \boldsymbol{\mu}_1 \\
\frac{\partial L}{\partial \mathbf{x}_k} &= -\mathbf{H}_k^T \mathbf{R}_k^{-1} [\mathbf{y}_k - \mathcal{H}_k(\mathbf{x}_k)] + \boldsymbol{\lambda}_k - \mathbf{M}_k^T \boldsymbol{\lambda}_{k+1} \\
\frac{\partial L}{\partial \mathbf{e}_k} &= -\boldsymbol{\lambda}_{k+1} + \boldsymbol{\mu}_k - \boldsymbol{\mu}_{k+1} \\
\frac{\partial L}{\partial \mathbf{x}_N} &= -\mathbf{H}_N^T \mathbf{R}_N^{-1} [\mathbf{y}_N - \mathcal{H}_N(\mathbf{x}_N)] + \boldsymbol{\lambda}_N \\
\frac{\partial L}{\partial \mathbf{e}_N} &= \boldsymbol{\mu}_N
\end{aligned} \tag{53}
$$

From the end condition at $k = N$ we get $\boldsymbol{\lambda}_{N+1} = \boldsymbol{\mu}_{N+1} = 0$. The first equation in Eq. (53) above can be rewritten as

$$\frac{\partial L}{\partial \mathbf{x}_0} = \mathbf{B}^{-1}\left(\mathbf{x}_0 - \mathbf{x}_0^f + \mathbf{e}_0\right) - \boldsymbol{\lambda}_0 \tag{54}$$

and to obtain the gradient of $L$ with respect to $\mathbf{e}_0$, we need to iterate backward the fourth equation of Eq. (53) from $k = N$ to $k = 0$ to get an expression for $\boldsymbol{\mu}_0$,

$$\boldsymbol{\mu}_0 = \sum_{i=1}^{N} \boldsymbol{\lambda}_i \tag{55}$$

and then we get

$$\frac{\partial L}{\partial \mathbf{e}_0} = -\mathbf{B}^{-1}\left(\mathbf{x}_0 - \mathbf{x}_0^f - \mathbf{e}_0\right) + \mathbf{Q}^{-1}\left(\mathbf{e}_0 - \mathbf{e}_0^f\right) - \sum_{i=1}^{N} \boldsymbol{\lambda}_i. \tag{56}$$

The introduction of $\mathbf{e}_0$ in the model background penalty term (i.e., $J_B$) introduces a direct coupling between $\mathbf{x}_0$ and $\mathbf{e}_0$ in Eqs. (54) and (56). It is interesting to note that if we had not introduced this dependence in $J_B$, the resulting minimization conditions would have been the same as for the weak constraint case with a constant model bias considered in Griffith and Nichols (1996).

Unexplained model biases are unavoidable and they dictate the use of a weak constraint approach. Neglecting to include random model errors may result in overfitting the analyses and degrading the forecast skill, as was first demonstrated by Wergen (1992). Ménard and Daley (1996) diagnosed the effect of a strong constraint in 4D-Var using Kalman smoother theory. The variational formalism of weak contraints was first introduced by Sasaki (1970). The first implementation was done by Bennett and co-workers (Bennett 1992; Bennett et al. 1993, 1996, 1997) using the representer method which reduced the size of the assimilation problem to the number of observations. Amodei (1995) and Courtier (1997) also introduced an extension of the 4D-PSAS that accounts for random model error, by increasing the size of the control state vector. Zupanski (1997) applied a weak constraint 4D-Var to an operational limited area model with model error represented by a first-order Markov process but in which the estimated random component is defined at a coarser resolution in time and space.

At ECMWF a research effort on weak constraint 4D-Var spanning several years was conducted by Trémolet (2003, 2006, 2007). Using a simple bias evolution model, namely a piece-wise constant, Trémolet investigated issues related to operational implementation such as the reasons for the limited success of model error estimation. He first noted that the model bias error covariance proposed by Dee and da Silva (Eq. 44) has little noticeable impact on the forecast, as the cumulative effect of the model-error forcing approximately compensates for the differences in initial conditions. Using the parallel form of the bias estimation problem (Eqs. 38, 39, 42, and 43) Trémolet noted that the basic difference between the model bias increment and the state increment comes from the leftmost matrix appearing in the gain matrices $\mathbf{L}$ and $\mathbf{K}^*$. If $\mathbf{Q}$ is taken to be proportional to $\mathbf{P}^f$, the initial condition increment and the model bias increment are constrained in the same direction and

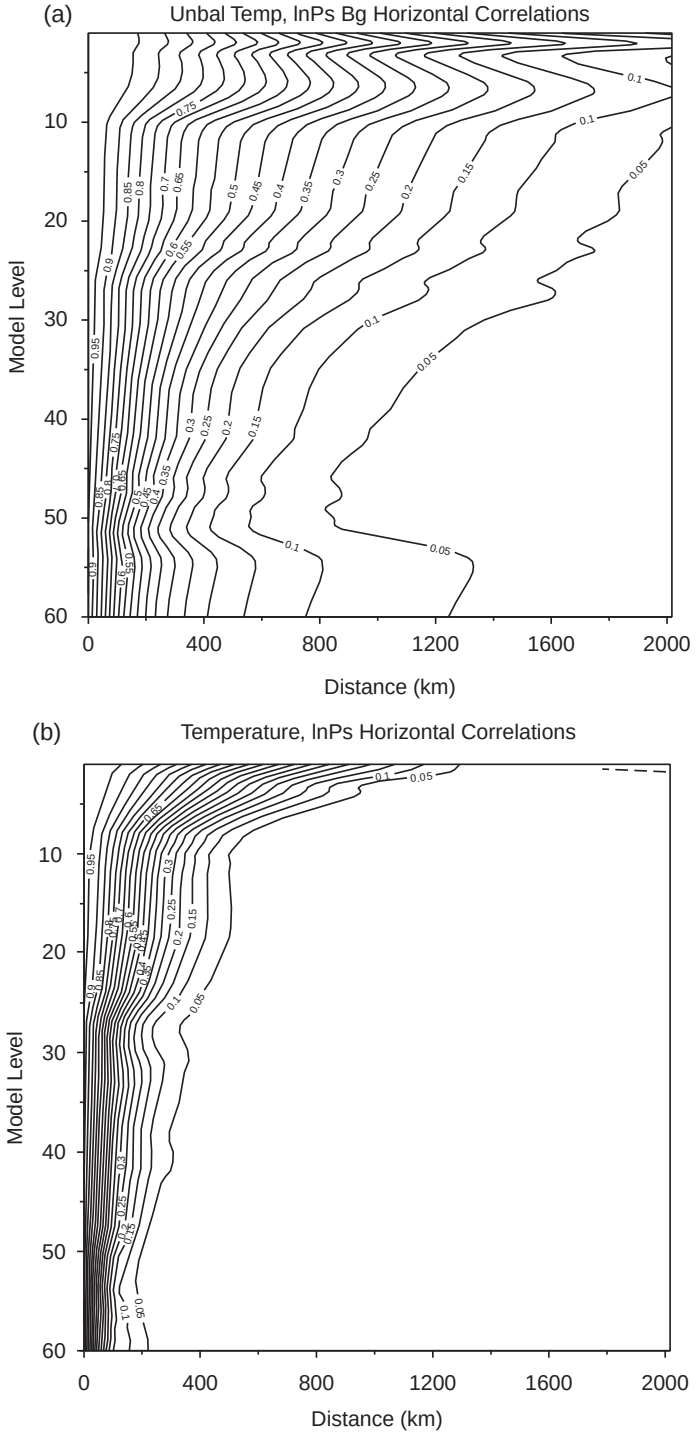



**Fig. 3** Horizontal temperature correlations for the background error (*panel* **a**) and for the model error (*panel* **b**) (Trémolet 2007; © Royal Meteorological Society)

are of opposite sign. The model bias is restricted to the same subspace as the initial condition increments and they differ only in their relative amplitude. An alternative method was presented by Trémolet. Since model error arises as a tendency in the model prognostic equation, model tendencies derived from an ensemble of forecasts should represent a distribution of the possible evolutions of the atmosphere from the true state. The spread of these tendencies may be a good first guess for the model error covariance. Figure 3 shows an example of the model error covariance horizontal correlation for temperature (panel b) as a function of height compared with the background error covariance (panel a). The model error covariance has much smaller scales than the background error covariance and thus provides an additional source of information for the model bias estimation.

## 6 Conclusions

A standard assumption in data assimilation is that neither observations nor the model have systematic errors, i.e., biases. In reality, models and observations often have systematic errors that cannot be neglected in comparison with the error standard deviation. In some cases, as in satellite remote sensing, systematic errors can be as large as the random component. When observations or the model have biases, an assimilation scheme based on the standard assumption will produce analyses that are still biased although it may be somewhat reduced. The presence of biases does in effect reduce the ability of observations to be used effectively in a model no matter how the error statistics are prescribed. Removing biases at the source or by a bias correction scheme is one way to produce an analysis that is unbiased. Over the years and, particularly recently, significant progress has been made to include on-line bias estimation and correction schemes in the assimilation system. Although data assimilation theory can be formulated where both observations and models have systematic errors, the outstanding issue of bias estimation is the problem of identifying model and observation bias from innovation statistics which can only be solved by using additional information. The experience with operational models that is now building up may provide further insights on this particular problem, and on the error statistics that are needed to address the problem of bias estimation.

From a mathematical point of view, bias estimation of model error and observation error on-line with state estimation can be formulated in Kalman filtering form, 3-D variational (3D-Var) form, and 4-D variational (4D-Var) strong and weak constraint forms. In a Kalman filter form, the estimation can be formulated either as a sequential process where the bias estimation is performed first and then state estimation follows, or in parallel where both bias and state estimates are computed concurrently and independently of each other. No special property or assumption aside from linearity is needed for sequential and parallel estimation steps to occur.

The so-called bias separation introduced by Friedland (1969) actually refers to the Kalman filter evolution of the error covariances of the bias parameters that can

be separated from the evolution of the state and cross state-bias error covariances which occur in the case of a constant bias treated as a strong constraint. The parallel form of the estimation problem has allowed clarification of some theoretical and practical issues about the bias estimation problem. In particular, it has been shown that, in effect, the observations are not used twice (once for the bias estimation and once for the state estimation) despite the fact that the observations may appear twice in the combined state-bias estimation algorithm. Also, it is now clear that if the model bias covariance error is based on the background error covariance, the correction of the state actually compensates to a large extent the correction on the model bias, and results in very little improvement in the forecast. It is thus important that the subspaces spanned by the bias error covariance and state error covariance be different. These new findings may shed some light on the outstanding issue of bias estimation – how can we separate observation bias from model bias? More work in that direction needs to be done and implementation in an operational system should provide insights to this fundamental problem. The application of robust estimation theory to the bias estimation problem (e.g. Kitanidis 1987; Simon 2006; Gillijns and De Moor 2007) may be a promising avenue, as it would reduce our dependence on unknown or poorly known error statistics.

**Acknowledgments** The author wishes to thank Stephen Cohn for the careful review of the manuscript, and Olivier Talagrand and Dick Dee for their thoughtful review which helped clarify the assumptions and limitations built in these algorithms.

## References

Alexandersson, H. and A. Moberg, 1997. Homogenization of Swedish temperature data. Part I: Homogeneity test for linear trends. *Int. J. Climatol*., **17**, 25–34.

Amodei, L., 1995. Solution approché pour un problème d'assimilation des données météorologiques avec la prise en compte de l'erreur de modèle. *Comptes Rendues de l'Académie des Sciences*, **321**, Série IIa, 1087–1094.

Auligné, T., A.P. McNally and D.P. Dee, 2007. Adaptive bias correction for satellite data in a numerical weather prediction system. *Q. J. R. Meteorol. Soc*., **133**, 631–642.

Baek, S-J., B.R. Hunt, E. Kalnay, E. Ott and I. Szunyogh, 2006. Local ensemble filtering in the presence of model bias. *Tellus*, **58A**, 293–306.

Bennett, A.F., 1992. *Inverse Methods in Physical Oceanography*, Cambridge University Press, Cambridge, UK, 346pp.

Bennett, A.F., B.S. Chua and L.M. Leslie, 1996. Generalized inversion of a global numerical weather prediction model. *Meteorol. Atmos. Phys*., **60**, 165–178.

Bennett, A.F., B.S. Chua and L.M. Leslie, 1997. Generalized inversion of a global numerical weather prediction model. II: Analysis and implementation. *Meteorol. Atmos. Phys*., **61**, 129–140.

Bennett, A.F., L.H. Leslie, C.R. Hagelberg and P.E. Powers, 1993. Tropical cyclone prediction using a barotropic model initialized by a generalized inverse method. *Mon. Weather Rev*., **121**, 1714–1729.

Chepurin, G.A., J.A. Carton and D.P. Dee, 2005. Forecast model bias correction in ocean data assimilation. *Mon. Weather Rev*., **133**, 1328–1342.

Courtier, P., 1997. Dual formulation of four-dimensional variational assimilation. *Q. J. R. Meteorol. Soc*., **123**, 2449–2461.

Dee, D.P., 2004. Variational bias correction of radiance data in the ECMWF system. In *Proceedings of the ECMWF Workshop on Assimilation of High Spectral Resolution Sounders in NWP*, Reading, UK, 28 June–1 July 2004, pp 97–112.

Dee, D.P., 2005. Bias and data assimilation. *Q. J. R. Meteorol. Soc*., **131**, 3323–3343.

Dee, D.P. and A.M. da Silva, 1998. Data assimilation in presence of forecast bias. *Q. J. R. Meteorol. Soc*., **124**, 269–295.

Dee, D.P. and R. Todling, 2000. Data assimilation in the presence of forecast bias: The GEOS moisture analysis. *Mon. Weather Rev*., **128**, 3268–3282.

Derber, J.C., 1989. A variational continuous assimilation technique. *Mon. Weather Rev*., **117**, 2437–2446.

Derber, J.C. and W-S. Wu, 1998. The use of TOVS cloud-cleared radiances in the NCEP SSI analysis system. *Mon. Weather Rev*., **126**, 2287–2299.

Eyre, J.R., 1992. A bias correction scheme for simulated TOVS brightness temperatures. *Tech. Memo.*, **186**, ECMWF, Reading, UK.

Friedland, B., 1969. Treatment of bias in recursive filtering. *IEEE Trans. Autom. Contr.*, **AC-14**, 359–367.

Garand, L., G. Deblonde, D. Anselmo, J. Aparicio, A. Beaulne, J. Hallé, S. Macpherson and N. Wagneur, 2006. Experience with bias correction at CMC. *Proceedings of the ECMWF/EUMETSAT NWP-SAF Workshop on Bias Estimation and Correction in Data Assimilation*, Reading, UK, 8–11 November 2005, pp 153–162.

Gillijns, S. and B. De Moor, 2007. Model error estimation in ensemble data assimilation. *Nonlinear Process. Geophys*., **14**, 59–71.

Griffith, A.K. and N.K. Nichols, 1996. Accounting for model error in data assimilation using adjoint methods. In *Computational Differentiation: Techniques, Applications and Tools*, Berz, M., C. Bishof, G. Corliss and A. Greiwank (eds.), SIAM, Philadelphia, pp 195–204.

Haimberger, L., 2007. Homogenization of radiosonde temperature time series using innovation statistics. *J. Climate*, **20**, 1377–1403.

Kitanidis, P.K., 1987. Unbiased minimum-variance linear state estimation. *Automatica*, **23**, 775–778.

Lea, D.J., J-P. Drecourt, K. Haines and M.J. Martin, 2008. Ocean altimeter assimilation with observational- and model-bias correction. *Q. J. R. Meteorol. Soc*., **134**, 1761–1774.

Lewis, J.M., S. Lakshmivarahan and S.K. Dhall, 2006. *Dynamic Data Assimilation: A Least Square Approach*, Cambridge University Press, New York, 654pp.

Ménard, R., S. Chabrillat, C. Charrette, M. Charron, T. von Clarmann, D. Fonteyn, P. Gauthier, J. de Grandpré, A. Kallaur, J. Kaminski, J. McConnell, A. Robichaud, Y. Rochon, P. Vaillancourt and Y. Yang, 2007. *Coupled Chemical-Dynamical Data Assimilation*. ESA/ESTEC Contract No. 18560/04/NL/FF Final report, 458 pp. Executive summary available from http://esamultimedia.esa.int/docs/gsp/completed/C18560ExS.pdf.

Ménard, R., S.E. Cohn, L-P. Chang and P.M. Lyster, 2000. Assimilation of stratospheric chemical tracer observations using a Kalman filter. Part I: Formulation. *Mon. Weather Rev.*, **128**, 2654–2671.

Ménard, R. and R. Daley, 1996. The application of Kalman smoother theory to the estimation of 4D Var error statistics. *Tellus*, **48A**, 221–237.

Nichols, N.K., 2003. Treating model error in 3-D and 4-D data assimilation. In *Data Assimilation for the Earth System*, NATO Science Series: IV. Earth and Environmental Sciences 26, Swinbank, R., V. Shutyaev and W.A. Lahoz (eds.), Kluwer Academic Publishers, Dordrecht, The Netherlands, pp 127–135, 378pp.

Polavarapu, S., T.G. Shepherd, Y. Rochon and S. Ren, 2005. Some challenges of middle atmosphere data assimilation. *Q. J. R. Meteorol. Soc*., **131**, 3513–3527.

Radakovich, J.D., M.G. Bosilovich, J-D. Chern, A.M. daSilva, R. Todling, J. Joiner, M-L. Wu and P. Norris, 2004. Implementation of coupled skin temperature analysis and bias correction in the NASA/GMAO finite volume assimilation system (FvDAS). P1.3 in *Proceedings of the 8th AMS Symposium on Integrated Observing and Assimilation Systems of the Atmosphere, Oceans, and Land Surface*, Seattle, WA, USA, 12–15, January 2004.

Sasaki, Y., 1970. Some basic formalisms in numerical variational analysis. *Mon. Weather Rev*., **98**, 875–883.

Saunders, R., 2005. Sources of biases in infrared radiative transfer models. *Proceedings of the ECMWF/EUMETSAT NWP-SAF Workshop on Bias Estimation and Correction in Data assimilation*, 8–11 November 2005, pp 41–50.

Simon, D., 2006. *Optimal State Estimation: Kalman*, $H_\infty$, *and Nonlinear Approaches*, Wiley and Sons, New Jersey, 526pp.

Trémolet, Y., 2003. Model error in variational data assimilation. *Proceedings ECMWF Seminar on* "*Recent Developments in Data Assimilation for Atmosphere and Ocean*", Reading, UK, 8–12 September 2003, pp 361–367.

Trémolet, Y., 2006. Accounting for an imperfect model in 4D-Var. *Q. J. R. Meteorol. Soc*., **132**, 2483–2504.

Trémolet, Y., 2007. Model-error estimation in 4D-Var. *Q. J. R. Meteorol. Soc*., **133**, 1267–1280.

Watts, P.A. and A.P. McNally, 2004. Identification and correction of radiative transfer modeling errors for atmospheric sounders: AIRS and AMSU-A. *Proceedings of the ECMWF Workshop on Assimilation of High Resolution Sounders in NWP*. Reading, UK, 28 June–1 July, pp 23–38.

Wergen, W., 1992. The effect of model errors in variational assimilation. *Tellus*, **44A**, 297–313.

Zupanski, M., 1993. Regional four-dimensional variational data assimilation in quasi-operational forecasting environment. *Mon. Weather Rev*., **121**, 2396–2408.

Zupanski, D., 1997. A general weak constraint application to operational 4DVar data assimilation systems. *Mon. Weather Rev*., **125**, 2274–2292.

# The Principle of Energetic Consistency in Data Assimilation

**Stephen E. Cohn**

## 1 Introduction

The preceding chapters have illustrated two essential features of data assimilation. First, to extract all the information available in the observations requires all the sources of uncertainty – in the initial conditions, the dynamics, and the observations – to be identified and accounted for properly in the data assimilation process. This task is complicated by the fact that the non-linear dynamical system actually being observed is typically an infinite-dimensional (continuum) system, whereas at one's disposal is only a finite-dimensional (discrete) numerical model of the continuum system dynamics. Second, to formulate a computationally viable data assimilation algorithm requires some probabilistic assumptions and computational approximations to be made. Those made in four-dimensional variational (4D-Var) and ensemble Kalman filter (EnKF) methods have been discussed in Chapters *Variational Assimilation* (Talagrand) and *Ensemble Kalman Filter: Status and Potential* (Kalnay), respectively.

The need to make assumptions and approximations makes it difficult in practice to distinguish whether uncertainties perceived by a data assimilation scheme are genuine, arising from the initial conditions, continuum dynamics and observations, or are instead artificial uncertainties that arise from assumptions and approximations made in the algorithmic formulation of the scheme itself. It is even possible that the latter dominate. For instance, in an EnKF for atmospheric data assimilation, Houtekamer et al. (2005) have found that the "model error" perceived by the filter – the total uncertainty accumulated from all sources not represented explicitly in the filter formulation – is quite large. Houtekamer and Mitchell (2005, pp. 3284–3285) go on to report that, when measured in a linearized total energy norm, this uncertainty is comparable to what would be incurred by neglecting model "physics" entirely. They conclude that much of it may originate in the analysis step, i.e., in

S.E. Cohn (✉)
Global Modeling and Assimilation Office, NASA Goddard Space Flight Center, Greenbelt, MD, 20771, USA
e-mail: stephen.e.cohn@nasa.gov

W. Lahoz et al. (eds.), *Data Assimilation*, DOI 10.1007/978-3-540-74703-1_7,

assumptions and approximations made in the EnKF formulation for assimilating the observations themselves. Considering that this uncertainty likely stems from a multitude of sources beyond the discrete dynamical model, Houtekamer and Mitchell (2005, p. 3285) suggest referring to it as "system error" rather than model error.

This example serves to illustrate the fact that current data assimilation methodologies lack a mechanism for distinguishing clearly between artificial and genuine sources of uncertainty. Such a mechanism would require a general principle depending only on known properties of the continuum system being observed, not on any assumptions or approximations made in the formulation of the data assimilation scheme itself. The present chapter states such a principle, called here the *principle of energetic consistency* (PEC), demonstrates its validity for a wide range of nonlinear continuum dynamics, and illustrates its application to distinguishing sources of uncertainty in EnKF methods. This and related applications of the PEC are discussed in Sect. 2, while supporting theoretical results are deferred mostly to Sects. 3, 4, and 5 and three appendices. Concluding remarks are given in Sect. 6.

## *1.1 Applications*

The key assumption of the PEC is that the non-linear continuum system being observed has total energy as a scalar invariant property, and which can be expressed in some state variables, called energy variables of the system, as the square of the norm on a separable Hilbert space. For example, for the hydrostatic atmospheric primitive equation dynamics discussed in Sect. 2.1, one set of energy variables is comprised of $s_1 = u\sqrt{p_*}$, $s_2 = v\sqrt{p_*}$, $s_3 = \sqrt{Tp_*}$ and $s_4 = \sqrt{p_*}$, where $u$ and $v$ are the zonal and meridional wind components, respectively, $T$ is temperature, and $p_* = p_s - p_t$, with $p_s$ the surface pressure and $p_t$ the (constant) top pressure. The PEC can be made to apply also to systems having only total mass as a scalar invariant, for instance to the assimilation of any number of chemically interacting tracers, by taking the square root of the mass density of each tracer as a state variable.

Applying the PEC to a data assimilation scheme requires the state variables of the scheme to be chosen to be (discretized) energy variables. As discussed in Sects. 2.2 and 2.4, this requires no explicit change of variables in an existing numerical model of the continuum dynamics, but it does require the observation operators to be expressed in terms of energy variables. When the state variables of a data assimilation scheme are chosen to be energy variables, the norm in which quantities are measured represents actual total energy rather than a linearized total energy.

The principle of energetic consistency is stated precisely in Sect. 2.1. Briefly, suppose that the state variables used to describe the continuum system being observed are energy variables for the system, for instance $\mathbf{s} = (s_1, s_2, s_3, s_4)^T$ in the example above, where the superscript $T$ denotes transposition. Then the total energy of the continuum system at time $t$ is $E(t) = \|\mathbf{s}(t)\|^2$, where $\|\cdot\|$ denotes a Hilbert space norm, and being a scalar invariant, *E(t)* is a property of the system itself, not of the choice of state variables. The PEC states that

$$||\bar{\mathbf{s}}(t)||^2 + \operatorname{tr}\mathcal{P}(t) = \mathcal{E}E(t)\,,$$

where $\bar{\mathbf{s}}(t)$ and $\mathcal{P}(t)$ are, respectively, the mean and covariance operator of $\mathbf{s}(t)$, the symbol tr denotes the trace operator, $\mathcal{E}$ is the expectation operator, and it is assumed that $\mathcal{E}E(t) < \infty$. The trace of the covariance operator is called the total variance, or total uncertainty, in the system state $\mathbf{s}(t)$. Thus the PEC partitions the expected value of the total energy of the system into two parts, a "certain" part, namely the total energy $\|\bar{\mathbf{s}}(t)\|^2$ of the mean state, and an "uncertain" part, namely the total variance $\operatorname{tr}\mathcal{P}(t)$. By way of this partitioning, the PEC says that the mean state and the covariance operator must be energetically consistent. Mathematically, the PEC is the extension to second-order Hilbert space-valued random variables (defined in Appendix 1) of the familiar result that $\bar{x}^2 + \sigma^2 = \mathcal{E}x^2$ for a scalar random variable $x$ with mean $\bar{x} = \mathcal{E}x$ and variance $\sigma^2 = \mathcal{E}(x - \bar{x})^2$.

The principle of energetic consistency is a general statement that requires little in order to be valid. It is not a statement about any particular continuum dynamics, but rather about a large class of dynamics. However, if the dynamics are also conservative, i.e., if $E(t) = E(t_0)$ for every initial state, where $t > t_0$ and $t_0$ is the initial time, then the PEC implies immediately that

$$||\bar{\mathbf{s}}(t)||^2 + \operatorname{tr}\mathcal{P}(t) = ||\bar{\mathbf{s}}(t_0)||^2 + \operatorname{tr}\mathcal{P}(t_0).$$

This statement of energy conservation is an exact dynamical link between just the first two moments of the continuum system state. It says that the total variance of the continuum state can increase (decrease) only as a result of extracting energy from (inserting energy into) the mean state, with the change in total variance balanced exactly by the change in total energy of the mean state. Special cases of the PEC written essentially as this statement of energy conservation have been recognized and used for different purposes by Kraichnan (1961), Epstein (1969), Fleming (1971), Cohn (1993, pp. 3131–3132), and Cohn (2009).

In Sect. 2.2 of the present chapter, it is shown that a conditional version of the PEC holds:

$$||\bar{\mathbf{s}}^k(t)||^2 + \operatorname{tr}\mathcal{P}^k(t) = \mathcal{E}(E(t)|\mathbf{y}^k)\,,$$

where $\bar{\mathbf{s}}^k(t)$ and $\mathcal{P}^k(t)$ are, respectively, the conditional mean and conditional covariance operator of $\mathbf{s}(t)$. Here the conditioning is on arbitrary observation vectors $\mathbf{y}_i = \mathbf{y}(t_i)$, $i = 1, \ldots, k$, and $\mathbf{y}^k = (\mathbf{y}_1^T, \ldots, \mathbf{y}_k^T)^T$ denotes the vector of all the observations up to time $t_k$. Like the PEC itself, the conditional version is a general statement, requiring little for validity, in particular requiring no assumptions on the relationship between the observations and the continuum state.

Ensemble Kalman filters are designed to calculate a discrete approximation to the conditional mean and covariance operator, under a number of assumptions (e.g. Anderson and Anderson 1999). The generality of the conditional version of the PEC is what makes it useful for distinguishing genuine and artificial sources of uncertainty in EnKF schemes. The conditional version of the PEC does not apply

directly to 4D-Var methods, however, because these are designed to approximate the conditional mode, under a number of assumptions, rather than the conditional mean. Some remarks on application of the PEC to 4D-Var methods are given in Sect. 6.

It is shown also in Sect. 2.2 that, as in the finite-dimensional case, the conditional mean state is the minimum variance state estimate: it minimizes the expected value of the total energy of the estimation error, under the sole assumption that this expectation is finite. More generally, for any choice of state variables that are not energy variables, the conditional mean state still minimizes the expected value of linearized energy norms of the estimation error. However, unlike the actual total energy of the estimation error, a linearized energy norm does not measure an intrinsic property of the observed continuum system, but only a property of the choice of state variables. Thus the fact that the conditional mean state is the minimum variance state estimate provides by itself a good reason to choose the state variables of an EnKF scheme to be energy variables.

Section 2.3 gives relationships that the PEC implies for arbitrary discretizations of the continuum dynamics, and Sect. 2.4 applies these to provide relationships that are supposed to be satisfied by EnKF schemes. The relationships corresponding to those that hold for conservative continuum dynamics are especially useful for testing the effect of the various assumptions and approximations made in EnKF schemes. When what is supposed to be a conservative continuum environment is simulated numerically, for instance in the case of an atmospheric model by turning off the model physics and simulating only the dynamics, these relationships can be used to verify whether or not a given assumption or approximation creates an artificial energetic source (or sink) of uncertainty. This kind of diagnostic test is completely analogous to energy conservation tests run on a numerical model of the dynamics during model development.

Section 2.4 uses these relationships to obtain theoretical results on some common approximations as artificial sources or sinks of uncertainty, including limited ensemble size, use of the sample covariance, covariance localization, the linear Kalman-type analysis update, and perhaps most importantly, use of a discrete dynamical model. Section 2.4 concludes with an analysis showing that a significant loss of total variance can occur as a result of even slight, but spurious, numerical dissipation typical of discrete model dynamics. The analysis shows further that, because the assimilation of observations continues to feed energy into small spatial scales, only to be dissipated away again, spuriously, during subsequent model integration of each ensemble member, the total variance can decay exponentially. Thus the interaction between spurious model dissipation and the assimilation of observations can cause ensemble collapse and filter divergence if left untreated.

This spurious loss of total variance, compounded by the assimilation of observations, is a problem not only for ensemble Kalman filtering per se. It has been observed to occur also for a full-rank Kalman filter in a study of stratospheric constituent data assimilation by Ménard et al. (2000) and Ménard and Chang (2000), making itself evident in that case by the presence of total mass as a supposedly conserved scalar. This problem may explain much of the need for the large "system error" term invoked by Houtekamer and Mitchell (2005, p. 3285), and for the "covariance inflation" factor proposed by Anderson and Anderson (1999, p. 2747)

which has become a common design feature in EnKF schemes. The analysis of Sect. 2.4 suggests a way to remedy this problem directly, in essence by undoing the spurious dissipation that acts on the ensemble perturbations. That the proposed remedy can be properly "tuned" is assured by the statement of energy conservation provided by the PEC.

## *1.2 Theory*

The essential requirement for the principle of energetic consistency to be valid is for the state of the non-linear system being observed to exist as a second-order Hilbert space-valued random variable over a closed time interval. The remaining sections of this chapter give general hypotheses under which this is the case. The emphasis is on continuum dynamical systems that are deterministic and may conserve total energy, but have a random initial state, since it is in the conservative case that the PEC has the most immediate applications indicated above. In the language of partial differential equations, this means that of primary interest in the rest of the chapter is the stochastic initial-value problem for non-linear hyperbolic systems. The parabolic case is encountered more frequently than the hyperbolic case in the literature on stochastic partial differential equations, but parabolic systems are usually not conservative and have a fundamentally different character than hyperbolic ones. Loss of total variance due to spurious dissipation in an otherwise conservative system is an illustration of this difference.

Section 3 gives the main theoretical results for stochastic initial-value problems on an arbitrary separable Hilbert space. Section 3.1 describes the abstract problem setting, and Sect. 3.2 summarizes the theory of Hilbert space-valued random variables which is given in more detail in Appendix 1. Theorem 1 in Sect. 3.3 states hypotheses under which the stochastic initial-value problem defines a second-order Hilbert space-valued random variable over a closed time interval, with conservative dynamics as a special case. Section 3.4 discusses the simplification of Theorem 1 that occurs if it is assumed that the total energy of every realization of the initial state is bounded by a constant. Such an assumption yields a convenient characterization of the system state, and also restricts the class of probability distributions that the system state can have at any time. For instance, the state cannot be Gaussian-distributed under such an assumption.

Section 4 shows how Theorem 1 is applied to verify the PEC for classical solutions of non-linear systems of differential equations. The stochastic initial-value problem for ordinary differential equations is treated in Sect. 4.1, and for symmetric hyperbolic partial differential equations in Sect. 4.2. For the hyperbolic case, well-posedness of the stochastic initial-value problem turns out generally to require boundedness of the total energy of every realization of the initial state. Thus the solution is not Gaussian-distributed at any time, but it can be characterized in a convenient way.

The results of Sect. 4.2 are applied to the global non-linear shallow-water equations as a concrete example in Sect. 5. For the shallow-water equations, $s_1 = u\sqrt{\Phi}$, $s_2 = v\sqrt{\Phi}$ and $s_3 = \Phi$, where $\Phi$ is the geopotential, comprise a set of energy

variables. Smoothness conditions satisfied by every realization of the solution $\mathbf{s} = (s_1, s_2, s_3)^T$ of the stochastic initial-value problem are given. Every realization of the geopotential field is bounded from below by a single positive constant, and a characterization of the solution is used to show how such a random field can be constructed for the initial condition. The trace of the covariance operator can be expressed as

$$\operatorname{tr} \mathcal{P}(t) = \int_S \operatorname{tr} \mathbf{P}_t(\mathbf{x}, \mathbf{x})\, a^2 \cos\phi \, \mathrm{d}\phi \, \mathrm{d}\lambda,$$

where the integration is over the sphere $S$ of radius $a$, $\mathbf{P}_t$ is the $3 \times 3$ covariance matrix of the stochastic shallow-water state at time $t$, $\mathbf{x} = (\lambda, \phi)$ denotes location on the sphere with $\lambda$ the longitude and $\phi$ the latitude, and $\operatorname{tr} \mathbf{C}$ denotes the trace, or sum of the diagonal elements, of a matrix $\mathbf{C}$.

Appendix 1 covers the theory of Hilbert space-valued random variables, Appendix 2 treats the theory of families of Hilbert spaces which are needed to handle spherical geometry, and Appendix 3 summarizes mathematical concepts and definitions used in the text.

## 2 The Principle of Energetic Consistency: Some Applications

### *2.1 The Principle of Energetic Consistency*

Denote by $\mathbf{s} = (s_1, \ldots, s_n)^T$ the state vector of the continuum system whose state is to be estimated. Assume that the state variables $s_1, \ldots, s_n$ are energy variables for the system. By this it is meant that there is a real, separable Hilbert space $\mathcal{H}$, with inner product and corresponding norm denoted by $(\cdot, \cdot)$ and $\| \cdot \|$, respectively, such that the total energy $E = \|\mathbf{s}\|^2$ is a scalar invariant of the system, i.e., a property of the system itself and not of any choice of state variables.

For example, $n = 4$ in the case of hydrostatic atmospheric dynamics mentioned in the Introduction, and $s_1 = u\sqrt{p_*}$, $s_2 = v\sqrt{p_*}$, $s_3 = \sqrt{Tp_*}$ and $s_4 = \sqrt{p_*}$ are energy variables. This is seen by writing the total energy integral for a (shallow) hydrostatic atmosphere as

$$E = \int \int \int_0^1 \mathbf{s}^T \mathbf{A}\mathbf{s} \, \mathrm{d}\sigma \, a^2 \mathrm{d}S,$$

where $\mathbf{s} = (s_1, s_2, s_3, s_4)^T$, $\mathbf{A}$ is the diagonal matrix

$$\mathbf{A} = \tfrac{1}{g} \operatorname{diag}\left(\tfrac{1}{2}, \tfrac{1}{2}, c_p, \phi_s\right),$$

$g$ is the acceleration due to gravity, $c_p$ is the specific heat of (dry) air at constant pressure, $\phi_s$ is the surface geopotential, $a$ is the Earth radius, the double integral is over the sphere with element of surface area $a^2 \mathrm{d}S$, and $\sigma = (p - p_t)/p_*$ is the vertical

coordinate where $p$ is pressure; cf. Kasahara (1974, Eq. 5.18 and p. 516).[1] In this case, $\mathcal{H}$ is the Hilbert space of real 4-vectors with fourth component independent of the vertical coordinate $\sigma$, with inner product

$$(\mathbf{f}, \mathbf{g}) = \int \int \int_0^1 \mathbf{f}^T \mathbf{A}\mathbf{g} \, \mathrm{d}\sigma \, a^2 \mathrm{d}S$$

and corresponding norm $\|\mathbf{g}\| = (\mathbf{g}, \mathbf{g})^{1/2} < \infty$, for all $\mathbf{f}, \mathbf{g} \in \mathcal{H}$. Similarly, $n = 3$ for shallow-water dynamics, and $s_1 = u\sqrt{\Phi}$, $s_2 = v\sqrt{\Phi}$ and $s_3 = \Phi$ are energy variables. As discussed further in Sect. 5, in this case $\mathcal{H}$ is the Hilbert space of real 3-vectors with inner product

$$(\mathbf{f}, \mathbf{g}) = \int \int \mathbf{f}^T \mathbf{g} \, a^2 \mathrm{d}S$$

and corresponding norm $\|\mathbf{g}\| = (\mathbf{g}, \mathbf{g})^{1/2} < \infty$, for all $\mathbf{f}, \mathbf{g} \in \mathcal{H}$.

Let $(\Omega, \mathcal{F}, P)$ be a complete probability space, and assume that the system state $\mathbf{s} = \mathbf{s}(t)$ is an $\mathcal{H}$-valued random variable, for all time $t$ in a closed time interval $\mathcal{T} = [t_0, T]$. This means that, for each $t \in \mathcal{T}$ and each $\mathbf{g} \in \mathcal{H}$, $(\mathbf{g}, \mathbf{s}(t))$ is a scalar (real) random variable on $(\Omega, \mathcal{F}, P)$. Randomness of the system state may arise, for instance, from a random initial condition $\mathbf{s}(t_0)$, from uncertain parameters in the system dynamics, or from stochastic forcing of the system dynamics.

Since $\mathbf{s}(t)$ is an $\mathcal{H}$-valued random variable, the total energy $E(t) = \|\mathbf{s}(t)\|^2$ is a scalar random variable, for all $t \in \mathcal{T}$. Assume that $\mathcal{E}E(t) < \infty$ for all $t \in \mathcal{T}$, where $\mathcal{E}$ denotes the expectation operator, which is defined only for scalar random variables on $(\Omega, \mathcal{F}, P)$. It follows (see Appendices 1a–1c for details) from the stated assumptions that, for all $t \in \mathcal{T}$, there exists a unique element $\bar{\mathbf{s}}(t) \in \mathcal{H}$ such that

$$(\mathbf{g}, \bar{\mathbf{s}}(t)) = \mathcal{E}(\mathbf{g}, \mathbf{s}(t)) \tag{1}$$

for all $\mathbf{g} \in \mathcal{H}$, called the mean of $\mathbf{s}(t)$, and a unique bounded linear operator $\mathcal{P}(t) : \mathcal{H} \to \mathcal{H}$ such that

$$(\mathbf{f}, \mathcal{P}(t)\mathbf{g}) = \mathcal{E}\left[(\mathbf{f}, \mathbf{s}(t) - \bar{\mathbf{s}}(t))(\mathbf{g}, \mathbf{s}(t) - \bar{\mathbf{s}}(t))\right] \tag{2}$$

---

[1] Staniforth et al. (2003) point out that the total energy of such an atmosphere is actually $E + E'$, where $E'$ is the constant

$$E' = \frac{1}{g} \int \int \int_0^1 \phi_s p_t \, \mathrm{d}\sigma \, a^2 \mathrm{d}S = \frac{p_t}{g} \int \int \phi_s \, a^2 \mathrm{d}S,$$

and they give corresponding expressions for $E$ and $E'$ for deep and/or non-hydrostatic atmospheres. Also, note that a moisture variable is not considered to be an atmospheric state variable for the purposes of this chapter, since moisture in the atmosphere contributes only indirectly to the total energy integral. Thus, choosing state variables to be energy variables does not imply a choice of moisture variables. Dee and da Silva (2003) discuss the many practical considerations involved in the choice of moisture variables.

for all $\mathbf{f}, \mathbf{g} \in \mathcal{H}$, called the covariance operator of $\mathbf{s}(t)$, and further that these are related by

$$\|\bar{\mathbf{s}}(t)\|^2 + \operatorname{tr} \mathcal{P}(t) = \mathcal{E}E(t)\,. \tag{3}$$

The covariance operator is self-adjoint and positive semidefinite, and also trace class. That is, the trace of $\mathcal{P}(t)$, defined for all $t \in \mathcal{T}$ by

$$\operatorname{tr} \mathcal{P}(t) = \sum_{i=1}^{\infty} (\mathbf{g}_i, \mathcal{P}(t)\mathbf{g}_i)\,,$$

where $\{\mathbf{g}_i\}_{i=1}^{\infty}$ is any countable orthonormal basis for $\mathcal{H}$, is finite and independent of basis. Further,

$$\operatorname{tr} \mathcal{P}(t) = \mathcal{E}\|\mathbf{s}(t) - \bar{\mathbf{s}}(t)\|^2\,, \tag{4}$$

for all $t \in \mathcal{T}$.

Equation (3) is the principle of energetic consistency (PEC) in its strong form.[2] It says that the sum of the total energy of the mean state and the total uncertainty in the system, as measured by the total variance $\operatorname{tr} \mathcal{P}(t)$, equals the expected value of the total energy of the system. It is clear that each of the three terms in Eq. (3) is a property of the dynamical system itself, not of the choice of state variables.

If $\|\mathbf{s}(t)\|^2$ is not a scalar invariant but the other two assumptions are satisfied for all $t \in \mathcal{T}$, i.e., if $\mathbf{s}(t)$ is an $\mathcal{H}$-valued random variable, with $\mathcal{H}$ a real, separable Hilbert space, and if $\mathcal{E}\|\mathbf{s}(t)\|^2 < \infty$, then the mean state and covariance operator still exist uniquely and satisfy

$$||\bar{\mathbf{s}}(t)||^2 + \operatorname{tr} \mathcal{P}(t) = \mathcal{E}||\mathbf{s}(t)||^2$$

for all $t \in \mathcal{T}$. This is a weak version of the PEC. It is weak because none of the terms here has an intrinsic physical meaning: each measures a property only of the state variables $s_1, \ldots, s_n$ chosen to describe the dynamical system, not a property of

---

[2] To derive the PEC in the finite-dimensional case, apply the expectation operator to the identity

$$||\mathbf{s}||^2 = ||\bar{\mathbf{s}}||^2 + 2(\bar{\mathbf{s}}, \mathbf{s}') + ||\mathbf{s}'||^2\,,$$

where $\mathbf{s}' = \mathbf{s} - \bar{\mathbf{s}}$ and the time argument has been omitted, to obtain

$$||\bar{\mathbf{s}}||^2 + \mathcal{E}||\mathbf{s}'||^2 = \mathcal{E}||\mathbf{s}||^2\,.$$

Since $(\mathbf{f}, \mathbf{g}) = \mathbf{f}^T\mathbf{B}\mathbf{g}$ for some symmetric positive definite matrix $\mathbf{B}$ and all vectors $\mathbf{f}, \mathbf{g} \in \mathcal{H}$ in case $\mathcal{H}$ is finite-dimensional, it follows from the definition given by Eq. (2) that in this case the covariance operator $\mathcal{P}$ has matrix representation $\mathbf{P} = \mathcal{E}\mathbf{s}'\mathbf{s}'^T\mathbf{B}$, where the expectation operator applied to a matrix of random variables is defined to act elementwise as usual. Therefore

$$\mathcal{E}||\mathbf{s}'||^2 = \mathcal{E}\mathbf{s}'^T\mathbf{B}\mathbf{s}' = \mathcal{E}\operatorname{tr}\mathbf{s}'\mathbf{s}'^T\mathbf{B} = \operatorname{tr}\mathcal{E}\mathbf{s}'\mathbf{s}'^T\mathbf{B} = \operatorname{tr}\mathbf{P}\,.$$

the system itself. In particular, the total variance $\mathrm{tr}\,\mathcal{P}(t)$ is not an intrinsic property of the system if $\|\mathbf{s}(t)\|^2$ is not a scalar invariant of the system.

The assumption that $\|\mathbf{s}(t)\|^2$ is a scalar invariant makes the principle of energetic consistency (Eq. 3) a strong statement about the dynamical system itself. It is also a general statement: no assumptions have been made on the system dynamics other than the simple ones stated above, whose purpose is mainly to guarantee existence of two moments of the system state. Similarly, the probability distribution function of $(\mathbf{g}, \mathbf{s}(t))$ for any $\mathbf{g} \in \mathcal{H}$ and $t \in \mathcal{T}$ has been left essentially free.

That only the first two moments of the system state appear in the PEC is due to the fact that, by definition, the total energy $E(t) = \|\mathbf{s}(t)\|^2$ is quadratic in the energy variables $s_1, \ldots, s_n$. Note that the total energy integral for a hydrostatic atmosphere happens to be a cubic polynomial in the variables $u$, $v$, $T$ and $p*$, which are typical model variables for hydrostatic models. Also, for the shallow-water equations, the total energy integral happens to be a cubic polynomial in the variables $u$, $v$ and $\Phi$. It is not difficult to show that for a total energy integral that is an $m$th-order polynomial in some state variables $\widetilde{s}_1, \ldots, \widetilde{s}_n$, there is a relationship much like the PEC among moments of these state variables up to order $m$ only, provided all the moments up to order $m$ exist, and the expected value of the total energy. A theory based on such a relationship would lack the simplicity of the one presented in this chapter, which relies heavily on the assumption that the total energy is the square of a Hilbert space norm. More importantly from a practical point of view, 4D-Var and EnKF methods are designed to approximate the evolution of just the first two moments of the system state: the PEC as stated is suited specifically to current data assimilation practice. Further, choosing the state variables to be energy variables is natural in the case of EnKF methods, because these are based on the minimum variance optimality criterion, as discussed next.

## 2.2 Minimum Variance State Estimation

In addition to the assumptions stated in Sect. 2.1, assume now that for all $t \in \mathcal{T}$, the state vector $\mathbf{s}(t)$ is jointly distributed with some real, random $p_i$-vectors $\mathbf{y}_i = \mathbf{y}(t_i)$, $i = 1, \ldots, k$, called observation vectors, where $t_1 < \cdots < t_k$ are time instants in $\mathcal{T}$. This means simply that the $p_i$ components of each vector $\mathbf{y}_i$ are scalar random variables on the probability space $(\Omega, \mathcal{F}, P)$. This is the case, for instance, if the observations are related to the state according to

$$\mathbf{y}(t_i) = \mathbf{h}_{t_i}(\mathbf{s}(t_i); \mathbf{w}(t_i)) \tag{5}$$

for $i = 1, \ldots, k$, where $\mathbf{w}(t_1), \ldots, \mathbf{w}(t_k)$ are real random vectors, provided that the observation operators $\mathbf{h}_{t_1}, \ldots, \mathbf{h}_{t_k}$ are continuous in both arguments. Note that observation operators that are linear in variables such as winds, temperature or surface

pressure are non-linear in the energy variables. The time $t_k$ can be thought of as representing the current observation time. Let $\mathbf{y}^k = (\mathbf{y}_1^T, \ldots, \mathbf{y}_k^T)^T$ denote the random $p^k$-vector consisting of all the observations up to time $t_k$, where $p^k = \sum_{i=1}^k p_i$.

Recall that if $r$ is a scalar random variable, then the conditional expectation $\mathcal{E}(r|\mathbf{y}^k)$ is also a scalar random variable. Thus, $\mathcal{E}((\mathbf{g}, \mathbf{s}(t))|\mathbf{y}^k)$ is a scalar random variable, for each $\mathbf{g} \in \mathcal{H}$ and all $t \in \mathcal{T}$. In fact, since $\mathcal{E}E(t) = \mathcal{E}\|\mathbf{s}(t)\|^2 < \infty$ for all $t \in \mathcal{T}$, it follows that there exists an $\mathcal{H}$-valued random variable $\bar{\mathbf{s}}^k(t)$, called the conditional mean of $\mathbf{s}(t)$, such that for all $\mathbf{g} \in \mathcal{H}$ and $t \in \mathcal{T}$,

$$\mathcal{E}||\bar{\mathbf{s}}^k(t)||^2 < \infty\,,$$

and

$$(\mathbf{g}, \bar{\mathbf{s}}^k(t)) = \mathcal{E}((\mathbf{g}, \mathbf{s}(t))|\mathbf{y}^k) \tag{6}$$

with probability one.[3] Similarly, there exists a bounded, self-adjoint, positive semi-definite, trace class, random linear operator $\mathcal{P}^k(t) : \mathcal{H} \to \mathcal{H}$, called the conditional covariance operator of $\mathbf{s}(t)$, such that for all $\mathbf{f}, \mathbf{g} \in \mathcal{H}$ and $t \in \mathcal{T}$,

$$(\mathbf{f}, \mathcal{P}^k(t)\mathbf{g}) = \mathcal{E}\left[(\mathbf{f}, \mathbf{s}(t) - \bar{\mathbf{s}}^k(t))(\mathbf{g}, \mathbf{s}(t) - \bar{\mathbf{s}}^k(t))|\mathbf{y}^k\right] \tag{7}$$

with probability one, and also

---

[3] This follows from Appendix 1d. Note first that

$$s^k[\mathbf{g}] = \mathcal{E}((\mathbf{g}, \mathbf{s})|\mathbf{y}^k)\,,$$

where the time argument has been omitted, defines a random linear functional on $\mathcal{H}$. Let $\{\mathbf{g}_i\}_{i=1}^\infty$ be a countable orthonormal basis for $\mathcal{H}$. Then

$$\left(s^k[\mathbf{g}_i]\right)^2 \le \mathcal{E}((\mathbf{g}_i, \mathbf{s})^2|\mathbf{y}^k)$$

by the Schwarz inequality, and taking expectations gives

$$\mathcal{E}\left(s^k[\mathbf{g}_i]\right)^2 \le \mathcal{E}(\mathbf{g}_i, \mathbf{s})^2\,,$$

for $i = 1, 2, \ldots$. Therefore,

$$\sum_{i=1}^\infty \mathcal{E}\left(s^k[\mathbf{g}_i]\right)^2 \le \sum_{i=1}^\infty \mathcal{E}(\mathbf{g}_i, \mathbf{s})^2 = \mathcal{E}\sum_{i=1}^\infty (\mathbf{g}_i, \mathbf{s})^2 = \mathcal{E}||\mathbf{s}||^2 < \infty\,.$$

Hence by the construction of Appendix 1d, there exists an $\mathcal{H}$-valued random variable $\bar{\mathbf{s}}^k$ such that $\mathcal{E}\|\bar{\mathbf{s}}^k\|^2 < \infty$ and, for all $\mathbf{g} \in \mathcal{H}$,

$$(\mathbf{g}, \bar{\mathbf{s}}^k) = s^k[\mathbf{g}]$$

with probability one. The construction shows that $\bar{\mathbf{s}}^k$ is defined uniquely on the set of $\omega \in \Omega$ where $\sum_{i=1}^\infty \left(s^k[\mathbf{g}_i]\right)^2 < \infty$, which must have probability measure one.

$$\operatorname{tr} \mathcal{P}^k(t) = \mathcal{E}(\|\mathbf{s}(t) - \bar{\mathbf{s}}^k(t)\|^2 | \mathbf{y}^k) \tag{8}$$

with probability one.[4]

It follows that, for all $t \in \mathcal{T}$,

$$\|\bar{\mathbf{s}}^k(t)\|^2 + \operatorname{tr} \mathcal{P}^k(t) = \mathcal{E}(E(t)|\mathbf{y}^k) \tag{9}$$

with probability one. Equation (9) extends the principle of energetic consistency (Eq. 3) to include the effect of observing the dynamical system. Each of the terms in Eq. (9) is a scalar random variable, accounting for the observations, and each measures a property of the observed dynamical system. If $\|\mathbf{s}(t)\|^2$ is not a scalar invariant but the remaining assumptions are satisfied, with $\mathcal{E}\|\mathbf{s}(t)\|^2 < \infty$ for all $t \in \mathcal{T}$, then the corresponding weak version

$$||\bar{\mathbf{s}}^k(t)||^2 + \operatorname{tr} \mathcal{P}^k(t) = \mathcal{E}(||\mathbf{s}(t)||^2 | \mathbf{y}^k)$$

still holds, with probability one, for all $t \in \mathcal{T}$. However, the terms here then have no intrinsic physical meaning.

Now let $\tilde{\mathbf{s}}^k(t)$ be any $\mathcal{H}$-valued random variable depending on the observations $\mathbf{y}^k$, such that

$$\mathcal{E}\|\tilde{\mathbf{s}}^k(t)\|^2 < \infty \tag{10}$$

---

[4] To see this, first define the conditional covariance functional

$$C^k[\mathbf{f}, \mathbf{g}] = \mathcal{E}\left[(\mathbf{f}, \mathbf{s} - \bar{\mathbf{s}}^k)(\mathbf{g}, \mathbf{s} - \bar{\mathbf{s}}^k)|\mathbf{y}^k\right],$$

where the time argument has been omitted. The functional $C^k$ is a symmetric, positive semidefinite, random bilinear functional on $\mathcal{H}$. As in Eq. (64) of Appendix 1c,

$$\left|C^k[\mathbf{f}, \mathbf{g}]\right| \leq ||\mathbf{f}||\,||\mathbf{g}||\,\mathcal{E}(||\mathbf{s} - \bar{\mathbf{s}}^k||^2 | \mathbf{y}^k)$$

for all $\mathbf{f}, \mathbf{g} \in \mathcal{H}$. Therefore, for each $\omega \in \Omega$ where $\mathcal{E}(\|\mathbf{s} - \bar{\mathbf{s}}^k\|^2 | \mathbf{y}^k) < \infty$, there exists a unique bounded linear operator $\mathcal{P}^k : \mathcal{H} \to \mathcal{H}$ such that $(\mathbf{f}, \mathcal{P}^k \mathbf{g}) = C^k[\mathbf{f}, \mathbf{g}]$ for all $\mathbf{f}, \mathbf{g} \in \mathcal{H}$, and this operator is self-adjoint, positive semidefinite, and trace class, with

$$\operatorname{tr} \mathcal{P}^k = \mathcal{E}(||\mathbf{s} - \bar{\mathbf{s}}^k||^2 | \mathbf{y}^k).$$

But $\mathcal{E}(\|\mathbf{s} - \bar{\mathbf{s}}^k\|^2 | \mathbf{y}^k) < \infty$ with probability one, since

$$\mathcal{E}||\mathbf{s} - \bar{\mathbf{s}}^k||^2 \leq 2\mathcal{E}||\mathbf{s}||^2 + 2\mathcal{E}||\bar{\mathbf{s}}^k||^2 < \infty$$

by the parallelogram law. Thus the set of $\omega \in \Omega$ where $\mathcal{E}\left(\|\mathbf{s} - \bar{\mathbf{s}}^k\|^2 | \mathbf{y}^k\right) = \infty$ has probability measure zero. Upon defining $\mathcal{P}^k$ to be the zero operator on this set, it follows that $\mathcal{P}^k$ is bounded, self-adjoint, positive semidefinite and trace class for all $\omega \in \Omega$, and that Eqs. (7) and (8) hold with probability one.

for all $t \in \mathcal{T}$, which will be called an *estimate* of the system state $\mathbf{s}(t)$. It follows from the parallelogram law (see Appendix 3d) that

$$\mathcal{E}||\mathbf{s}(t) - \widetilde{\mathbf{s}}^k(t)||^2 \leq 2\mathcal{E}E(t) + 2\mathcal{E}||\widetilde{\mathbf{s}}^k(t)||^2 < \infty\,.$$

The scalar random variable $\|\mathbf{s}(t) - \widetilde{\mathbf{s}}^k(t)\|^2$ is the total energy of the *estimation error* $\mathbf{s}(t) - \widetilde{\mathbf{s}}^k(t)$. An estimate of $\mathbf{s}(t)$ is called a *minimum variance estimate* if it minimizes the expected value of the total energy of the estimation error, $\mathcal{E}\|\mathbf{s}(t) - \widetilde{\mathbf{s}}^k(t)\|^2$, over all estimates $\widetilde{\mathbf{s}}^k(t)$. Thus, by definition, the property of being a minimum variance estimate is an intrinsic property of the observed dynamical system.

The conditional mean state $\bar{\mathbf{s}}^k(t)$ is an $\mathcal{H}$-valued random variable depending on $\mathbf{y}^k$, and it was already shown to satisfy Eq. (10). Thus the conditional mean state is an estimate of the system state. Furthermore, by essentially the same argument as in the finite-dimensional case (e.g. Jazwinski 1970, p. 149, Theorem 5.3 or Cohn 1997, pp. 282–283), one has

$$\mathcal{E}||\mathbf{s}(t) - \widetilde{\mathbf{s}}^k(t)||^2 = \mathcal{E}||\mathbf{s}(t) - \bar{\mathbf{s}}^k(t)||^2 + \mathcal{E}||\bar{\mathbf{s}}^k(t) - \widetilde{\mathbf{s}}^k(t)||^2$$

for all $t \in \mathcal{T}$.[5] Therefore,

---

[5] This follows by taking expectations on the identity

$$||\mathbf{s} - \widetilde{\mathbf{s}}^k||^2 = ||\mathbf{s} - \bar{\mathbf{s}}^k||^2 + 2(\mathbf{s} - \bar{\mathbf{s}}^k, \bar{\mathbf{s}}^k - \widetilde{\mathbf{s}}^k) + ||\bar{\mathbf{s}}^k - \widetilde{\mathbf{s}}^k||^2\,,$$

where the time argument has been omitted, and noting that $\mathcal{E}(\mathbf{s} - \bar{\mathbf{s}}^k, \bar{\mathbf{s}}^k - \widetilde{\mathbf{s}}^k) = 0$ since

$$\mathcal{E}\left[(\mathbf{s} - \bar{\mathbf{s}}^k, \bar{\mathbf{s}}^k - \widetilde{\mathbf{s}}^k)|\mathbf{y}^k\right] = 0$$

with probability one. The latter equality can be shown in the infinite-dimensional case as follows. Let $\{\mathbf{g}_i\}_{i=1}^{\infty}$ be a countable orthonormal basis for $\mathcal{H}$. Then

$$\begin{aligned}\mathcal{E}\left[(\mathbf{s} - \bar{\mathbf{s}}^k, \bar{\mathbf{s}}^k - \widetilde{\mathbf{s}}^k)|\mathbf{y}^k\right] &= \mathcal{E}\left[\sum_{i=1}^{\infty}(\mathbf{g}_i, \mathbf{s} - \bar{\mathbf{s}}^k)(\mathbf{g}_i, \bar{\mathbf{s}}^k - \widetilde{\mathbf{s}}^k)|\mathbf{y}^k\right] \\ &= \sum_{i=1}^{\infty}\mathcal{E}\left[(\mathbf{g}_i, \mathbf{s} - \bar{\mathbf{s}}^k)(\mathbf{g}_i, \bar{\mathbf{s}}^k - \widetilde{\mathbf{s}}^k)|\mathbf{y}^k\right],\end{aligned}$$

since

$$\mathcal{E}\sum_{i=1}^{\infty}|(\mathbf{g}_i, \mathbf{s} - \bar{\mathbf{s}}^k)(\mathbf{g}_i, \bar{\mathbf{s}}^k - \widetilde{\mathbf{s}}^k)| \leq \left(\mathcal{E}||\mathbf{s} - \bar{\mathbf{s}}^k||^2\right)^{1/2}\left(\mathcal{E}||\bar{\mathbf{s}}^k - \widetilde{\mathbf{s}}^k||^2\right)^{1/2} < \infty\,;$$

cf. Doob (1953, Property $CE_5$, p. 23). But for each $i = 1, 2, \ldots,$

$$\mathcal{E}\left[(\mathbf{g}_i, \mathbf{s} - \bar{\mathbf{s}}^k)(\mathbf{g}_i, \bar{\mathbf{s}}^k - \widetilde{\mathbf{s}}^k)|\mathbf{y}^k\right] = (\mathbf{g}_i, \bar{\mathbf{s}}^k - \widetilde{\mathbf{s}}^k)\,\mathcal{E}\left[(\mathbf{g}_i, \mathbf{s} - \bar{\mathbf{s}}^k)|\mathbf{y}^k\right] = 0$$

with probability one, since $\bar{\mathbf{s}}^k - \widetilde{\mathbf{s}}^k$ depends only on $\mathbf{y}^k$ and since $\mathcal{E}\left[(\mathbf{g}_i, \mathbf{s} - \bar{\mathbf{s}}^k)|\mathbf{y}^k\right] = 0$ with probability one.

$$\mathcal{E}||\mathbf{s}(t) - \widetilde{\mathbf{s}}^k(t)||^2 \geq \mathcal{E}||\mathbf{s}(t) - \bar{\mathbf{s}}^k(t)||^2$$

for all $t \in \mathcal{T}$, with equality if, and only if, $\widetilde{\mathbf{s}}^k(t) = \bar{\mathbf{s}}^k(t)$ with probability one. Thus the conditional mean state is always a minimum variance state estimate, and any minimum variance state estimate is identical to the conditional mean state with probability one.

The conditional mean state has the following additional properties. First, taking expectations in Eq. (6) and using Eq. (1) gives

$$\mathcal{E}(\mathbf{g}, \bar{\mathbf{s}}^k(t)) = \mathcal{E}(\mathbf{g}, \mathbf{s}(t)) = (\mathbf{g}, \bar{\mathbf{s}}(t))$$

for all $\mathbf{g} \in \mathcal{H}$ and $t \in \mathcal{T}$. Thus the conditional mean state $\bar{\mathbf{s}}^k(t)$ has mean $\bar{\mathbf{s}}(t)$. Equivalently, the conditional mean state is an unbiased estimate of the system state.

It follows that

$$\mathcal{E}||\bar{\mathbf{s}}^k(t) - \bar{\mathbf{s}}(t)||^2 = \mathcal{E}||\bar{\mathbf{s}}^k(t)||^2 - ||\bar{\mathbf{s}}(t)||^2$$

for all $t \in \mathcal{T}$. Taking expectations in Eq. (9) gives

$$\mathcal{E}\|\bar{\mathbf{s}}^k(t)\|^2 + \mathcal{E}\,\mathrm{tr}\,\mathcal{P}^k(t) = \mathcal{E}E(t) \tag{11}$$

for all $t \in \mathcal{T}$, which is yet another extension of the PEC. Combining these two results with Eqs. (3) and (4) gives

$$\mathcal{E}||\mathbf{s}(t) - \bar{\mathbf{s}}(t)||^2 = \mathrm{tr}\,\mathcal{P}(t) = \mathcal{E}\,\mathrm{tr}\,\mathcal{P}^k(t) + \mathcal{E}||\bar{\mathbf{s}}^k(t) - \bar{\mathbf{s}}(t)||^2\,.$$

Therefore,

$$\mathcal{E}\,\mathrm{tr}\,\mathcal{P}^k(t) \leq \mathrm{tr}\,\mathcal{P}(t) \tag{12}$$

for all $t \in \mathcal{T}$. This means that in the expected value sense, the act of observing can only reduce total variance or, possibly, leave it unchanged. Also,

$$\mathcal{E}\|\bar{\mathbf{s}}^k(t) - \bar{\mathbf{s}}(t)\|^2 \leq \mathcal{E}\|\mathbf{s}(t) - \bar{\mathbf{s}}(t)\|^2 \tag{13}$$

for all $t \in \mathcal{T}$. This means that the conditional mean state can only be more concentrated about its mean than is the system state itself or, possibly, as concentrated. The inequalities given by Eqs. (12) and (13) still hold if $\|\mathbf{s}(t)\|^2$ is not a scalar invariant, but in that case the inequalities have no physical interpretation intrinsic to the dynamical system.

## 2.3 Discretization

The principle of energetic consistency (Eq. 3) and its extension to include the effect of observations (Eq. 9) are general relationships that apply independently of the continuum dynamics and the observations. However, they are not yet quite in a form directly applicable to computational methods for data assimilation, which are necessarily discrete. To maintain generality, a generic discretization will now be introduced.

Let $\mathcal{H}^N$ be an $N$-dimensional subspace of $\mathcal{H}$. For instance, $\mathcal{H}^N$ could be the space of all $\mathbf{g} \in \mathcal{H}$ that are constant on grid boxes of a numerical model of the continuum dynamics, or the space of all $\mathbf{g} \in \mathcal{H}$ obtained by a fixed spectral truncation. Then $\mathcal{H}^N$ is a (finite-dimensional) Hilbert space, under the same inner product and corresponding norm as that of $\mathcal{H}$.

Let $\Pi$ be the orthogonal projection operator from $\mathcal{H}$ onto $\mathcal{H}^N$. Thus $\Pi\mathbf{g} \in \mathcal{H}^N$ and $(\Pi\mathbf{f}, \mathbf{g} - \Pi\mathbf{g}) = 0$ for all $\mathbf{f}, \mathbf{g} \in \mathcal{H}$. Denote by $\mathbf{s}_r(t) = \Pi\mathbf{s}(t)$ the "resolved" part of the state $\mathbf{s}(t)$ and by $\mathbf{s}_u(t) = \mathbf{s}(t) - \mathbf{s}_r(t)$ the "unresolved" part, for all $t \in \mathcal{T}$. Then the total energy $E(t) = \|\mathbf{s}(t)\|^2$ is the sum of the total energy in the resolved scales, $E_r(t) = \|\mathbf{s}_r(t)\|^2$, and that in the unresolved scales, $E_u(t) = \|\mathbf{s}_u(t)\|^2$:

$$E(t) = E_r(t) + E_u(t)\,,$$

for all $t \in \mathcal{T}$. The components $s_1, \ldots, s_N$ of $\mathbf{s}_r(t)$ will be called discretized energy variables.

From the definition of $\mathbf{s}_r(t)$ and the fact that $\mathbf{s}(t)$ is an $\mathcal{H}$-valued random variable it follows that $\mathbf{s}_r(t)$ is an $\mathcal{H}^N$-valued random variable, and further that

$$\mathcal{E}||\mathbf{s}_r(t)||^2 = \mathcal{E}E_r(t) \leq \mathcal{E}E(t) < \infty$$

for all $t \in \mathcal{T}$. Therefore, $\mathbf{s}_r(t)$ has mean $\bar{\mathbf{s}}_r(t) \in \mathcal{H}^N$ and covariance operator $\mathcal{P}_r(t) : \mathcal{H}^N \to \mathcal{H}^N$, defined uniquely for all $t \in \mathcal{T}$ by the relationships

$$(\mathbf{g}, \bar{\mathbf{s}}_r(t)) = \mathcal{E}(\mathbf{g}, \mathbf{s}_r(t)) \tag{14}$$

and

$$(\mathbf{f}, \mathcal{P}_r(t)\mathbf{g}) = \mathcal{E}\left[(\mathbf{f}, \mathbf{s}_r(t) - \bar{\mathbf{s}}_r(t))(\mathbf{g}, \mathbf{s}_r(t) - \bar{\mathbf{s}}_r(t))\right]\,, \tag{15}$$

respectively, for all $\mathbf{f}, \mathbf{g} \in \mathcal{H}^N$. It follows that the principle of energetic consistency holds for $\mathbf{s}_r(t)$:

$$\|\bar{\mathbf{s}}_r(t)\|^2 + \operatorname{tr}\mathcal{P}_r(t) = \mathcal{E}E_r(t) \tag{16}$$

for all $t \in \mathcal{T}$, where

$$\operatorname{tr} \mathcal{P}_r(t) = \sum_{i=1}^{N} (\mathbf{g}_i, \mathcal{P}_r(t)\mathbf{g}_i) = \mathcal{E}||\mathbf{s}_r(t) - \bar{\mathbf{s}}_r(t)||^2 ,$$

and $\{\mathbf{g}_i\}_{i=1}^N$ is any orthonormal basis for $\mathcal{H}^N$.

The discretized PEC (Eq. 16) can be written in some equivalent ways. For instance, it can be verified that $\bar{\mathbf{s}}_r(t) = \Pi\bar{\mathbf{s}}(t)$ and that $\mathcal{P}_r(t) = \Pi\widetilde{\mathcal{P}}(t)$, where $\widetilde{\mathcal{P}}(t) : \mathcal{H}^N \to \mathcal{H}$ denotes the restriction of $\mathcal{P}(t)$ to $\mathcal{H}^N$, i.e., $\widetilde{\mathcal{P}}(t)\mathbf{g} = \mathcal{P}(t)\mathbf{g}$ for all $\mathbf{g} \in \mathcal{H}^N$. Also, viewing elements of $\mathcal{H}^N$ as real $N$-vectors, the inner product on $\mathcal{H}^N$ must be given by a real, symmetric positive definite matrix $\mathbf{B}$,

$$(\mathbf{f}, \mathbf{g}) = \mathbf{f}^T\mathbf{B}\mathbf{g} ,$$

with corresponding norm $\|\mathbf{f}\| = (\mathbf{f}, \mathbf{f})^{1/2}$, for all $\mathbf{f}, \mathbf{g} \in \mathcal{H}^N$.[6] Then $\mathbf{s}_r(t)$ is viewed as a vector of real random variables and it follows from the definition given by Eq. (14) that

$$\bar{\mathbf{s}}_r(t) = \mathcal{E}\mathbf{s}_r(t) ,$$

where the expectation operator applied to a vector of random variables is defined to act componentwise, so that

$$||\bar{\mathbf{s}}_r(t)||^2 = \bar{\mathbf{s}}_r^T(t)\mathbf{B}\bar{\mathbf{s}}_r(t)$$

for all $t \in \mathcal{T}$. Further, it follows from the definition given by Eq. (15) that $\mathcal{P}_r(t)$ has matrix representation

$$\mathbf{P}_r(t) = \mathcal{E}\left[(\mathbf{s}_r(t) - \bar{\mathbf{s}}_r(t))(\mathbf{s}_r(t) - \bar{\mathbf{s}}_r(t))^T\right]\mathbf{B} ,$$

where the expectation operator applied to a matrix of random variables is defined to act elementwise, so that

---

[6] For instance, if $\mathcal{H}^N$ consists of the elements of $\mathcal{H}$ that are constant on grid volumes $V_j$ of a numerical model of hydrostatic atmospheric dynamics with an unstaggered grid, then the matrix $\mathbf{B}$ is the block-diagonal matrix with diagonal blocks

$$\mathbf{B}_j = \int\int\int_{V_j} \mathbf{A}\, \mathrm{d}\sigma\, a^2 \mathrm{d}S ,$$

where the diagonal matrix $\mathbf{A}$ was defined in Sect. 2.1. In the general case, $\mathcal{H}^N$ is isometrically isomorphic to the Hilbert space $\mathcal{G}^N$ of real $N$-vectors with the stated inner product and corresponding norm; cf. Reed and Simon (1972, Theorem II.7, p. 47). Thus, viewing the elements of $\mathcal{H}^N$ as real $N$-vectors means it is understood that an isometric isomorphism has been applied to elements of $\mathcal{H}^N$ to obtain elements of $\mathcal{G}^N$. Then $\mathcal{H}^N$-valued random variables become $\mathcal{G}^N$-valued random variables, because an isometric isomorphism is norm-continuous.

$$\operatorname{tr} \mathcal{P}_r(t) = \operatorname{tr} \mathbf{P}_r(t)$$

for all $t \in \mathcal{T}$.

Just as Eq. (9) followed from Eq. (3), it follows from Eq. (16) that

$$\|\bar{\mathbf{s}}_r^k(t)\|^2 + \operatorname{tr} \mathcal{P}_r^k(t) = \mathcal{E}\left(E_r(t)|\mathbf{y}^k\right) \tag{17}$$

with probability one, for all $t \in \mathcal{T}$. Here the conditional mean $\bar{\mathbf{s}}_r^k(t)$ of $\mathbf{s}_r(t)$ is an $\mathcal{H}^N$-valued random variable with $\mathcal{E}\|\bar{\mathbf{s}}_r^k(t)\|^2 < \infty$ and satisfies

$$\left(\mathbf{g}, \bar{\mathbf{s}}_r^k(t)\right) = \mathcal{E}\left((\mathbf{g}, \mathbf{s}_r(t))|\mathbf{y}^k\right)$$

with probability one, for all $t \in \mathcal{T}$ and $\mathbf{g} \in \mathcal{H}^N$. The conditional covariance operator $\mathcal{P}_r^k(t) : \mathcal{H}^N \to \mathcal{H}^N$ of $\mathbf{s}_r(t)$ is a bounded, self-adjoint, positive semidefinite, trace class, random linear operator satisfying

$$\left(\mathbf{f}, \mathcal{P}_r^k(t)\mathbf{g}\right) = \mathcal{E}\left[(\mathbf{f}, \mathbf{s}_r(t) - \bar{\mathbf{s}}_r^k(t))(\mathbf{g}, \mathbf{s}_r(t) - \bar{\mathbf{s}}_r^k(t))|\mathbf{y}^k\right]$$

and

$$\operatorname{tr} \mathcal{P}_r^k(t) = \mathcal{E}\left(||\mathbf{s}_r(t) - \bar{\mathbf{s}}_r^k(t)||^2|\mathbf{y}^k\right),$$

both with probability one, for all $t \in \mathcal{T}$ and $\mathbf{f}, \mathbf{g} \in \mathcal{H}^N$. The equivalent ways of writing Eq. (16) described in the preceding paragraph apply similarly to Eq. (17). For instance, the discrete conditional mean state can be defined as $\bar{\mathbf{s}}_r^k(t) = \Pi\bar{\mathbf{s}}^k(t)$, and the discrete conditional covariance operator $\mathcal{P}_r^k(t)$ can be represented as a random matrix, i.e., a matrix of random variables. It is clear that the discrete conditional mean state is an unbiased estimate of the discrete system state $\mathbf{s}_r(t)$.

Finally, let $\tilde{\mathbf{s}}_r^k(t)$ denote any $\mathcal{H}^N$-valued random variable depending on the observations $\mathbf{y}^k$, such that $\mathcal{E}\|\tilde{\mathbf{s}}_r^k(t)\|^2 < \infty$ for all $t \in \mathcal{T}$. As in Sect. 2.2, it follows that the expected value of the total energy of the estimation error $\mathbf{s}_r(t) - \tilde{\mathbf{s}}_r^k(t)$ is given by

$$\mathcal{E}||\mathbf{s}_r(t) - \tilde{\mathbf{s}}_r^k(t)||^2 = \mathcal{E}||\mathbf{s}_r(t) - \bar{\mathbf{s}}_r^k(t)||^2 + \mathcal{E}||\bar{\mathbf{s}}_r^k(t) - \tilde{\mathbf{s}}_r^k(t)||^2 < \infty$$

for all $t \in \mathcal{T}$, and is therefore minimized (uniquely, with probability one) over all estimates $\tilde{\mathbf{s}}_r^k(t)$ by the discrete conditional mean state $\bar{\mathbf{s}}_r^k(t)$. With this minimization as the optimality criterion, the objective of data assimilation is thus to calculate $\bar{\mathbf{s}}_r^k(t)$. Taking expectations in Eq. (17) gives

$$\mathcal{E}||\bar{\mathbf{s}}_r^k(t)||^2 + \mathcal{E} \operatorname{tr} \mathcal{P}_r^k(t) = \mathcal{E}E_r(t)$$

for all $t \in \mathcal{T}$, from which discrete counterparts of Eqs. (12) and (13) follow.

One difficulty in attempting to calculate $\bar{\mathbf{s}}_r^k(t)$ through data assimilation is that the observations, for instance as given by Eq. (5), depend on both the resolved and

unresolved parts of the continuum system state $\mathbf{s}(t)$. This dependence leads in turn to the observation error due to unresolved scales, which is part of the so-called representativeness error (e.g. Janjić and Cohn 2006, and references therein).

## *2.4 Application to Ensemble Kalman Filter Methods*

### 2.4.1 General Formulation

Suppose now that a discrete model of the non-linear continuum dynamics is given, in the general form

$$\mathbf{x}_{t_i} = \mathcal{M}_{t_i,t_{i-1}}(\mathbf{x}_{t_{i-1}})$$

for $i = 1, \ldots, K$, with $t_K > t_k$, where $\mathcal{M}_{t_i,t_{i-1}} : \mathcal{D}_{\mathbf{x}} \to \mathcal{D}_{\mathbf{x}}$ is continuous and $\mathcal{D}_{\mathbf{x}}$ is a domain (a connected open set) in $\mathbb{R}^N$. Typically the state variables $x_1, \ldots, x_N$ are not discretized energy variables as defined in the preceding subsection, but there is a known, continuous and invertible transformation $\widetilde{\mathbf{x}} = \widetilde{\mathbf{x}}(\mathbf{x})$ from $\mathcal{D}_{\mathbf{x}}$ to a domain $\mathcal{D}_{\widetilde{\mathbf{x}}} \in \mathbb{R}^N$, with continuous inverse $\mathbf{x} = \mathbf{x}(\widetilde{\mathbf{x}})$, such that $\tilde{x}_1, \ldots, \tilde{x}_N$ are discretized energy variables. For instance, the simple transformation from typical atmospheric model variables $(u, v, T, p_*)$ defined on the model grid to gridded energy variables $(u\sqrt{p_*}, v\sqrt{p_*}, \sqrt{Tp_*}, \sqrt{p_*})$ has the required properties, provided that $T$ and $p_*$ remain bounded from below by positive constants. Then the given model is equivalent to the model

$$\widetilde{\mathbf{x}}_{t_i} = \widetilde{\mathcal{M}}_{t_i,t_{i-1}}(\widetilde{\mathbf{x}}_{t_{i-1}})$$

for $i = 1, \ldots, K$, where $\widetilde{\mathcal{M}}_{t_i,t_{i-1}} : \mathcal{D}_{\widetilde{\mathbf{x}}} \to \mathcal{D}_{\widetilde{\mathbf{x}}}$ is continuous and is obtained as the composition

$$\widetilde{\mathcal{M}}_{t_i,t_{i-1}} = \widetilde{\mathbf{x}} \circ \mathcal{M}_{t_i,t_{i-1}} \circ \mathbf{x}\,,$$

for $i = 1, \ldots, K$. Applying the principle of energetic consistency to a given model thus requires no real change to the model, but only a change of variables before and after each observation time to process the observations, which in general are related non-linearly to the energy variables as in Eq. (5). Henceforth the tildes will be omitted, including that for the domain $\mathcal{D}_{\widetilde{\mathbf{x}}}$, and it is to be understood that the model variables are discretized energy variables.

Denote by $\mathcal{G}^N$ the Hilbert space of real $N$-vectors with inner product $(\mathbf{f}, \mathbf{g}) = \mathbf{f}^T\mathbf{B}\mathbf{g}$ and corresponding norm $\|\mathbf{f}\| = (\mathbf{f}, \mathbf{f})^{1/2}$, for all $\mathbf{f}, \mathbf{g} \in \mathcal{G}^N$, where $\mathbf{B}$ is the real, symmetric positive definite matrix defined in the preceding subsection. Note that the same symbols are used for the inner product and norm on $\mathcal{H}$, but no confusion should arise because the context will be clear. View the $\mathcal{H}^N$-valued random variable $\mathbf{s}_r(t_0)$ of the preceding subsection as a $\mathcal{G}^N$-valued random variable and let

$$\mathbf{x}_{t_0} = \mathbf{s}_r(t_0)\,,$$

where it is assumed that $\mathbf{s}_r(t_0) \in \mathcal{D}_{\mathbf{x}}$. Denote by

$$E_d(t_i) = ||\mathbf{x}_{t_i}||^2$$

the total energy of the resulting discrete model state, for $i = 0, \ldots, K$. Since $\mathcal{M}_{t_i,t_{i-1}}$ is continuous, $\mathbf{x}_{t_i}$ is a $\mathcal{G}^N$-valued random variable and $E_d(t_i)$ is a scalar random variable, for $i = 1, \ldots, K$.

Suppose now that the non-linear continuum dynamics are conservative, $E(t) = E(t_0)$ for all $t \in \mathcal{T} = [t_0, T]$, as is the case for dry hydrostatic atmospheric dynamics and for shallow-water dynamics. In this case the principle of energetic consistency (Eq. 3) reads

$$||\bar{\mathbf{s}}(t)||^2 + \operatorname{tr} \mathcal{P}(t) = \mathcal{E}E(t_0) = \text{const.}\,.$$

Suppose also for now that the given discrete dynamical model is conservative,

$$||\mathcal{M}_{t_i,t_{i-1}}(\mathbf{f})||^2 = ||\mathbf{f}||^2$$

for all $\mathbf{f} \in \mathcal{D}_{\mathbf{x}}$ and for $i = 1, \ldots, K$. Then

$$\mathcal{E}E_d(t_i) = \mathcal{E}E_d(t_0) = \mathcal{E}E_r(t_0) < \infty \tag{18}$$

for $i = 1, \ldots, K$. It follows immediately that the mean state $\overline{\mathbf{x}}_{t_i} = \mathcal{E}\mathbf{x}_{t_i} \in \mathcal{G}^N$ and covariance matrix $\mathbf{P}_{t_i} : \mathcal{G}^N \to \mathcal{G}^N$ defined by

$$\mathbf{P}_{t_i} = \mathcal{E}\left[(\mathbf{x}_{t_i} - \overline{\mathbf{x}}_{t_i})(\mathbf{x}_{t_i} - \overline{\mathbf{x}}_{t_i})^T\right]\mathbf{B}$$

exist for $i = 0, \ldots, K$, and that they are related by

$$\|\overline{\mathbf{x}}_{t_i}\|^2 + \operatorname{tr} \mathbf{P}_{t_i} = \mathcal{E}E_r(t_0) = \text{const.}\,, \tag{19}$$

for $i = 0, \ldots, K$. Thus, for both the continuum state and the modelled discrete state, the sum of the total energy of the mean state and the total variance is constant in time. Equation (19) is somewhat at odds with Eq. (16), whose right-hand side need not be constant in time when the continuum dynamics are conservative.

It follows similarly from Eq. (18) that

$$\|\overline{\mathbf{x}}^k_{t_i}\|^2 + \operatorname{tr} \mathbf{P}^k_{t_i} = \mathcal{E}\left(E_r(t_0)|\mathbf{y}^k\right) \tag{20}$$

with probability one, for $i = 0, \ldots, K$. Here $\overline{\mathbf{x}}^k_{t_i}$ and $\mathbf{P}^k_{t_i}$ are, respectively, the mean state and covariance matrix of $\mathbf{x}_{t_i}$ conditioned on the observations $\mathbf{y}^k$, and are defined as in Sects. 2.2 and 2.3. The right-hand side of Eq. (20) is independent of the time $t_i$, so that in particular,

$$\|\overline{\mathbf{x}}^k_{t_{k+1}}\|^2 + \operatorname{tr} \mathbf{P}^k_{t_{k+1}} = \|\overline{\mathbf{x}}^k_{t_k}\|^2 + \operatorname{tr} \mathbf{P}^k_{t_k} \tag{21}$$

with probability one. In traditional filtering notation this is written

$$||\mathbf{x}^f_{k+1}||^2 + \operatorname{tr} \mathbf{P}^f_{k+1} = ||\mathbf{x}^a_k||^2 + \operatorname{tr} \mathbf{P}^a_k$$

with probability one, where $\mathbf{x}^a_k$ and $\mathbf{P}^a_k$ are, respectively, the conditional mean analysis and conditional analysis error covariance matrix at time $t_k$, and $\mathbf{x}^f_{k+1}$ and $\mathbf{P}^f_{k+1}$ are, respectively, the conditional mean forecast and conditional forecast error covariance matrix to time $t_{k+1}$, where all the conditioning is on the observations up to time $t_k$.

### 2.4.2 Ensemble Behaviour Between Observation Times

An ensemble version of Eq. (21) is satisfied exactly, independently of ensemble size $L$, by an appropriately formulated EnKF scheme. Assume that $\overline{\mathbf{x}}^k_{t_k} \in \mathcal{D}_{\mathbf{x}}$ with probability one.[7] Then let $\left\{\mathbf{x}^k_{t_k}(l)\right\}_{l=1}^L$ be a sample of the $\mathcal{G}^N$-valued random variable $\overline{\mathbf{x}}^k_{t_k}$ with $\mathbf{x}^k_{t_k}(l) \in \mathcal{D}_{\mathbf{x}}$ for $l = 1, \ldots, L$, and define

$$\mathbf{x}^k_{t_{k+1}}(l) = \mathcal{M}_{t_{k+1},t_k}\left(\mathbf{x}^k_{t_k}(l)\right)$$

for $l = 1, \ldots, L$. Also, for $i = k$ and $i = k+1$, define

$$\widehat{E}^k_{t_i} = \frac{1}{L}\sum_{l=1}^{L} ||\mathbf{x}^k_{t_i}(l)||^2\,,$$
$$\widehat{\mathbf{x}}^k_{t_i} = \frac{1}{L}\sum_{l=1}^{L} \mathbf{x}^k_{t_i}(l)\,,$$

and

$$\widehat{\mathbf{P}}^k_{t_i} = \frac{1}{L}\sum_{l=1}^{L}\left(\mathbf{x}^k_{t_i}(l) - \widehat{\mathbf{x}}^k_{t_i}\right)\left(\mathbf{x}^k_{t_i}(l) - \widehat{\mathbf{x}}^k_{t_i}\right)^T \mathbf{B}\,.$$

By manipulating the sums it follows that

$$\widehat{E}^k_{t_i} = \|\widehat{\mathbf{x}}^k_{t_i}\|^2 + \operatorname{tr} \widehat{\mathbf{P}}^k_{t_i} \tag{22}$$

for $i = k$ and $i = k+1$. But $\|\mathbf{x}^k_{t_{k+1}}(l)\|^2 = \|\mathbf{x}^k_{t_k}(l)\|^2$ for $l = 1, \ldots, L$, since $\mathcal{M}_{t_{k+1},t_k}$ is conservative, and therefore

[7] If $\mathcal{D}_{\mathbf{x}}$ is convex then $\overline{\mathbf{x}}_{t_i} \in \mathcal{D}_{\mathbf{x}}$ for $i = 0, \ldots, K$ (e.g. Cohn 2009, p. 454), and therefore $\overline{\mathbf{x}}^k_{t_i} \in \mathcal{D}_{\mathbf{x}}$ with probability one, for $i = 0, \ldots, K$.

$$\widehat{E}^k_{t_{k+1}} = \widehat{E}^k_{t_k} .$$

This statement of energy conservation thus reads

$$\|\widehat{\mathbf{x}}^k_{t_{k+1}}\|^2 + \operatorname{tr} \widehat{\mathbf{P}}^k_{t_{k+1}} = \|\widehat{\mathbf{x}}^k_{t_k}\|^2 + \operatorname{tr} \widehat{\mathbf{P}}^k_{t_k} , \tag{23}$$

which corresponds to Eq. (21).

Equation (22) does not hold if the conditional covariance estimate $\widehat{\mathbf{P}}^k_{t_i}$ is replaced there by the sample covariance $L/(L-1)\widehat{\mathbf{P}}^k_{t_i}$. Therefore, Eq. (23) does not generally hold if either or both conditional covariance estimates there are replaced by the corresponding sample covariances. The sample covariance $L/(L-1)\widehat{\mathbf{P}}^k_{t_{k+1}}$ has been used traditionally in EnKF schemes, on the grounds that it is unbiased as an estimator of $\mathbf{P}^k_{t_{k+1}}$, but it is clear that this violates energy conservation by artificially increasing the total energy $\widehat{E}^k_{t_{k+1}}$ of the ensemble at each observation time. This increase can be significant for typically small ensemble sizes. For instance, for $L = 100$ it is about 0.1% per observation time, or more than 4% per 10 days with observations every 6 h, in case $\operatorname{tr} \widehat{\mathbf{P}}^k_{t_{k+1}} = 0.1\|\widehat{\mathbf{x}}^k_{t_{k+1}}\|^2$. Von Storch and Zwiers (1999, p. 87) give other reasons why the sample covariance should be used only with caution in general.

Now let $\mathbf{C}$ be an $N \times N$ correlation matrix, i.e., a symmetric positive semidefinite matrix with unit diagonal. Then since the trace of a square matrix is the sum of its diagonal elements, it follows that

$$\operatorname{tr} (\mathbf{C} \circ \mathbf{P}) = \operatorname{tr} \mathbf{P}$$

for any $N \times N$ matrix $\mathbf{P}$, where the symbol $\circ$ is used to denote the Hadamard (elementwise) product of two matrices. Thus, Eq. (23) still holds if $\widehat{\mathbf{P}}^k_{t_{k+1}}$ is replaced there by a "localized" conditional covariance estimate

$$\widetilde{\mathbf{P}}^k_{t_{k+1}} = \mathbf{C} \circ \widehat{\mathbf{P}}^k_{t_{k+1}} .$$

The covariance localization approach introduced by Houtekamer and Mitchell (2001, Eq. 6) and studied further by Mitchell et al. (2002) approximates this formula, reducing computational effort, but the degree to which the approximation might in effect violate energy conservation is not known. The effect of the alternative localization approach of Ott et al. (2004) on energy conservation is also not known.

### 2.4.3 Ensemble Behaviour at Observation Times

To see what is supposed to happen at observation times, assume for the moment that $E(t_0)$ is simply a constant, i.e., is not a random variable. Thus each realization of the continuum system state has the same total energy at the initial time $t_0$, hence at all times $t \in \mathcal{T}$ since the continuum dynamics were assumed to be conservative. Then

assume that $E_d(t_0) = E_r(t_0)$ is also a constant. Thus each realization of the modelled discrete state has the same total energy,

$$||\mathbf{x}_{t_i}||^2 = E_d(t_i) = E_d(t_0) = \text{const.}$$

for $i = 0, \ldots, K$, since the discrete model was also assumed to be conservative. On taking conditional expectations it follows that

$$\mathcal{E}(\|\mathbf{x}_{t_i}\|^2 | \mathbf{y}^k) = E_d(t_0) = \text{const.} \tag{24}$$

with probability one, for $i = 0, \ldots, K$, independently of the relationship between the observations and the continuum state. For an EnKF scheme, this means that the analysis update is supposed to leave the total energy of each ensemble member unchanged, at each observation time, regardless of how any assumed relationship between the observations and the discrete state is modelled. The assumption that $E_d(t_0)$ is constant can be implemented in an EnKF scheme by normalizing the total energy of each ensemble member to a constant at the initial time.

It follows from Eq. (24) that

$$\|\overline{\mathbf{x}}^k_{t_k}\|^2 + \operatorname{tr} \mathbf{P}^k_{t_k} = \|\overline{\mathbf{x}}^{k-1}_{t_k}\|^2 + \operatorname{tr} \mathbf{P}^{k-1}_{t_k} \tag{25}$$

with probability one, where $\overline{\mathbf{x}}^{k-1}_{t_k}$ and $\mathbf{P}^{k-1}_{t_k}$ are, respectively, the mean state and covariance matrix at time $t_k$, both conditioned on the observations up to time $t_{k-1}$. In traditional filtering notation, this is written

$$||\mathbf{x}^a_k||^2 + \operatorname{tr} \mathbf{P}^a_k = ||\mathbf{x}^f_k||^2 + \operatorname{tr} \mathbf{P}^f_k$$

with probability one. It can be verified that the usual Kalman-type analysis update formula for discrete linear observation operators does not satisfy this relationship. The reason it does not is that, interpreted probabilistically, the Kalman formula assumes that the discrete state is Gaussian-distributed, which is not possible if its total energy is a constant. On the other hand, on taking expectations in Eq. (25), it follows that

$$\mathcal{E}\|\overline{\mathbf{x}}^k_{t_k}\|^2 + \mathcal{E} \operatorname{tr} \mathbf{P}^k_{t_k} = \mathcal{E}\|\overline{\mathbf{x}}^{k-1}_{t_k}\|^2 + \mathcal{E} \operatorname{tr} \mathbf{P}^{k-1}_{t_k}\,, \tag{26}$$

which is satisfied for the linear Kalman update formula. Ensemble implementations of the Kalman formula do not leave the total energy of each ensemble member unchanged at observation times, and do not satisfy the ensemble version of Eq. (25), viz.,

$$||\widehat{\mathbf{x}}^k_{t_k}||^2 + \operatorname{tr} \widehat{\mathbf{P}}^k_{t_k} = ||\widehat{\mathbf{x}}^{k-1}_{t_k}||^2 + \operatorname{tr} \widehat{\mathbf{P}}^{k-1}_{t_k}$$

but at least they should satisfy

$$\mathcal{E}\|\widehat{\mathbf{x}}_{t_k}^k\|^2 + \mathcal{E}\,\mathrm{tr}\,\widehat{\mathbf{P}}_{t_k}^k = \mathcal{E}\|\widehat{\mathbf{x}}_{t_k}^{k-1}\|^2 + \mathcal{E}\,\mathrm{tr}\,\widehat{\mathbf{P}}_{t_k}^{k-1}\,. \tag{27}$$

Verifying Eq. (27) would require carrying out numerical experiments with many random samples of the discrete initial state.

More generally now, suppose that there are constants $E_{\min}$ and $E_{\max}$ such that $E_{\min} \le E_d(t_0) \le E_{\max}$. This is the case if it is assumed that the total energy of every realization of the continuum initial state is bounded from above. It follows that

$$E_{\min} \le ||\overline{\mathbf{x}}_{t_k}^{k-1}||^2 + \mathrm{tr}\,\mathbf{P}_{t_k}^{k-1} \le E_{\max}$$

and

$$E_{\min} \le \|\overline{\mathbf{x}}_{t_k}^{k}\|^2 + \mathrm{tr}\,\mathbf{P}_{t_k}^{k} \le E_{\max}\,, \tag{28}$$

both with probability one, independently of the relationship between the observations and the continuum state or the discrete state, actual or assumed. Again, ensemble implementations of the Kalman update formula cannot satisfy

$$E_{\min} \le ||\widehat{\mathbf{x}}_{t_k}^{k}||^2 + \mathrm{tr}\,\widehat{\mathbf{P}}_{t_k}^{k} \le E_{\max}\,,$$

because a Gaussian-distributed state cannot have total energy bounded from above by a constant. However, Eq. (28) implies that

$$E_{\min} \le \mathcal{E}||\overline{\mathbf{x}}_{t_k}^{k}||^2 + \mathcal{E}\,\mathrm{tr}\,\mathbf{P}_{t_k}^{k} \le E_{\max}\,,$$

which is satisfied for the linear Kalman update formula. Therefore, an ensemble implementation of this formula should satisfy

$$E_{\min} \le \mathcal{E}||\widehat{\mathbf{x}}_{t_k}^{k}||^2 + \mathcal{E}\,\mathrm{tr}\,\widehat{\mathbf{P}}_{t_k}^{k} \le E_{\max}\,.$$

### 2.4.4 Ensemble Behaviour for Dissipative Models

Numerical models almost always exhibit some spurious dissipation. Typical numerical dissipation is mild and may be self-limiting in the context of deterministic prediction (e.g. Lin and Rood 1997, p. 2490; Lin 2004, p. 2303), but it can pose a serious problem in the context of filtering. Ménard et al. (2000, pp. 2658–2661) have found for a full Kalman filter for assimilating tracer observations on isentropic surfaces that a small, spurious dissipation in the numerical advection model causes a large, state-dependent loss of total variance (with the total variance defined slightly differently there than in the present chapter), even without assimilating the observations. Ménard and Chang (2000, p. 2676) found, moreover, that this spurious loss of variance is made worse by the assimilation of observations. That the loss of total variance due to spurious numerical dissipation is a generic problem for filtering can be understood in the EnKF context in the following way.

The individual ensemble members $\{\mathbf{x}^k_{t_k}(l)\}^L_{l=1}$ are supposed to be spatially somewhat rough, particularly near the observations, since they are supposed to be samples from a probability distribution that includes the effect of observation error, which is local. The ensemble mean $\widehat{\mathbf{x}}^k_{t_k}$ is more spatially smooth than the individual ensemble members since it is their average. Thus the spurious loss of total energy of each of the ensemble perturbations $\{\mathbf{x}^k_{t_k}(l) - \widehat{\mathbf{x}}^k_{t_k}\}^L_{l=1}$ up to the next observing time is usually far more than for a "typical" state, such as the ensemble mean, since a significant fraction of the total energy of each of the ensemble perturbations is supposed to be concentrated near grid scale, where numerical dissipation usually acts most strongly. The total variance is just the ensemble average of the total energy of the ensemble perturbations, and thus is usually lost to spurious numerical dissipation much more rapidly than is the total energy of the ensemble mean. Moreover, this loss of total variance need not be self-limiting, because the ensemble analysis update is supposed to inject energy into the perturbations, from the observation error near grid scale, at each observation time, only to be dissipated away again. If this argument is correct, then for large times $t_k - t_0$ one should expect an exponential decay of total variance.

Ménard et al. (2000) addressed the problem of loss of total variance due to spurious numerical dissipation by utilizing the fact that for tracer dynamics there is a partial differential equation for variance evolution that can be discretized directly. This was found to give results superior to simply adding an artificial "model error" term for instance (Ménard and Chang 2000, p. 2682). Unfortunately, such an equation does not exist for much more general dynamics. The argument above can be formalized in the following way, which also leads to a general approach for addressing the problem.

Suppose that the given discrete model is dissipative, in the sense that

$$||\mathcal{M}_{t_{k+1},t_k}(\mathbf{x})||^2 \leq ||\mathbf{x}||^2$$

for all $\mathbf{x} \in \mathcal{D}_{\mathbf{x}}$. Here $\mathcal{M}_{t_{k+1},t_k}$ will be thought of as the solution operator from time $t_k$ to time $t_{k+1}$ for a system of ordinary differential equations,

$$\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} + \mathbf{f}(\mathbf{x}, t) = \mathbf{d}(\mathbf{x}, t)\,,$$

where $\mathrm{d}\mathbf{x}/\mathrm{d}t + \mathbf{f} = \mathbf{0}$ is a conservative model of the continuum dynamics, $(\mathbf{x}, \mathbf{f}(\mathbf{x}, t)) = 0$ for all $\mathbf{x} \in \mathcal{D}_{\mathbf{x}}$, and where $\mathbf{d}$ is dissipative and defined throughout $\mathcal{G}^N$,

$$(\mathbf{x}, \mathbf{d}(\mathbf{x}, t)) \leq 0 \tag{29}$$

for all $\mathbf{x} \in \mathcal{G}^N$. Thus,

$$\frac{\mathrm{d}}{\mathrm{d}t}||\mathbf{x}||^2 = \frac{\mathrm{d}}{\mathrm{d}t}(\mathbf{x}, \mathbf{x}) = 2\left(\mathbf{x}, \frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t}\right) = 2(\mathbf{x}, \mathbf{d}(\mathbf{x}, t)) \leq 0\,.$$

The ensemble members $\{\mathbf{x}(l)\}^L_{l=1}$ are supposed to satisfy the equation

$$\frac{\mathrm{d}\mathbf{x}(l)}{\mathrm{d}t} + \mathbf{f}(\mathbf{x}(l)) = \mathbf{d}(\mathbf{x}(l)),$$

where time arguments are omitted for notational convenience. Then the ensemble mean $\widehat{\mathbf{x}} = \frac{1}{L}\sum_{l=1}^{L}\mathbf{x}(l)$ satisfies

$$\frac{\mathrm{d}\widehat{\mathbf{x}}}{\mathrm{d}t} + \widehat{\mathbf{f}} = \widehat{\mathbf{d}},$$

where $\widehat{\mathbf{f}} = \frac{1}{L}\sum_{l=1}^{L}\mathbf{f}(\mathbf{x}(l))$ and $\widehat{\mathbf{d}} = \frac{1}{L}\sum_{l=1}^{L}\mathbf{d}(\mathbf{x}(l))$. Assume that $\mathbf{d}$ is linear in a neighbourhood of $\widehat{\mathbf{x}}$ that includes all the ensemble members, i.e.,

$$\mathbf{d}(\mathbf{x}(l)) = \mathbf{d}(\widehat{\mathbf{x}}) + \mathbf{D}(\mathbf{x}(l) - \widehat{\mathbf{x}}) \tag{30}$$

for $l = 1, \ldots, L$, where $\mathbf{D}$ is the Jacobian matrix

$$\mathbf{D} = \mathbf{D}_{\widehat{\mathbf{x}}} = \left.\frac{\partial \mathbf{d}(\mathbf{x})}{\partial \mathbf{x}}\right|_{\mathbf{x}=\widehat{\mathbf{x}}}. \tag{31}$$

Taking ensemble averages in Eq. (30) gives $\widehat{\mathbf{d}} = \mathbf{d}(\widehat{\mathbf{x}})$, and therefore

$$\frac{\mathrm{d}(\mathbf{x}(l) - \widehat{\mathbf{x}})}{\mathrm{d}t} + \mathbf{f}(\mathbf{x}(l)) - \widehat{\mathbf{f}} = \mathbf{D}(\mathbf{x}(l) - \widehat{\mathbf{x}}) \tag{32}$$

for $l = 1, \ldots, L$. Thus,

$$\frac{\mathrm{d}}{\mathrm{d}t}||\mathbf{x}(l) - \widehat{\mathbf{x}}||^2 = -2(\mathbf{x}(l) - \widehat{\mathbf{x}}, \mathbf{f}(\mathbf{x}(l)) - \widehat{\mathbf{f}}) + 2(\mathbf{x}(l) - \widehat{\mathbf{x}}, \mathbf{D}(\mathbf{x}(l) - \widehat{\mathbf{x}}))$$

for $l = 1, \ldots, L$. Since the total variance is

$$\operatorname{tr}\widehat{\mathbf{P}} = \frac{1}{L}\sum_{l=1}^{L}||\mathbf{x}(l) - \widehat{\mathbf{x}}||^2,$$

and since $(\mathbf{x}(l), \mathbf{f}(\mathbf{x}(l))) = 0$ for $l = 1, \ldots, L$, it follows that

$$\frac{1}{2}\frac{\mathrm{d}\operatorname{tr}\widehat{\mathbf{P}}}{\mathrm{d}t} = (\widehat{\mathbf{x}}, \widehat{\mathbf{f}}) + \frac{1}{L}\sum_{l=1}^{L}(\mathbf{x}(l) - \widehat{\mathbf{x}}, \mathbf{D}(\mathbf{x}(l) - \widehat{\mathbf{x}})). \tag{33}$$

Note that in case $\mathbf{f}$ is linear, then $(\widehat{\mathbf{x}}, \widehat{\mathbf{f}}) = (\widehat{\mathbf{x}}, \mathbf{f}(\widehat{\mathbf{x}})) = 0$.

Now, if $\mathbf{d}$ is linear over *all* of $\mathcal{G}^N$, then the Jacobian matrix $\mathbf{D}$ is independent of $\widehat{\mathbf{x}}$, and also $\mathbf{D}(\mathbf{x}(l) - \widehat{\mathbf{x}}) = \mathbf{d}(\mathbf{x}(l) - \widehat{\mathbf{x}})$. Therefore, if Eq. (29) holds as a strict inequality for at least one of the ensemble perturbations, i.e., if

$$(\mathbf{x}(l) - \widehat{\mathbf{x}}, \mathbf{d}(\mathbf{x}(l) - \widehat{\mathbf{x}})) < 0$$

for at least one $l \in [1, L]$, then

$$(\mathbf{x}(l) - \widehat{\mathbf{x}}, \mathbf{D}(\mathbf{x}(l) - \widehat{\mathbf{x}})) = (\mathbf{x}(l) - \widehat{\mathbf{x}}, \mathbf{d}(\mathbf{x}(l) - \widehat{\mathbf{x}})) < 0$$

for at least one $l \in [1, L]$, so it follows from Eq. (33) that

$$\frac{1}{2}\frac{\mathrm{d}\,\mathrm{tr}\,\widehat{\mathbf{P}}}{\mathrm{d}t} < (\widehat{\mathbf{x}}, \widehat{\mathbf{f}})\,. \tag{34}$$

Thus the effect of linear model dissipation is to reduce the total variance $\mathrm{tr}\,\widehat{\mathbf{P}}$. However, spurious numerical dissipation is typically non-linear, particularly when expressed in terms of energy variables. Also, Ménard et al. (2000, Fig. 1) found the loss of total variance to depend strongly on the state estimate, which in the present context requires the Jacobian matrix $\mathbf{D}$ to depend on the ensemble mean $\widehat{\mathbf{x}}$. Thus $\mathbf{d}$ has been allowed to be non-linear, and is assumed for the purpose of this analysis to be linear only in a neighbourhood of the ensemble mean. This local linearity assumption (Eq. 30) can be justified if the total energy of each of the ensemble perturbations is small relative to the total energy of the ensemble mean.

Instead of taking $\mathbf{d}$ to be linear, assume that

$$\frac{1}{L}\sum_{l=1}^{L}(\mathbf{x}(l), \mathbf{d}(\mathbf{x}(l))) \leq (1 + \varepsilon)(\widehat{\mathbf{x}}, \mathbf{d}(\widehat{\mathbf{x}})) \tag{35}$$

for some time-independent constant $\varepsilon > 0$, i.e., that the non-linear dissipation acts more strongly on the ensemble members, on average, than on the ensemble mean, for all time, as should be the case if the ensemble members are periodically updated with observation error near grid scale. Then from Eq. (30) and the fact that $\widehat{\mathbf{d}} = \mathbf{d}(\widehat{\mathbf{x}})$ it follows that

$$\begin{aligned}\frac{1}{L}\sum_{l=1}^{L}(\mathbf{x}(l) - \widehat{\mathbf{x}}, \mathbf{D}(\mathbf{x}(l) - \widehat{\mathbf{x}})) &= \frac{1}{L}\sum_{l=1}^{L}(\mathbf{x}(l) - \widehat{\mathbf{x}}, \mathbf{d}(\mathbf{x}(l)) - \mathbf{d}(\widehat{\mathbf{x}})) \\ &= \frac{1}{L}\sum_{l=1}^{L}(\mathbf{x}(l), \mathbf{d}(\mathbf{x}(l))) - (\widehat{\mathbf{x}}, \mathbf{d}(\widehat{\mathbf{x}})) \\ &\leq \varepsilon(\widehat{\mathbf{x}}, \mathbf{d}(\widehat{\mathbf{x}}))\,.\end{aligned} \tag{36}$$

If Eq. (29) holds as a strict inequality for the ensemble mean, i.e., if

$$(\widehat{\mathbf{x}}, \mathbf{d}(\widehat{\mathbf{x}})) < 0\,,$$

then it follows from Eqs. (33) and (36) that Eq. (34) still holds, and thus that the effect of the non-linear dissipation is indeed to reduce the total variance $\mathrm{tr}\,\widehat{\mathbf{P}}$.

To see why the effect can actually be an exponential loss of total variance, suppose that in fact there is a time-independent constant $\delta > 0$ such that

$$(\widehat{\mathbf{x}}, \mathbf{d}(\widehat{\mathbf{x}})) \leq -\delta\,.$$

It follows that there is a time-independent constant $C > 0$ such that

$$C \operatorname{tr} \widehat{\mathbf{P}} \leq -(\widehat{\mathbf{x}}, \mathbf{d}(\widehat{\mathbf{x}}))\,.$$

Combining this result with Eqs. (33) and (36) yields

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\operatorname{tr} \widehat{\mathbf{P}} \leq (\widehat{\mathbf{x}}, \widehat{\mathbf{f}}) - \varepsilon C \operatorname{tr} \widehat{\mathbf{P}}\,.$$

In case $\mathbf{f}$ is linear, so that $(\widehat{\mathbf{x}}, \widehat{\mathbf{f}}) = 0$, it follows that

$$\operatorname{tr} \widehat{\mathbf{P}}^k_{t_k} \leq e^{-2\varepsilon C(t_k - t_0)} \operatorname{tr} \widehat{\mathbf{P}}^0_{t_0}.$$

In the general nonlinear case, if $|(\widehat{\mathbf{x}}, \widehat{\mathbf{f}})| \leq \alpha$ for some time-independent constant $\alpha$, then

$$\operatorname{tr} \widehat{\mathbf{P}}^k_{t_k} \leq \frac{\alpha}{\varepsilon C}\left[1 - e^{-2\varepsilon C(t_k - t_0)}\right] + e^{-2\varepsilon C(t_k - t_0)} \operatorname{tr} \widehat{\mathbf{P}}^0_{t_0}.$$

Non-linearity can thus prevent decay to zero, although even the crude bound $\alpha/\varepsilon C$ may be small.

Spurious loss of total variance can be eliminated by undoing the spurious dissipation that acts on the ensemble perturbations. Denote by $\widetilde{\mathbf{P}}^k_{t_{k+1}}$ the ensemble covariance matrix that would have been obtained in case $\mathbf{D} = \mathbf{0}$, starting from $\widetilde{\mathbf{P}}^k_{t_k} = \widehat{\mathbf{P}}^k_{t_k}$. Then it follows from Eq. (32) that, to first order in $\Delta t_k = t_{k+1} - t_k$,

$$\widetilde{\mathbf{P}}^k_{t_{k+1}} = (\mathbf{I} - \Delta t_k \mathbf{D}) \widehat{\mathbf{P}}^k_{t_{k+1}} (\mathbf{I} - \Delta t_k \mathbf{D}^*)\,, \tag{37}$$

where $\widehat{\mathbf{P}}^k_{t_{k+1}}$ is the ensemble covariance matrix obtained with dissipation, $\mathbf{I}$ is the identity matrix, and $\mathbf{D}^*$ is the adjoint of $\mathbf{D}$ with respect to the inner product on $\mathcal{G}^N$,

$$\mathbf{D}^* = \mathbf{B}^{-1}\mathbf{D}^T\mathbf{B}\,.$$

The dissipation correction formula (Eq. 37) can be thought of as a generalization of the idea of covariance inflation (Anderson and Anderson 1999, p. 2747). Covariance inflation addresses general filter divergence problems simply by multiplying $\widehat{\mathbf{P}}^k_{t_{k+1}}$ by a number $\alpha = \alpha(\Delta t_k)$ slightly larger than one, thus amplifying all spatial scales equally. Covariance inflation did not perform well in experiments of Ménard et al. (2000, p. 2666), because it led to too much growth of variance away from the sparse observations. In the dissipation correction formula, the amplification is selective,

acting most strongly on the scales that have been most damped by spurious numerical dissipation, and only weakly or not at all on the larger scales where numerical models have little or no spurious dissipation.

For complex models, arriving at an appropriate formulation for the matrix $\mathbf{D} = \mathbf{D}_{\widehat{\mathbf{x}}}$ may not be a simple matter, and surely would involve some trial and error. However, an appropriate formulation would guarantee that the principle of energetic consistency is satisfied in the form of Eq. (23) extremely well on replacing $\widehat{\mathbf{P}}^k_{t_{k+1}}$ with $\widetilde{\mathbf{P}}^k_{t_{k+1}}$, provided that $|(\widehat{\mathbf{x}}^k_{t_k}, \mathbf{d}(\widehat{\mathbf{x}}^k_{t_k}))| \ll \|\widehat{\mathbf{x}}^k_{t_k}\|^2$.

## 3 The Principle of Energetic Consistency

### *3.1 Problem Setting*

Let $\mathcal{H}$ be a real, separable Hilbert space, with inner product and corresponding norm denoted by $(\cdot, \cdot)$ and $\|\cdot\|$, respectively. Recall from Appendix 3d that every separable Hilbert space has a countable orthonormal basis, and that every orthonormal basis of a separable Hilbert space has the same number of elements $N \leq \infty$, the dimension of the space. Let $\{\mathbf{h}_i\}_{i=1}^N$ be an orthonormal basis for $\mathcal{H}$, where $N = \dim \mathcal{H} \leq \infty$ is the dimension of $\mathcal{H}$.

Let $\mathcal{S}$ be any non-empty set in $\mathcal{B}(\mathcal{H})$, where $\mathcal{B}(\mathcal{H})$ denotes the Borel field generated by the open sets in $\mathcal{H}$, i.e., $\mathcal{B}(\mathcal{H})$ is the smallest $\sigma$-algebra of subsets of $\mathcal{H}$ containing all the sets that are open in $\mathcal{H}$. In particular, $\mathcal{S} \subset \mathcal{H}$, $\mathcal{S}$ can be all of $\mathcal{H}$, and $\mathcal{S}$ can be any open or closed set in $\mathcal{H}$.

Let $t_0$ and $T$ be two times with $-\infty < t_0 < T < \infty$, and let $\mathcal{T}$ be a time set bounded by and including $t_0$ and $T$. For instance, $\mathcal{T} = [t_0, T]$ in the case of continuous-time dynamics, and $\mathcal{T} = [t_0, t_1, \ldots, t_K = T]$ in the discrete-time case. The set $\mathcal{T}$ is allowed to depend on the set $\mathcal{S}$, $\mathcal{T} = \mathcal{T}(\mathcal{S})$.

Let $\mathbf{N}_{t,t_0}$ be a map from $\mathcal{S}$ into $\mathcal{H}$ (written $\mathbf{N}_{t,t_0} : \mathcal{S} \to \mathcal{H}$) for all times $t \in \mathcal{T}$, i.e., for all $\mathbf{s}_{t_0} \in \mathcal{S}$ and $t \in \mathcal{T}$, $\mathbf{N}_{t,t_0}(\mathbf{s}_{t_0})$ is defined and

$$\mathbf{s}_t = \mathbf{N}_{t,t_0}(\mathbf{s}_{t_0}) \tag{38}$$

is in $\mathcal{H}$, so that $\|\mathbf{s}_t\| < \infty$. Assume that $\mathbf{N}_{t,t_0}$ is continuous and bounded for all $t \in \mathcal{T}$. Continuity means that for every $t \in \mathcal{T}$, $\mathbf{s}_{t_0} \in \mathcal{S}$ and $\varepsilon > 0$, there is a $\delta > 0$ such that if $\|\mathbf{s}_{t_0} - \mathbf{s}'_{t_0}\| < \delta$ and $\mathbf{s}'_{t_0} \in \mathcal{S}$, then $\|\mathbf{N}_{t,t_0}(\mathbf{s}_{t_0}) - \mathbf{N}_{t,t_0}(\mathbf{s}'_{t_0})\| < \varepsilon$. Boundedness means that there is a constant $M = M_{t,t_0}$ such that

$$||\mathbf{N}_{t,t_0}(\mathbf{s}_{t_0})|| \leq M_{t,t_0}||\mathbf{s}_{t_0}||$$

for all $\mathbf{s}_{t_0} \in \mathcal{S}$ and $t \in \mathcal{T}$. Continuity and boundedness are equivalent if $\mathbf{N}_{t,t_0}$ is a linear operator.

In the applications of Sect. 4, $\mathbf{N}_{t,t_0}$ will be the solution operator of a well-posed initial-value problem, for the state vector $\mathbf{s}$ of a non-linear, deterministic

system of partial ($N = \dim \mathcal{H} = \infty$) or ordinary ($N = \dim \mathcal{H} < \infty$) differential equations ($\mathcal{T} = [t_0, T]$). Recall that continuity of the solution operator is part of the (Hadamard) definition of well-posedness of the initial-value problem for continuous-time or discrete-time dynamical systems: not only must there exist sets $\mathcal{S}$ and $\mathcal{T} = \mathcal{T}(\mathcal{S})$, taken here to be defined as above, and a unique solution $\mathbf{s}_t \in \mathcal{H}$ for all $\mathbf{s}_{t_0} \in \mathcal{S}$ and $t \in \mathcal{T}$, which taken together define the solution operator, but the solution must also depend continuously on the initial data. In Sect. 4.1 on ordinary differential equations, $\mathcal{H}$ will be Euclidean space $\mathbb{R}^N$ and the set $\mathcal{S}$ for the initial conditions will be an open subset of $\mathbb{R}^N$. In Sect. 4.2 on partial differential equations, $\mathcal{H}$ will be the space $L^2(D)$ of square-integrable vectors on the spatial domain $D$ of the problem, and $\mathcal{S}$ will be an open subset of an appropriate Sobolev space contained in $L^2(D)$.

The operator $\mathbf{N}_{t,t_0}$ is called isometric or conservative (in the norm $\|\cdot\|$ on $\mathcal{H}$) if

$$||\mathbf{N}_{t,t_0}(\mathbf{s}_{t_0})|| = ||\mathbf{s}_{t_0}||$$

for all $\mathbf{s}_{t_0} \in \mathcal{S}$ and $t \in \mathcal{T}$, and the differential (or difference) equations that express the dynamics of a well-posed initial-value problem are called conservative if the solution operator of the problem is conservative. With $\mathbf{s}_t \in \mathcal{H}$ defined for all $\mathbf{s}_{t_0} \in \mathcal{S}$ and $t \in \mathcal{T}$ by Eq. (38), the quantity

$$E_t = \|\mathbf{s}_t\|^2 = (\mathbf{s}_t, \mathbf{s}_t) < \infty \tag{39}$$

satisfies $E_t \leq M_{t,t_0}^2 E_{t_0}$ for all $t \in \mathcal{T}$ under the assumption of boundedness, and is constant in time, $E_t = E_{t_0}$ for all $t \in \mathcal{T}$, in the conservative case.

It will be seen in Sect. 3.3 that, in essence, the principle of energetic consistency is a statement about continuous, bounded transformations of Hilbert space, with conservative transformations as an important special case. Thus, applied to bounded solution operators, it becomes a statement about well-posed initial-value problems. It is important to recognize that the quantity $E_t$ defined in Eq. (39) is quadratic in $\mathbf{s}_t$. For non-linear systems of differential equations that express physical laws, there is often a choice of dependent (state) variables such that $E_t$ is the physical total energy, in which case the dynamics are conservative in the norm on $\mathcal{H}$ if the physical system is closed. However, in the rest of this chapter it will not be assumed that $E_t$ represents a physical total energy, nor that it is a scalar invariant. Rather, it will be simplest to proceed with the abstract hypotheses stated in the present subsection, and to treat the conservative case as special.

## *3.2 Scalar and Hilbert Space-Valued Random Variables*

Before stating the principle of energetic consistency in the setting of Sect. 3.1, some probability concepts will first be summarized. For details, see Appendices 1a–1c and 3c.

Let $(\Omega, \mathcal{F}, P)$ be a complete probability space, with $\Omega$ the sample space, $\mathcal{F}$ the event space and $P$ the probability measure. The event space consists of subsets of the set $\Omega$, called events or measurable sets, which are those subsets on which the probability measure is defined. Denote by $\mathcal{E}$ the expectation operator.

A (scalar) random variable is a map $r : \Omega \to \mathbb{R}^e$ that is measurable, i.e., an extended real-valued function $r$, defined for all $\omega \in \Omega$, that satisfies

$$\{\omega \in \Omega : r(\omega) \leq x\} \in \mathcal{F}$$

for all $x \in \mathbb{R}$. Thus, if $r$ is a random variable then its probability distribution function

$$F_r(x) = P(\{\omega \in \Omega : r(\omega) \leq x\})$$

is defined for all $x \in \mathbb{R}$. If $r$ is a random variable then $r^2$ is a random variable.

Suppose that $r$ is a random variable. Then the expectation $\mathcal{E}|r|$ is defined and $\mathcal{E}|r| \leq \infty$. If $\mathcal{E}|r| < \infty$, then the expectation $\mathcal{E}r$ is defined and called the mean of $r$, and $|\mathcal{E}r| \leq \mathcal{E}|r| < \infty$. If $\mathcal{E}r^2 < \infty$, then $r$ is called second-order, the mean $\bar{r} = \mathcal{E}r$ and variance $\sigma^2 = \mathcal{E}(r - \bar{r})^2$ of $r$ are defined, and

$$\mathcal{E}r^2 = \bar{r}^2 + \sigma^2 . \tag{40}$$

An $\mathcal{H}$-valued random variable is a map $\mathbf{r} : \Omega \to \mathcal{H}$ such that

$$\{\omega \in \Omega : \mathbf{r}(\omega) \in B\} \in \mathcal{F}$$

for every set $B \in \mathcal{B}(\mathcal{H})$. A map $\mathbf{r} : \Omega \to \mathcal{H}$ is an $\mathcal{H}$-valued random variable if, and only if, $(\mathbf{h}, \mathbf{r})$ is a scalar random variable for every $\mathbf{h} \in \mathcal{H}$, that is, if and only if

$$\{\omega \in \Omega : (\mathbf{h}, \mathbf{r}(\omega)) \leq x\} \in \mathcal{F}$$

for all $\mathbf{h} \in \mathcal{H}$ and $x \in \mathbb{R}$. If $\mathbf{r}$ is an $\mathcal{H}$-valued random variable then $\|\mathbf{r}\|$ is a scalar random variable. An $\mathcal{H}$-valued random variable $\mathbf{r}$ is called second-order if $\|\mathbf{r}\|$ is a second-order scalar random variable, i.e., if $\mathcal{E}\|\mathbf{r}\|^2 < \infty$. If $\mathbf{r}$ is a second-order $\mathcal{H}$-valued random variable then $(\mathbf{h}, \mathbf{r})$ is a second-order scalar random variable, i.e., $\mathcal{E}(\mathbf{h}, \mathbf{r})^2 < \infty$, for all $\mathbf{h} \in \mathcal{H}$.

Suppose that $\mathbf{r}$ is a second-order $\mathcal{H}$-valued random variable. Then there exists a unique element $\bar{\mathbf{r}} \in \mathcal{H}$, called the mean of $\mathbf{r}$, such that $\mathcal{E}(\mathbf{h}, \mathbf{r}) = (\mathbf{h}, \bar{\mathbf{r}})$ for all $\mathbf{h} \in \mathcal{H}$. Also, $\mathbf{r}' = \mathbf{r} - \bar{\mathbf{r}}$ is a second-order $\mathcal{H}$-valued random variable with mean $\mathbf{0} \in \mathcal{H}$, and

$$\mathcal{E}||\mathbf{r}||^2 = ||\bar{\mathbf{r}}||^2 + \mathcal{E}||\mathbf{r}'||^2 .$$

Furthermore, there exists a unique bounded linear operator $\mathcal{P} : \mathcal{H} \to \mathcal{H}$, called the covariance operator of $\mathbf{r}$, such that

$$\mathcal{E}(\mathbf{g}, \mathbf{r}')(\mathbf{h}, \mathbf{r}') = (\mathbf{g}, \mathcal{P}\mathbf{h})$$

for all $\mathbf{g}, \mathbf{h} \in \mathcal{H}$. The covariance operator $\mathcal{P}$ is self-adjoint and positive semidefinite, i.e., $(\mathbf{g}, \mathcal{P}\mathbf{h}) = (\mathcal{P}\mathbf{g}, \mathbf{h})$ and $(\mathbf{h}, \mathcal{P}\mathbf{h}) \geq 0$ for all $\mathbf{g}, \mathbf{h} \in \mathcal{H}$. It is also trace class, i.e., the sum $\sum_{i=1}^{N}(\mathbf{h}_i, \mathcal{P}\mathbf{h}_i)$ is finite and independent of the orthonormal basis $\{\mathbf{h}_i\}_{i=1}^{N}$, $N = \dim \mathcal{H} \leq \infty$, chosen for $\mathcal{H}$. This sum is called the trace of $\mathcal{P}$:

$$\operatorname{tr} \mathcal{P} = \sum_{i=1}^{N}(\mathbf{h}_i, \mathcal{P}\mathbf{h}_i) < \infty\,.$$

In addition, there exists an orthonormal basis for $\mathcal{H}$ which consists of eigenvectors $\{\tilde{\mathbf{h}}_i\}_{i=1}^{N}$ of $\mathcal{P}$,

$$\mathcal{P}\tilde{\mathbf{h}}_i = \lambda_i \tilde{\mathbf{h}}_i$$

for $i = 1, 2, \ldots, N$, and the corresponding eigenvalues $\{\lambda_i\}_{i=1}^{N}$ are all non-negative. It follows that

$$\lambda_i = (\tilde{\mathbf{h}}_i, \mathcal{P}\tilde{\mathbf{h}}_i) = \mathcal{E}(\tilde{\mathbf{h}}_i, \mathbf{r}')^2 = \sigma_i^2\,,$$

where $\sigma_i^2$ is the variance of the second-order scalar random variable $(\tilde{\mathbf{h}}_i, \mathbf{r})$, for $i = 1, 2, \ldots, N$, and that

$$\operatorname{tr} \mathcal{P} = \sum_{i=1}^{N} \sigma_i^2 = \mathcal{E}||\mathbf{r}'||^2\,.$$

Thus the trace of $\mathcal{P}$ is also called the total variance of the second-order $\mathcal{H}$-valued random variable $\mathbf{r}$, and

$$\mathcal{E}\|\mathbf{r}\|^2 = \|\bar{\mathbf{r}}\|^2 + \mathcal{E}\|\mathbf{r}'\|^2 = \|\bar{\mathbf{r}}\|^2 + \sum_{i=1}^{N} \sigma_i^2 = \|\bar{\mathbf{r}}\|^2 + \operatorname{tr} \mathcal{P}\,. \tag{41}$$

Equation (41) generalizes Eq. (40), which holds for second-order scalar random variables, to the case of second-order $\mathcal{H}$-valued random variables.

Suppose that $\mathcal{R} \in \mathcal{B}(\mathcal{H})$. An $\mathcal{R}$-valued random variable is a map $\mathbf{r} : \Omega \to \mathcal{R}$ such that

$$\{\omega \in \Omega : \mathbf{r}(\omega) \in C\} \in \mathcal{F}$$

for every set $C \in \mathcal{B}_{\mathcal{R}}(\mathcal{H})$, where

$$\mathcal{B}_{\mathcal{R}}(\mathcal{H}) = \{B \in \mathcal{B}(\mathcal{H}) : B \subset \mathcal{R}\}.$$

Every $\mathcal{R}$-valued random variable is an $\mathcal{H}$-valued random variable, and every $\mathcal{H}$-valued random variable $\mathbf{r}$ with $\mathbf{r}(\omega) \in \mathcal{R}$ for all $\omega \in \Omega$ is an $\mathcal{R}$-valued random variable. An $\mathcal{R}$-valued random variable $\mathbf{r}$ is called second-order if $\|\mathbf{r}\|$ is a second-order scalar random variable. Thus every second-order $\mathcal{R}$-valued random variable is a second-order $\mathcal{H}$-valued random variable, and every second-order $\mathcal{H}$-valued random variable $\mathbf{r}$ with $\mathbf{r}(\omega) \in \mathcal{R}$ for all $\omega \in \Omega$ is a second-order $\mathcal{R}$-valued random variable. Finally, if $\mathbf{r}$ is an $\mathcal{R}$-valued random variable and $\mathbf{N}$ is a continuous map from $\mathcal{R}$ into $\mathcal{H}$, then $\mathbf{N}(\mathbf{r})$ is an $\mathcal{H}$-valued random variable.

## *3.3 The Principle of Energetic Consistency in Hilbert Space*

Referring now back to Sect. 3.1, consider for $\mathbf{s}_{t_0}$ not just a single element of $\mathcal{S}$, but rather a whole collection of elements $\mathbf{s}_{t_0}(\omega)$ indexed by the probability variable $\omega \in \Omega$. Suppose at first that $\mathbf{s}_{t_0}$ is simply a map $\mathbf{s}_{t_0} : \Omega \to \mathcal{S}$, i.e., that $\mathbf{s}_{t_0}(\omega)$ is defined for all $\omega \in \Omega$ and $\mathbf{s}_{t_0}(\omega) \in \mathcal{S}$ for all $\omega \in \Omega$. Then since $\mathbf{N}_{t,t_0} : \mathcal{S} \to \mathcal{H}$ for all $t \in \mathcal{T}$, it follows that $\mathbf{s}_t = \mathbf{N}_{t,t_0}(\mathbf{s}_{t_0}) : \Omega \to \mathcal{H}$ for all $t \in \mathcal{T}$, with

$$\mathbf{s}_t(\omega) = \mathbf{N}_{t,t_0}(\mathbf{s}_{t_0}(\omega))$$

and $\|\mathbf{s}_t(\omega)\| < \infty$, for all $\omega \in \Omega$ and $t \in \mathcal{T}$.

Suppose further that $\mathbf{s}_{t_0}$ is an $\mathcal{S}$-valued random variable. Then it follows from the continuity assumption on $\mathbf{N}_{t,t_0}$ that $\mathbf{s}_t$ is an $\mathcal{H}$-valued random variable, and therefore that $E_t = \|\mathbf{s}_t\|^2$ is a scalar random variable, for all $t \in \mathcal{T}$.

Suppose still further that $\mathbf{s}_{t_0}$ is a second-order $\mathcal{S}$-valued random variable, $\mathcal{E}E_{t_0} = \mathcal{E}\|\mathbf{s}_{t_0}\|^2 < \infty$. Then from the boundedness assumption on $\mathbf{N}_{t,t_0}$,

$$||\mathbf{s}_t(\omega)||^2 \leq M_{t,t_0}^2 ||\mathbf{s}_{t_0}(\omega)||^2$$

for all $\omega \in \Omega$ and $t \in \mathcal{T}$, it follows that

$$\mathcal{E}E_t = \mathcal{E}||\mathbf{s}_t||^2 \leq M_{t,t_0}^2 \mathcal{E}||\mathbf{s}_{t_0}||^2 < \infty$$

for all $t \in \mathcal{T}$. Therefore, $\mathbf{s}_t$ is a second-order $\mathcal{H}$-valued random variable, with mean $\bar{\mathbf{s}}_t \in \mathcal{H}$, covariance operator $\mathcal{P}_t : \mathcal{H} \to \mathcal{H}$, and

$$\mathcal{E}||\mathbf{s}_t||^2 = ||\bar{\mathbf{s}}_t||^2 + \operatorname{tr} \mathcal{P}_t \,,$$

for all $t \in \mathcal{T}$. Thus the principle of energetic consistency has been established:

**Theorem 1** *Let $\mathcal{H}$, $\mathcal{S}$, $\mathcal{T}$ and $\mathbf{N}_{t,t_0}$ be as stated in Sect. 3.1, with $\mathbf{N}_{t,t_0}$ continuous and bounded for all $t \in \mathcal{T}$, and let $\mathcal{E}$ be the expectation operator on a complete probability space $(\Omega, \mathcal{F}, P)$. If $\mathbf{s}_{t_0}$ is a second-order $\mathcal{S}$-valued random variable, then for all $t \in \mathcal{T}$, (i) $\mathbf{s}_t = \mathbf{N}_{t,t_0}(\mathbf{s}_{t_0})$ is a second-order $\mathcal{H}$-valued random variable, (ii) $E_t = \|\mathbf{s}_t\|^2$ is a scalar random variable, (iii) $\mathbf{s}_t$ has mean $\bar{\mathbf{s}}_t \in \mathcal{H}$ and covariance*

*operator* $\mathcal{P}_t : \mathcal{H} \to \mathcal{H}$, *(iv)*

$$\mathcal{E}E_t = ||\bar{\mathbf{s}}_t||^2 + tr\,\mathcal{P}_t\,,$$

*and (v)*

$$\|\bar{\mathbf{s}}_t\|^2 + tr\,\mathcal{P}_t \leq M^2_{t,t_0}(\|\bar{\mathbf{s}}_{t_0}\|^2 + tr\,\mathcal{P}_{t_0})\,. \tag{42}$$

*If, in addition,* $\mathbf{N}_{t,t_0}$ *is conservative, then (vi)*

$$\|\bar{\mathbf{s}}_t\|^2 + tr\,\mathcal{P}_t = \|\bar{\mathbf{s}}_{t_0}\|^2 + tr\,\mathcal{P}_{t_0} \tag{43}$$

*for all* $t \in \mathcal{T}$.

It is in the conservative case that the principle of energetic consistency is most useful, because in that case, Eq. (43) provides an equality against which approximate moment evolution schemes can be compared, as discussed in Sect. 2.4 and in Cohn (2009). In case $\mathbf{N}_{t,t_0}$ is only bounded, for example in the presence of dissipation, or for initial-boundary value problems with a net flux of energy across the boundaries, Eq. (42) still provides an upper bound on the total variance $\mathrm{tr}\,\mathcal{P}_t$.

## *3.4 A Natural Restriction on* $\mathcal{S}$

Suppose for the moment that $\mathbf{s}_{t_0}$ is an $\mathcal{S}$-valued random variable, not necessarily second-order. When the squared norm on $\mathcal{H}$ represents a physical total energy, it is natural to impose the restriction that every possible initial state $\mathbf{s}_{t_0}(\omega)$, $\omega \in \Omega$, has total energy less than some finite maximum amount, say $E_* < \infty$, i.e., that $\mathcal{S} \subset \mathcal{H}_{E_*}$, where $\mathcal{H}_E$ is defined for all $E > 0$ as the open set

$$\mathcal{H}_E = \{\mathbf{s} \in \mathcal{H} : \|\mathbf{s}\|^2 < E\}\,. \tag{44}$$

Otherwise, given any total energy $E$, no matter how large, there would be a non-zero probability that $\mathbf{s}_{t_0}$ has total energy greater than or equal to $E$:

$$P(\{\omega \in \Omega : ||\mathbf{s}_{t_0}(\omega)||^2 \geq E\}) > 0.$$

Of course, it can be argued that since this probability would be very small for $E$ very large, it may be acceptable as an approximation not to impose this restriction. On the other hand, as discussed in Sect. 4.2 and illustrated in Sect. 5, for classical solutions of hyperbolic systems of partial differential equations, it is necessary to require that $\mathcal{S} \subset \mathcal{H}_{E_*}$ for some $E_* < \infty$ just to ensure well-posedness. Thus the restriction is often not only natural, but also necessary. It also simplifies matters, as discussed next, for it makes $\mathbf{s}_{t_0}$ second-order automatically and gives $\mathbf{s}_t = \mathbf{N}_{t,t_0}(\mathbf{s}_{t_0})$ some additional desirable properties, and it also yields a convenient characterization of $\mathbf{s}_{t_0}$ and $\mathbf{s}_t$.

Suppose that $\mathbf{s}_{t_0}$ is an $\mathcal{S}$-valued random variable, and that $\mathcal{S} \subset \mathcal{H}_{E_*}$ for some $E_* < \infty$. Thus $\|\mathbf{s}_{t_0}(\omega)\|^2 < E_*$ for all $\omega \in \Omega$, and therefore $\mathcal{E}\|\mathbf{s}_{t_0}\|^2 < E_*$, i.e., $\mathbf{s}_{t_0}$ is a second-order $\mathcal{S}$-valued random variable. Therefore, for all $t \in \mathcal{T}$, $\mathbf{s}_t = \mathbf{N}_{t,t_0}(\mathbf{s}_{t_0})$ is a second-order $\mathcal{H}$-valued random variable, in fact with $\mathbf{s}_t(\omega) \in \mathcal{H}_E$ for all $\omega \in \Omega$, where $E = E_*$ in the conservative case and $E = M^2_{t,t_0} E_*$ in the merely bounded case. Since $\mathcal{H}_E$ is an open set in $\mathcal{H}$, $\mathcal{H}_E \in \mathcal{B}(\mathcal{H})$. Therefore, for all $t \in \mathcal{T}$, $\mathbf{s}_t$ is an $\mathcal{H}_E$-valued random variable. Further, for all $p > 0$ and $t \in \mathcal{T}$, $\mathcal{E}\|\mathbf{s}_t\|^p < E^{p/2}$. Thus $\|\mathbf{s}_t\|$ has finite moments of all orders, for all $t \in \mathcal{T}$.

Now suppose that $\mathbf{s}$ is an $\mathcal{H}_E$-valued random variable, for some $E < \infty$. Then since $\mathbf{s}$ is also an $\mathcal{H}$-valued random variable, $(\mathbf{h}_i, \mathbf{s})$ is a scalar random variable for $i = 1, \ldots, N$, where $\{\mathbf{h}_i\}_{i=1}^N$ is any orthonormal basis for $\mathcal{H}$ and $N = \dim \mathcal{H} \leq \infty$. Since $\mathbf{s}(\omega) \in \mathcal{H}$ for all $\omega \in \Omega$, $\mathbf{s}(\omega)$ has the representation

$$\mathbf{s}(\omega) = \sum_{i=1}^{N} (\mathbf{h}_i, \mathbf{s}(\omega)) \mathbf{h}_i$$

for each $\omega \in \Omega$, and by Parseval's relation,

$$||\mathbf{s}(\omega)||^2 = \sum_{i=1}^{N} (\mathbf{h}_i, \mathbf{s}(\omega))^2 < E$$

for each $\omega \in \Omega$.

It is shown in Appendix 1d that if $\{s_i\}_{i=1}^N$ is any collection of scalar random variables with $\sum_{i=1}^N \mathcal{E} s_i^2 < \infty$, where $N = \dim \mathcal{H} \leq \infty$, then there is a second-order $\mathcal{H}$-valued random variable $\tilde{\mathbf{s}}$ such that $(\mathbf{h}_i, \tilde{\mathbf{s}}(\omega)) = s_i(\omega)$ for $i = 1, \ldots, N$ and for all $\omega \in \Omega$ with $\sum_{i=1}^N s_i^2(\omega) < \infty$. Therefore, if $\{s_i\}_{i=1}^N$ is any collection of scalar random variables with $\sum_{i=1}^N s_i^2(\omega) < E$ for all $\omega \in \Omega$, then there is a second-order $\mathcal{H}$-valued random variable $\tilde{\mathbf{s}}$ such that $(\mathbf{h}_i, \tilde{\mathbf{s}}(\omega)) = s_i(\omega)$ for $i = 1, \ldots, N$ and for *all* $\omega \in \Omega$, in which case $\tilde{\mathbf{s}}(\omega) = \sum_{i=1}^N s_i(\omega) \mathbf{h}_i$ for all $\omega \in \Omega$, and so by Parseval's relation, this $\tilde{\mathbf{s}}$ is an $\mathcal{H}_E$-valued random variable.

Thus, a map $\mathbf{s} : \Omega \to \mathcal{H}$ is an $\mathcal{H}_E$-valued random variable if, and only if,

$$\mathbf{s}(\omega) = \sum_{i=1}^{N} s_i(\omega) \mathbf{h}_i$$

for all $\omega \in \Omega$, where $\{s_i\}_{i=1}^N$ is a collection of scalar random variables with

$$\sum_{i=1}^{N} s_i^2(\omega) < E$$

for all $\omega \in \Omega$, in which case

$$s_i(\omega) = (\mathbf{h}_i, \mathbf{s}(\omega))$$

for $i = 1, \ldots, N$ and for all $\omega \in \Omega$. In particular, $|s_i(\omega)| < E^{1/2}$ for $i = 1, \ldots, N$ and for all $\omega \in \Omega$, which is a strong restriction on the scalar random variables $s_i = (\mathbf{h}_i, \mathbf{s})$. It implies immediately that the probability distribution functions

$$F_{(\mathbf{h}_i, \mathbf{s})}(x) = P(\{\omega \in \Omega : (\mathbf{h}_i, \mathbf{s}(\omega)) \leq x\})$$

must satisfy

$$F_{(\mathbf{h}_i, \mathbf{s})}(x) = \begin{cases} 0 \text{ if } x \leq -E^{1/2} \\ 1 \text{ if } x \geq E^{1/2} \end{cases}$$

for $i = 1, \ldots, N$. Thus $(\mathbf{h}_i, \mathbf{s})$ cannot be Gaussian-distributed, for instance, for any $i = 1, \ldots, N$. Also, since $\|\mathbf{s}(\omega)\| < E^{1/2}$ for all $\omega \in \Omega$, the probability distribution function

$$F_{||\mathbf{s}||}(x) = P(\{\omega \in \Omega : ||\mathbf{s}(\omega)|| \leq x\})$$

of the scalar random variable $\|\mathbf{s}\|$ must satisfy

$$F_{||\mathbf{s}||}(x) = \begin{cases} 0 \text{ if } x \leq 0 \\ 1 \text{ if } x \geq E^{1/2} \end{cases} \quad .$$

The characterization of $\mathcal{H}_E$-valued random variables given above will be used in Sect. 5 to construct an $\mathcal{H}_E$-valued random variable $\mathbf{s}_{t_0}$ for the shallow-water equations. This will guarantee directly that the random initial geopotential field is positive.

## 4 The Principle of Energetic Consistency for Differential Equations

### *4.1 Ordinary Differential Equations*

Consider a non-linear system of ordinary differential equations

$$\frac{\mathrm{d}\mathbf{s}}{\mathrm{d}t} + \mathbf{f}(\mathbf{s}, t) = \mathbf{0}\,, \tag{45}$$

where $\mathbf{f} : \mathcal{S}_1 \times \mathcal{T}_1 \to \mathbb{R}^N$, with $\mathcal{S}_1$ an open connected set in $\mathbb{R}^N$, possibly all of $\mathbb{R}^N$, and with $\mathcal{T}_1 = [t_0, T_1]$ and $0 < T_1 - t_0 < \infty$. Take $\mathcal{H} = \mathbb{R}^N$, with $(\cdot, \cdot)$ denoting the Euclidean inner product, $(\mathbf{g}, \mathbf{h}) = \mathbf{g}^T\mathbf{h}$ for all $\mathbf{g}, \mathbf{h} \in \mathbb{R}^N$, and $\|\cdot\|$ the Euclidean norm, $\|\mathbf{h}\| = (\mathbf{h}^T\mathbf{h})^{1/2}$ for all $\mathbf{h} \in \mathbb{R}^N$.

Assume that $\mathbf{f}$ is of class $C(\mathcal{S}_1 \times \mathcal{T}_1)$, i.e., that $\mathbf{f}$ is continuous on its domain of definition $\mathcal{S}_1 \times \mathcal{T}_1$. Assume also that $\mathbf{f}$ is bounded on $\mathcal{S}_1 \times \mathcal{T}_1$:

$$||\mathbf{f}(\mathbf{s}, t)|| < C_1$$

for some constant $C_1$, for all $\mathbf{s} \in \mathcal{S}_1$ and $t \in \mathcal{T}_1$. Note that the latter assumption follows from the former one if $\mathcal{S}_1 \subset \mathcal{H}_E$ for some $E < \infty$, where $\mathcal{H}_E$ was defined in Eq. (44). Assume finally that $\mathbf{f}$ is Lipschitz continuous in its first argument, uniformly in time, i.e., that there is a constant $C_2$ such that

$$||\mathbf{f}(\mathbf{r}, t) - \mathbf{f}(\mathbf{s}, t)|| \leq C_2||\mathbf{r} - \mathbf{s}||$$

for all $\mathbf{r}, \mathbf{s} \in \mathcal{S}_1$ and $t \in \mathcal{T}_1$.

A real $N$-vector function $\mathbf{s} = \mathbf{s}(t)$ defined on an interval $\mathcal{T}_* = [t_0, T_*]$, $T_* \in (t_0, T_1]$, is called a (continuous) solution of Eq. (45) if, for all $t \in \mathcal{T}_*$, (i) $\mathbf{s}(t) \in \mathcal{S}_1$, (ii) $\mathbf{s}(t)$ is continuous, and (iii) $\mathbf{s}(t)$ satisfies Eq. (45) pointwise. It follows from the continuity assumption on $\mathbf{f}$ that if $\mathbf{s}$ is a solution on an interval $\mathcal{T}_*$, then $\mathrm{d}\mathbf{s}/\mathrm{d}t$ is continuous on $\mathcal{T}_*$, and so

$$\frac{\mathrm{d}}{\mathrm{d}t}\|\mathbf{s}\|^2 = \frac{\mathrm{d}}{\mathrm{d}t}(\mathbf{s}, \mathbf{s}) = 2\left(\mathbf{s}, \frac{\mathrm{d}\mathbf{s}}{\mathrm{d}t}\right) = -2(\mathbf{s}, \mathbf{f}(\mathbf{s}, t)) \tag{46}$$

is also continuous on $\mathcal{T}_*$, hence integrable on $\mathcal{T}_*$. Similarly, if $\mathbf{r}$ and $\mathbf{s}$ are two solutions on an interval $\mathcal{T}_*$, then by the Schwarz inequality,

$$||\mathbf{r} - \mathbf{s}|| \left|\frac{\mathrm{d}||\mathbf{r} - \mathbf{s}||}{\mathrm{d}t}\right| = |(\mathbf{r} - \mathbf{s}, \mathbf{f}(\mathbf{r}, t) - \mathbf{f}(\mathbf{s}, t))| \leq ||\mathbf{r} - \mathbf{s}||\,||\mathbf{f}(\mathbf{r}, t) - \mathbf{f}(\mathbf{s}, t)||$$

for all $t \in \mathcal{T}_*$, and so by integrating it follows from the Lipschitz continuity assumption that

$$||\mathbf{r}(t) - \mathbf{s}(t)|| \leq \mathrm{e}^{C_2(t-t_0)}||\mathbf{r}(t_0) - \mathbf{s}(t_0)||$$

for all $t \in \mathcal{T}_*$. Thus, if $\mathbf{r}(t_0) = \mathbf{s}(t_0)$ then $\mathbf{r}(t) = \mathbf{s}(t)$ for all $t \in \mathcal{T}_*$: for each $\mathbf{s}_{t_0} \in \mathcal{S}_1$ there exists at most one solution $\mathbf{s}(t)$ defined on an interval $\mathcal{T}_*$, such that $\mathbf{s}(t_0) = \mathbf{s}_{t_0}$. The inequality also shows that if such a solution exists, then it depends continuously on $\mathbf{s}_{t_0}$, for all $t \in \mathcal{T}_*$.

The continuity and boundedness assumptions on $\mathbf{f}$ together imply that, for each $\mathbf{s}_{t_0} \in \mathcal{S}_1$, there does exist a solution $\mathbf{s}(t)$ with $\mathbf{s}(t_0) = \mathbf{s}_{t_0}$, and that it remains in existence either until time $t = T_1$ or the first time that the solution hits the boundary $\partial\mathcal{S}_1$ of $\mathcal{S}_1$, where $\mathbf{f}$ may not be defined, whichever is smaller (e.g. Coddington and Levinson 1955, pp. 6, 15). Thus, if $\mathcal{S}_1 = \mathbb{R}^N$, then the solution exists until time $T_1$. This time can be arbitrarily large, for instance if $\mathbf{f}$ is independent of time. More generally, a minimum existence time can be found by noting that the solution $\mathbf{s}(t)$ with $\mathbf{s}(t_0) = \mathbf{s}_{t_0} \in \mathcal{S}_1$ must satisfy the integral equation

$$\mathbf{s}(t) = \mathbf{s}_{t_0} - \int_{t_0}^{t} \mathbf{f}(\mathbf{s}(\tau), \tau)\, \mathrm{d}\tau\,,$$

and so

$$||\mathbf{s}(t) - \mathbf{s}_{t_0}|| < C_1(t - t_0)$$

by the boundedness assumption on $\mathbf{f}$, for as long as the solution exists. Denoting by $\rho(\mathbf{s}_{t_0})$ the Euclidean distance from any $\mathbf{s}_{t_0} \in \mathcal{S}_1$ to $\partial\mathcal{S}_1$,

$$\rho(\mathbf{s}_{t_0}) = \inf_{\mathbf{h}\in\partial\mathcal{S}_1} ||\mathbf{h} - \mathbf{s}_{t_0}||\,,$$

it follows that $\|\mathbf{s}(t) - \mathbf{s}_{t_0}\| < \rho(\mathbf{s}_{t_0})$ if $t - t_0 \leq \rho(\mathbf{s}_{t_0})/C_1$, and so the solution exists on $\mathcal{T}_* = [t_0, T_*]$ for

$$T_* = T_*(\mathbf{s}_{t_0}) = \min(T_1, t_0 + \rho(\mathbf{s}_{t_0})/C_1)\,.$$

Note that $\rho(\mathbf{s}_{t_0}) > 0$ for each $\mathbf{s}_{t_0} \in \mathcal{S}_1$ since $\mathcal{S}_1$ is an open set, and therefore $T_* > t_0$.

The principle of energetic consistency requires a set $\mathcal{S} \in \mathcal{B}(\mathcal{H}) = \mathcal{B}(\mathbb{R}^N)$ for the initial data and a time interval $\mathcal{T} = [t_0, T]$ such that, for *every* $\mathbf{s}_{t_0} \in \mathcal{S}$, the corresponding solution exists on $\mathcal{T}$, i.e., every solution must exist for the same minimum amount of time $T - t_0 > 0$, independently of the location of $\mathbf{s}_{t_0} \in \mathcal{S}$. If $\mathcal{S}_1 = \mathbb{R}^N$, then take $\mathcal{S} = \mathbb{R}^N$ and $\mathcal{T} = [t_0, T_1]$. Otherwise, let $\mathcal{S}$ be any open set in $\mathbb{R}^N$ which is contained in the interior of $\mathcal{S}_1$, and denote by $\rho_{\mathcal{S}}$ the minimum Euclidean distance from the boundary of $\mathcal{S}$ to that of $\mathcal{S}_1$. Then

$$\rho(\mathbf{s}_{t_0}) > \rho_{\mathcal{S}} = \inf_{\mathbf{s}\in\mathcal{S}} \rho(\mathbf{s}) > 0$$

for all $\mathbf{s}_{t_0} \in \mathcal{S}$, and setting

$$T = T_{\mathcal{S}} = \min(T_1, t_0 + \rho_{\mathcal{S}}/C_1)$$

and $\mathcal{T} = \mathcal{T}(\mathcal{S}) = [t_0, T]$, it follows that the unique solution $\mathbf{s}(t)$ corresponding to each $\mathbf{s}_{t_0} \in \mathcal{S}$ exists for all $t \in \mathcal{T}$. Denoting this solution by $\mathbf{s}_t = \mathbf{N}_{t,t_0}(\mathbf{s}_{t_0})$, it follows that $\mathbf{N}_{t,t_0}$ is defined uniquely on $\mathcal{S}$, as a continuous map from $\mathcal{S}$ into $\mathcal{H} = \mathbb{R}^N$, for all $t \in \mathcal{T}$.

It follows from Eq. (46) that the solution operator $\mathbf{N}_{t,t_0}$ is conservative if

$$(\mathbf{s}, \mathbf{f}(\mathbf{s}, t)) = 0$$

for all $\mathbf{s} \in \mathcal{S}_1$ and $t \in \mathcal{T}$. More generally, it follows that if there is a constant $C_3$ such that

$$|(\mathbf{s}, \mathbf{f}(\mathbf{s}, t))| \leq C_3||\mathbf{s}||^2$$

for all $\mathbf{s} \in \mathcal{S}_1$ and $t \in \mathcal{T}$, then $\mathbf{N}_{t,t_0}$ is bounded, with

$$||\mathbf{s}_t|| = ||\mathbf{N}_{t,t_0}(\mathbf{s}_{t_0})|| \leq \mathrm{e}^{C_3(t-t_0)}||\mathbf{s}_{t_0}||$$

for all $t \in \mathcal{T}$. Note that if $\mathbf{f}(\mathbf{s}, t)$ depends linearly on $\mathbf{s}$ near the origin of coordinates $\mathbf{0} \in \mathbb{R}^N$, or if $\mathbf{0} \notin \mathcal{S}_1$, then by the boundedness assumption on $\mathbf{f}$ there is a constant $C_3$ such that $\|\mathbf{f}(\mathbf{s}, t))\| \leq C_3\|\mathbf{s}\|$ for all $\mathbf{s} \in \mathcal{S}_1$ and $t \in \mathcal{T}$, so that by the Schwarz inequality,

$$|(\mathbf{s}, \mathbf{f}(\mathbf{s}, t))| \leq ||\mathbf{s}||\,||\mathbf{f}(\mathbf{s}, t))|| \leq C_3||\mathbf{s}||^2$$

for all $\mathbf{s} \in \mathcal{S}_1$ and $t \in \mathcal{T}$, and therefore $\mathbf{N}_{t,t_0}$ is bounded for all $t \in \mathcal{T}$.

## *4.2 Symmetric Hyperbolic Partial Differential Equations*

### 4.2.1 The Deterministic Initial-Value Problem

Consider now a non-linear system of partial differential equations

$$\frac{\partial \mathbf{s}}{\partial t} + \mathbf{G}\mathbf{s} = \mathbf{0}\,, \tag{47}$$

where $\mathbf{G} = \mathbf{G}(\mathbf{s}) = \mathbf{G}(\mathbf{s}, \mathbf{x}, t)$ is a linear differential operator of first order in space variables $\mathbf{x} = (x_1, \ldots, x_d)^T$,

$$\mathbf{G} = \sum_{j=1}^{d} \mathbf{A}_j(\mathbf{s}, \mathbf{x}, t)\frac{\partial}{\partial x_j} + \mathbf{A}_{d+1}(\mathbf{s}, \mathbf{x}, t)\,,$$

and $\mathbf{A}_1, \ldots, \mathbf{A}_{d+1}$ are real $n \times n$ matrices. For simplicity assume that the $d$-dimensional spatial domain $D$ of the problem is

$$D = \{\mathbf{x} \in \mathbb{R}^d : |x_j| \leq L_j, j = 1, \ldots, d\}\,,$$

with periodic boundary conditions at the endpoints $x_j = \pm L_j, j = 1, \ldots, d$. Consider endpoint $x_j = L_j$ to be identified with endpoint $x_j = -L_j$, for each $j = 1, \ldots, d$, so that a continuous function on $D$ satisfies the periodic boundary conditions automatically. (Spherical geometry will be treated in Sect. 5.) Take $\mathcal{H} = L^2(D)$, the Hilbert space of real, Lebesgue square-integrable $n$-vectors on $D$, with inner product

$$(\mathbf{g}, \mathbf{h}) = \int_D \mathbf{g}^T(\mathbf{x})\mathbf{h}(\mathbf{x})\,\mathrm{d}x_1 \cdots \mathrm{d}x_d$$

for all $\mathbf{g}, \mathbf{h} \in L^2(D)$, and corresponding norm $\|\mathbf{h}\| = (\mathbf{h}, \mathbf{h})^{1/2}$ for all $\mathbf{h} \in L^2(D)$.

Assume that the matrices $\mathbf{A}_1, \ldots, \mathbf{A}_{d+1}$ are defined on all of $\mathbb{R}^n \times D \times \mathcal{T}_1$, where $\mathcal{T}_1 = [t_0, T_1]$ and $0 < T_1 - t_0 < \infty$. Assume further that each matrix is of class $C^\infty(\mathbb{R}^n \times D \times \mathcal{T}_1)$, i.e., that all of the matrix elements and all of their partial derivatives are continuous functions on $\mathbb{R}^n \times D \times \mathcal{T}_1$ and satisfy the periodic boundary conditions in the space variables. Assume finally that $\mathbf{A}_1, \ldots, \mathbf{A}_d$ (but not $\mathbf{A}_{d+1}$) are symmetric matrices.

A real $n$-vector function $\mathbf{s} = \mathbf{s}(\mathbf{x}, t)$ defined on $D \times \mathcal{T}_*$, with $\mathcal{T}_* = [t_0, T_*]$ and $T_* \in (t_0, T_1]$, is called a classical solution of the symmetric hyperbolic system (Eq. 47) if (i) $\mathbf{s} \in C^1(D) \cap C^1(\mathcal{T}_*)$ and (ii) $\mathbf{s}$ satisfies Eq. (47) pointwise in $D \times \mathcal{T}_*$. The condition $\mathbf{s} \in C^1(D) \cap C^1(\mathcal{T}_*)$ means that the components of the vector $\mathbf{s}$ and their first time and space derivatives are continuous on $D$ for each fixed $t \in \mathcal{T}_*$, are continuous on $\mathcal{T}_*$ for each fixed $\mathbf{x} \in D$, and satisfy the periodic boundary conditions. The initial condition for a classical solution is a real $n$-vector function $\mathbf{s}_{t_0} \in C^1(D)$.

Suppose for the moment that $\mathbf{s} = \mathbf{s}(\mathbf{x}, t)$ is a classical solution on $D \times \mathcal{T}_*$. Then

$$\frac{\mathrm{d}}{\mathrm{d}t}\|\mathbf{s}\|^2 = \frac{\mathrm{d}}{\mathrm{d}t}(\mathbf{s}, \mathbf{s}) = 2\left(\mathbf{s}, \frac{\partial \mathbf{s}}{\partial t}\right) = -2(\mathbf{s}, \mathbf{G}(\mathbf{s})\mathbf{s}) \tag{48}$$

is continuous on $\mathcal{T}_*$. Also, by using the symmetry of $\mathbf{A}_1, \ldots, \mathbf{A}_d$ and the periodic boundary conditions, an integration by parts gives

$$(\mathbf{s}, \mathbf{G}(\mathbf{s})\mathbf{s}) = \int_D \mathbf{s}^T(\mathbf{x}, t)\mathbf{B}(\mathbf{x}, t)\mathbf{s}(\mathbf{x}, t)\,\mathrm{d}x_1 \cdots \mathrm{d}x_d \tag{49}$$

for all $t \in \mathcal{T}_*$, where

$$\mathbf{B}(\mathbf{x}, t) = \mathbf{A}_{d+1}(\mathbf{s}, \mathbf{x}, t) - \tfrac{1}{2}\sum_{j=1}^{d} \frac{\mathrm{d}\mathbf{A}_j(\mathbf{s}, \mathbf{x}, t)}{\mathrm{d}x_j}$$

and

$$\frac{\mathrm{d}\mathbf{A}_j}{\mathrm{d}x_j} = \sum_{i=1}^{n} \frac{\partial \mathbf{A}_j}{\partial s_i}\frac{\partial s_i}{\partial x_j} + \frac{\partial \mathbf{A}_j}{\partial x_j}\,.$$

Further, since $\mathbf{s} \in C^1(D) \cap C^1(\mathcal{T}_*)$, the components of $\mathbf{s}$ and their first partial derivatives with respect to the space variables are bounded functions on $D \times \mathcal{T}_*$. Define $\beta_0 = \beta_0(\mathbf{s})$ by

$$\beta_0 = \max_{D \times \mathcal{T}_*} \sum_{i=1}^{n} |s_i(\mathbf{x}, t)|\,,$$

and $\beta_j = \beta_j(\mathbf{s})$ by

$$\beta_j = \max_{D\times\mathcal{T}_*} \sum_{i=1}^{n} \left| \frac{\partial s_i(\mathbf{x},t)}{\partial x_j} \right| ,$$

for $j = 1,\ldots,d$. Then it follows from Eq. (49) and the continuity assumption on the matrices $\mathbf{A}_1,\ldots,\mathbf{A}_{d+1}$ that there is a continuous function $C_1 = C_1(\beta_0,\ldots,\beta_d)$ such that

$$|(\mathbf{s}, \mathbf{G}(\mathbf{s})\mathbf{s})| \le C_1 ||\mathbf{s}||^2$$

for all $t \in \mathcal{T}_*$. Equation (48) then implies that

$$\|\mathbf{s}(\cdot,t)\| \le \mathrm{e}^{C_1(t-t_0)} \|\mathbf{s}(\cdot,t_0)\| \tag{50}$$

for all $t \in \mathcal{T}_*$.

A similar argument shows that if $\mathbf{r}$ and $\mathbf{s}$ are two classical solutions on $D \times \mathcal{T}_*$, then there is a continuous function $C_2 = C_2(\beta_0(\mathbf{r}),\ldots,\beta_d(\mathbf{r}),\beta_0(\mathbf{s}),\ldots,\beta_d(\mathbf{s}))$ such that

$$\|\mathbf{r}(\cdot,t) - \mathbf{s}(\cdot,t)\| \le \mathrm{e}^{C_2(t-t_0)} \|\mathbf{r}(\cdot,t_0) - \mathbf{s}(\cdot,t_0)\| \tag{51}$$

for all $t \in \mathcal{T}_*$. Therefore, for each $\mathbf{s}_{t_0} \in C^1(D)$, there exists at most one classical solution $\mathbf{s}$ on $D \times \mathcal{T}_*$ such that $\mathbf{s}(\mathbf{x},t_0) = \mathbf{s}_{t_0}(\mathbf{x})$ for all $\mathbf{x} \in D$. This inequality does not imply that if such a solution exists, then it depends continuously on $\mathbf{s}_{t_0}$ in the norm $\|\cdot\|$, unless $C_2$ can be made to depend only on $\mathbf{r}_{t_0}$ and $\mathbf{s}_{t_0}$. This is accomplished by means of the existence theory itself, discussed next.

Denote by $H^k = H^k(D)$, for $k = 0, 1, \ldots$, the Sobolev space of real $n$-vectors on $D$ with $k$ Lebesgue square-integrable derivatives on $D$. The spaces $H^k$ are Hilbert spaces, with inner product

$$(\mathbf{g},\mathbf{h})_{H^k} = \sum_{l=0}^{k} \sum_{l_1+\cdots+l_d=l} (D^l\mathbf{g}, D^l\mathbf{h})$$

for all $\mathbf{g},\mathbf{h} \in H^k$, where

$$D^l = \frac{\partial^l}{\partial x_1^{l_1} \cdots \partial x_d^{l_d}} ,$$

and corresponding norm $\|\mathbf{h}\|_{H^k} = (\mathbf{h},\mathbf{h})_{H^k}^{1/2}$ for all $\mathbf{h} \in H^k$. Note that $H^m \subset H^k \subset H^0 = \mathcal{H}$ for $0 \le k \le m$. The Sobolev lemma (e.g. Kreiss and Lorenz 1989, Appendix 3, pp. 371–387) says that if $\mathbf{h} \in H^k$ and $k \ge \left[\frac{d}{2}\right] + 1$, where $[y]$ denotes the largest integer less than or equal to $y$, then $\mathbf{h}$ is a bounded function on $D$, with bound

$$\max_{\mathbf{x}\in D} \sum_{i=1}^{n} |h_i(\mathbf{x})| \leq \alpha_k \|\mathbf{h}\|_{H^k} , \tag{52}$$

where the constant $\alpha_k$ depends on $L_1, \ldots, L_d$ but not on $\mathbf{h}$. It follows that if $\mathbf{h} \in H^k$ and $k \geq \left[\frac{d}{2}\right] + l + 1$ for some positive integer $l$, then all of the $l$th-order partial derivatives of $\mathbf{h}$ are bounded functions on $D$, with bound

$$\max_{\mathbf{x}\in D} \sum_{i=1}^{n} |D^l h_i(\mathbf{x})| \leq \alpha_k \|\mathbf{h}\|_{H^k} , \tag{53}$$

and in particular, $\mathbf{h} \in C^{l-1}(D)$, since otherwise the $l$th-order partial derivatives of $\mathbf{h}$ are not defined as bounded functions. Thus, for any non-negative integer $l$, $H^k = H^k(D) \subset C^l(D)$ if $k \geq \left[\frac{d}{2}\right] + l + 2$.

Suppose now that $\mathbf{s}_{t_0} \in H^k$ with $k \geq \left[\frac{d}{2}\right] + 3$. According to the existence theory for linear and quasi-linear symmetric hyperbolic systems (e.g. Courant and Hilbert 1962, pp. 668–676), there is a time interval $\mathcal{T}_* \subset \mathcal{T}_1$ for which Eq. (47) has a solution $\mathbf{s} \in H^k \cap C^1(\mathcal{T}_*)$ with $\mathbf{s}(\mathbf{x}, t_0) = \mathbf{s}_{t_0}(\mathbf{x})$ for all $\mathbf{x} \in D$, which is the classical solution since $H^k \subset C^1(D)$, and the solution remains in existence as long as $t \leq T_1$ and $\mathbf{s}(\cdot, t) \in H^k$. This is completely analogous to the situation for ordinary differential equations: the first time $t$ such that the solution $\mathbf{s}(\cdot, t) \notin H^k$, if such a time is reached, is the first time the solution hits the "boundary" of $H^k$, $\|\mathbf{s}(\cdot, t)\|_{H^k} = \infty$. Typically the first partial derivatives of the classical solution become unbounded in finite time, even if $\mathbf{s}_{t_0} \in C^\infty(D)$ (e.g. Lax 1973, Theorem 6.1, p. 37).

A minimum existence time for the solution $\mathbf{s} \in H^k \cap C^1(\mathcal{T}_*)$, $k \geq \left[\frac{d}{2}\right] + 3$, can be found in the following way. For any $\mathbf{s} \in H^k \cap C^1(\mathcal{T}_*)$,

$$\frac{\mathrm{d}}{\mathrm{d}t} ||\mathbf{s}||^2_{H^k} = -2(\mathbf{s}, \mathbf{G}(\mathbf{s})\mathbf{s})_{H^k}$$

is continuous on $\mathcal{T}_*$, as in Eq. (48). An integration by parts using the symmetry of the matrices $\mathbf{A}_1, \ldots, \mathbf{A}_d$, along with the Sobolev inequalities of Eqs. (52) and (53), shows that there is a function $\phi \in C^1([0, \infty))$ such that

$$|(\mathbf{s}, \mathbf{G}(\mathbf{s})\mathbf{s})_{H^k}| \leq \phi(||\mathbf{s}||_{H^k})||\mathbf{s}||_{H^k} ;$$

see Kreiss and Lorenz (1989, pp. 190–196) for details. It follows that the solution $\mathbf{s}(\cdot, t)$ exists in $H^k$ as long as $t \leq T_1$ and the solution $y(t)$ of the ordinary differential equation $\mathrm{d}y/\mathrm{d}t = \phi(y)$ with $y(t_0) = \|\mathbf{s}_{t_0}\|_{H^k}$ remains finite. Further, there is a time $T_2 > t_0$ depending continuously on $\|\mathbf{s}_{t_0}\|_{H^k}$, $T_2 = T_2(\|\mathbf{s}_{t_0}\|_{H^k}) \leq T_1$, for which $\|\mathbf{s}(\cdot, t)\|_{H^k}$ can be bounded in terms of $\|\mathbf{s}_{t_0}\|_{H^k}$, say

$$||\mathbf{s}(\cdot,t)||_{H^k} \leq \sqrt{2}\,||\mathbf{s}_{t_0}||_{H^k}\,,$$

for all $t \in [t_0, T_2]$ (e.g. Kreiss and Lorenz 1989, Lemma 6.4.4, p. 196). Then by the continuity of $T_2(\|\mathbf{s}_{t_0}\|_{H^k})$, it follows that if $\mathbf{s}_{t_0}$ is restricted to be in any bounded set in $H^k$, say if $\mathbf{s}_{t_0} \in H^k_E$ for some $E < \infty$, where

$$H^k_E = \{\mathbf{h} \in H^k : ||\mathbf{h}||^2_{H^k} < E\}\,,$$

then $T_2$ becomes independent of $\mathbf{s}_{t_0}$ (but depends on $E$), and the solution $\mathbf{s}$ corresponding to any $\mathbf{s}_{t_0} \in H^k_E$ satisfies

$$||\mathbf{s}(\cdot,t)||^2_{H^k} < 2E$$

for all $t \in [t_0, T_2]$. Also, since $H^k_E$ is open as a set in $H^k$, and since $\|\mathbf{h}\| \leq \|\mathbf{h}\|_{H^k}$ for all $\mathbf{h} \in H^k$, $H^k_E$ is open as a set in $L^2(D)$, and therefore $H^k_E \in \mathcal{B}(L^2(D))$.

### 4.2.2 The Solution Operator

Thus take $\mathcal{S} = H^k_E$ for any $E < \infty$ and $k \geq \left[\frac{d}{2}\right] + 3$, and take $\mathcal{T} = \mathcal{T}(\mathcal{S}) = [t_0, T_2]$. Then $\mathcal{S} \in \mathcal{B}(\mathcal{H}) = \mathcal{B}(L^2(D))$, and the unique classical solution $\mathbf{s}(\cdot,t)$ corresponding to each $\mathbf{s}_{t_0} \in \mathcal{S}$ exists in $H^k_{2E} \subset H^k \subset \mathcal{H} = L^2(D)$ for all $t \in \mathcal{T}$. Denoting this solution by $\mathbf{s}_t = \mathbf{N}_{t,t_0}(\mathbf{s}_{t_0})$, it follows that $\mathbf{N}_{t,t_0}$ is defined uniquely on $\mathcal{S}$, as a map from $\mathcal{S}$ into $\mathcal{H}$, for all $t \in \mathcal{T}$. Further, since $\mathbf{s}_t \in H^k_{2E}$ for all $t \in \mathcal{T}$, it follows from the Sobolev inequalities that the function $C_2$ in Eq. (51) depends only on $E$ and $\alpha_k$, and therefore the map $\mathbf{N}_{t,t_0}$ is continuous in the norm $\|\cdot\|$, for all $t \in \mathcal{T}$. Note that $\mathcal{S} = H^k_E \subset \mathcal{H}_E$, where $\mathcal{H}_E$ was defined in Eq. (44), since $\|\mathbf{h}\| \leq \|\mathbf{h}\|_{H^k}$ for all $\mathbf{h} \in H^k$. It was necessary to define $\mathcal{S}$ as a bounded set in $H^k$, and therefore as a bounded set in $L^2(D)$.

The solution operator $\mathbf{N}_{t,t_0}$ is bounded not only as a map from $\mathcal{S}$ into $\mathcal{H}$, with the function $C_1$ in Eq. (50) now being a constant depending only on $E$ and $\alpha_k$, but also as a map from $\mathcal{S}$ into $H^k$, with

$$||\mathbf{s}_t||_{H^k} = ||\mathbf{N}_{t,t_0}(\mathbf{s}_{t_0})||_{H^k} \leq \sqrt{2}\,||\mathbf{s}_{t_0}||_{H^k}$$

for all $t \in \mathcal{T}$. According to Eq. (48), the solution operator is conservative if the differential operator $\mathbf{G}$ is skew-symmetric,

$$(\mathbf{s}, \mathbf{G}(\mathbf{r})\mathbf{s}) = 0$$

for all $\mathbf{r}(\cdot,t), \mathbf{s}(\cdot,t) \in H^k$ and $t \in \mathcal{T}$. This conservation condition is met for an important class of symmetric hyperbolic systems (Lax 1973, p. 31), but often a change of dependent variables which destroys symmetry of the matrices $\mathbf{A}_1, \ldots, \mathbf{A}_d$ is necessary to obtain conservation in $\mathcal{H} = L^2(D)$, as will be the case for the shallow-water equations.

It has been shown that, for each $\mathbf{s}_{t_0} \in \mathcal{S} = H_E^k$, with $E < \infty$, $k \geq \left[\frac{d}{2}\right] + l + 2$ and $l \geq 1$, the unique corresponding solution $\mathbf{s} = \mathbf{s}(\mathbf{x}, t)$ is of class $C^l(D) \cap C^1(\mathcal{T})$, for $\mathcal{T} = [t_0, T]$ and an appropriately defined $T$ depending on $\alpha_k$ and $E$, and that $\|\mathbf{s}(\cdot, t)\|_{H^k}^2 < 2E$ for all $t \in \mathcal{T}$. It is important to have a condition to guarantee further that $\mathbf{s} \in C^l(D \times \mathcal{T})$, particularly for the shallow-water example. To this end, denote by $L^2(D \times \mathcal{T})$ the Hilbert space of real, Lebesgue square-integrable $n$-vectors on $D \times \mathcal{T}$, with inner product

$$(\mathbf{g}, \mathbf{h})_{\mathcal{T}} = \int_{t_0}^{T} (\mathbf{g}, \mathbf{h})\, \mathrm{d}t$$

for all $\mathbf{g}, \mathbf{h} \in L^2(D \times \mathcal{T})$, and corresponding norm $\|\mathbf{h}\|_{\mathcal{T}} = (\mathbf{h}, \mathbf{h})_{\mathcal{T}}^{1/2}$ for all $\mathbf{h} \in L^2(D \times \mathcal{T})$. Also, denote by $H^m(D \times \mathcal{T})$, for $m = 0, 1, \ldots$, the Sobolev space of real $n$-vectors on $D \times \mathcal{T}$ with $m$ Lebesgue square-integrable mixed space and time partial derivatives on $D \times \mathcal{T}$, with the Sobolev inner product and norm. Thus, for any non-negative integer $l$, $H^m(D \times \mathcal{T}) \subset C^l(D \times \mathcal{T})$ if $m \geq \left[\frac{d+1}{2}\right] + l + 2$. Now, the differential equations (Eq. 47) can be used to express all mixed space-time partial derivatives of the solution up to any order $m$ in terms of pure spatial partial derivatives up to order $m$. But

$$\int_{t_0}^{T} ||\mathbf{s}(\cdot, t)||_{H^k}^2\, \mathrm{d}t < 2E(T - t_0) < \infty$$

since $\|\mathbf{s}(\cdot, t)\|_{H^k}^2 < 2E$ for all $t \in \mathcal{T}$, and therefore $\mathbf{s} \in H^k(D \times \mathcal{T})$. Thus, for each $\mathbf{s}_{t_0} \in \mathcal{S} = H_E^k$, with $E < \infty$, $k \geq \left[\frac{d+1}{2}\right] + l + 2$ and $l \geq 1$, the unique corresponding solution $\mathbf{s} = \mathbf{s}(\mathbf{x}, t)$ is of class $C^l(D \times \mathcal{T})$.

### 4.2.3 The Stochastic Initial-Value Problem

With $\mathcal{H} = L^2(D)$, $\mathcal{S} = H_E^k$, $E < \infty$, $k \geq \left[\frac{d}{2}\right] + l + 2$ and $l \geq 1$, let $\mathcal{T} = [t_0, T]$ and $\mathbf{N}_{t,t_0}$ be as defined in Sect. 4.2.2, let $t \in \mathcal{T}$, and suppose that $\mathbf{s}_{t_0}$ is an $\mathcal{S}$-valued random variable. Since $\mathcal{S} \subset \mathcal{H}_E$, it follows from the discussion of Sect. 3.4 that $\mathbf{s}_{t_0}$ is a second-order $\mathcal{S}$-valued random variable. Therefore by Theorem 1, $\mathbf{s}_t = \mathbf{N}_{t,t_0}(\mathbf{s}_{t_0})$ is a second-order $\mathcal{H}$-valued random variable, with mean $\bar{\mathbf{s}}_t \in \mathcal{H}$ and covariance operator $\mathcal{P}_t : \mathcal{H} \to \mathcal{H}$, which are related by

$$||\bar{\mathbf{s}}_t||^2 + \mathrm{tr}\, \mathcal{P}_t \leq \mathrm{e}^{2C_1(t-t_0)}(||\bar{\mathbf{s}}_{t_0}||^2 + \mathrm{tr}\, \mathcal{P}_{t_0})\,,$$

where $C_1$ is the constant in Eq. (50). In fact,

$$||\mathbf{s}_t(\omega)||^2 \leq ||\mathbf{s}_t(\omega)||_{H^k}^2 \leq 2||\mathbf{s}_{t_0}(\omega)||_{H^k}^2 < 2E$$

for all $\omega \in \Omega$, and so $\mathbf{s}_t$ is a second-order $H^k$-valued random variable with

$$\mathcal{E}||\mathbf{s}_t||^2 = ||\bar{\mathbf{s}}_t||^2 + \operatorname{tr} \mathcal{P}_t \leq \mathcal{E}||\mathbf{s}_t||_{H^k}^2 \leq 2\mathcal{E}||\mathbf{s}_{t_0}||_{H^k}^2 < 2E\,.$$

The covariance operator $\mathcal{P}_t$ can be expressed in the following tangible way, which will lead also to a simple expression for $\operatorname{tr} \mathcal{P}_t$. Since $\mathcal{P}_t$ is a trace class operator, $\mathcal{P}_t$ is also a Hilbert-Schmidt operator. Since $\mathcal{H} = L^2(D)$ and $\mathcal{P}_t : \mathcal{H} \to \mathcal{H}$, it follows (e.g. Reed and Simon 1972, Theorem VI.23, p. 210) that there is a real $n \times n$ matrix function $\mathbf{P}_t \in L^2(D \times D)$, called the covariance matrix of $\mathbf{s}_t$, such that

$$(\mathcal{P}_t \mathbf{h})(\mathbf{x}) = \int_D \mathbf{P}_t(\mathbf{x}, \mathbf{y})\mathbf{h}(\mathbf{y})\, \mathrm{d}\mathbf{y}$$

for all $\mathbf{h} \in \mathcal{H}$, where $d\mathbf{y} = dy_1 \cdots dy_d$, and moreover, that

$$\int_D \int_D \operatorname{tr} \mathbf{P}_t(\mathbf{x}, \mathbf{y})\mathbf{P}_t^T(\mathbf{x}, \mathbf{y})\, \mathrm{d}\mathbf{x}\, \mathrm{d}\mathbf{y} = \sum_{i=1}^{\infty} \lambda_i^2(t) < \infty\,, \tag{54}$$

where $\operatorname{tr} \mathbf{A}$ denotes the trace of a matrix $\mathbf{A}$ and $\{\lambda_i(t)\}_{i=1}^{\infty}$ are the eigenvalues of the covariance operator $\mathcal{P}_t$. Thus

$$\mathcal{E}(\mathbf{g}, \mathbf{s}_t - \bar{\mathbf{s}}_t)(\mathbf{h}, \mathbf{s}_t - \bar{\mathbf{s}}_t) = (\mathbf{g}, \mathcal{P}_t \mathbf{h}) = \int_D \int_D \mathbf{g}^T(\mathbf{x})\mathbf{P}_t(\mathbf{x}, \mathbf{y})\mathbf{h}(\mathbf{y})\, \mathrm{d}\mathbf{x}\, \mathrm{d}\mathbf{y} \tag{55}$$

for all $\mathbf{g}, \mathbf{h} \in \mathcal{H}$. Since $\mathcal{P}_t$ is self-adjoint, the covariance matrix $\mathbf{P}_t$ has the symmetry property $\mathbf{P}_t^T(\mathbf{x}, \mathbf{y}) = \mathbf{P}_t(\mathbf{y}, \mathbf{x})$ for all $\mathbf{x}, \mathbf{y} \in D$.

Now let $\{\tilde{\mathbf{h}}_i(\cdot, t)\}_{i=1}^{\infty}$ denote the orthonormal eigenvectors (eigenfunctions) of $\mathcal{P}_t$ corresponding to the eigenvalues $\{\lambda_i(t)\}_{i=1}^{\infty}$,

$$\mathcal{P}_t \tilde{\mathbf{h}}_i(\cdot, t) = \lambda_i(t)\tilde{\mathbf{h}}_i(\cdot, t)$$

for $i = 1, 2, \ldots$. The eigenvalues are all non-negative since the covariance operator is positive semidefinite, and the eigenvectors form an orthonormal basis for $\mathcal{H}$ since the covariance operator is Hilbert-Schmidt. From Eq. (55) and the orthonormality of the eigenvectors, it follows that

$$\int_D \int_D \tilde{\mathbf{h}}_j^T(\mathbf{x}, t)\mathbf{P}_t(\mathbf{x}, \mathbf{y})\tilde{\mathbf{h}}_i(\mathbf{y}, t)\, \mathrm{d}\mathbf{x}\, \mathrm{d}\mathbf{y} = (\tilde{\mathbf{h}}_j(\cdot, t), \mathcal{P}_t \tilde{\mathbf{h}}_i(\cdot, t)) = \lambda_i(t)\delta_{ij}$$

for $i,j = 1, 2, \ldots$, where $\delta_{ij}$ is the Kronecker delta. Therefore, $\mathbf{P}_t$ has the representation

$$\mathbf{P}_t(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{\infty} \lambda_i(t)\tilde{\mathbf{h}}_i(\mathbf{x}, t)\tilde{\mathbf{h}}_i^T(\mathbf{y}, t)\,, \tag{56}$$

where the convergence is in $L^2(D \times D)$ as indicated in Eq. (54). Since the eigenvectors form an orthonormal basis for $\mathcal{H}$ and since $\mathcal{P}_t$ is trace class,

$$\operatorname{tr} \mathcal{P}_t = \sum_{i=1}^{\infty} \lambda_i(t) < \infty \,.$$

But according to Eq. (56),

$$\operatorname{tr} \mathbf{P}_t(\mathbf{x}, \mathbf{x}) = \sum_{i=1}^{\infty} \lambda_i(t) \tilde{\mathbf{h}}_i^T(\mathbf{x}, t) \tilde{\mathbf{h}}_i(\mathbf{x}, t) \,,$$

and therefore

$$\int_D \operatorname{tr} \mathbf{P}_t(\mathbf{x}, \mathbf{x}) \, d\mathbf{x} = \sum_{i=1}^{\infty} \lambda_i(t)$$

by the normality of the eigenvectors. Thus,

$$\operatorname{tr} \mathcal{P}_t = \int_D \operatorname{tr} \mathbf{P}_t(\mathbf{x}, \mathbf{x}) \, d\mathbf{x} \,. \tag{57}$$

Now recall that $\mathbf{s}_t$ is a second-order $H^k$-valued random variable. Therefore $\bar{\mathbf{s}}_t \in H^k$, $\mathcal{P}_t$ maps $H^k$ into $H^k$, and

$$\mathcal{E} ||\mathbf{s}_t||_{H^k}^2 = ||\bar{\mathbf{s}}_t||_{H^k}^2 + \operatorname{tr} \mathcal{P}_t \,.$$

Also, $\bar{\mathbf{s}}_t \in C^l(D)$ since $H^k \subset C^l(D)$. Further, since $\mathcal{P}_t$ maps $H^k$ into $H^k$, the eigenvectors of $\mathcal{P}_t$ form an orthonormal basis for $H^k$, and therefore they are all in $C^l(D)$.

Finally, let $\{\mathbf{h}_i\}_{i=1}^{\infty}$ be an orthonormal basis for $H^k$. Since $\mathbf{s}_{t_0}$ is an $H_E^k$-valued random variable, it follows from the discussion of Sect. 3.4 that

$$\mathbf{s}_{t_0}(\omega) = \sum_{i=1}^{\infty} (\mathbf{h}_i, \mathbf{s}_{t_0}(\omega))_{H^k} \mathbf{h}_i$$

for all $\omega \in \Omega$, with

$$||\mathbf{s}_{t_0}(\omega)||_{H^k}^2 = \sum_{i=1}^{\infty} (\mathbf{h}_i, \mathbf{s}_{t_0}(\omega))_{H^k}^2 < E \,.$$

for all $\omega \in \Omega$. It follows also that if $\{s_i\}_{i=1}^{\infty}$ is any collection of scalar random variables with

$$\sum_{i=1}^{\infty} s_i^2(\omega) < E$$

for all $\omega \in \Omega$, then $\sum_{i=1}^{\infty} s_i(\omega)\mathbf{h}_i$ is an $H_E^k$-valued random variable.

## 5 The Shallow-Water Equations

The global non-linear shallow-water equations written for the zonal and meridional velocity components and geopotential, $u$, $v$ and $\Phi$, respectively, are the system

$$\frac{\partial u}{\partial t} + \mathbf{V}\cdot\nabla u - \left(f + \frac{u}{a}\tan\phi\right)v + \frac{1}{a\cos\phi}\frac{\partial \Phi}{\partial \lambda} = 0$$
$$\frac{\partial v}{\partial t} + \mathbf{V}\cdot\nabla v + \left(f + \frac{u}{a}\tan\phi\right)u + \frac{1}{a}\frac{\partial \Phi}{\partial \phi} = 0$$
$$\frac{\partial \Phi}{\partial t} + \mathbf{V}\cdot\nabla\Phi + \Phi\nabla\cdot\mathbf{V} = 0\,,$$

where $\mathbf{V}$ is the wind vector, $\phi$ the latitude, $\lambda$ the longitude, $a$ the Earth radius, and $f$ the Coriolis parameter. The change of variable $\Phi = w^2/4$ yields the symmetric hyperbolic system

$$\frac{\partial \mathbf{s}}{\partial t} + \left[\mathbf{A}\frac{1}{a\cos\phi}\frac{\partial}{\partial\lambda} + \mathbf{B}\frac{1}{a}\frac{\partial}{\partial\phi} + \mathbf{C}\right]\mathbf{s} = 0\,, \tag{58}$$

where $\mathbf{s} = (u, v, w)^T$,

$$\mathbf{A} = \begin{bmatrix} u & 0 & \frac{1}{2}w \\ 0 & u & 0 \\ \frac{1}{2}w & 0 & u \end{bmatrix},$$
$$\mathbf{B} = \begin{bmatrix} v & 0 & 0 \\ 0 & v & \frac{1}{2}w \\ 0 & \frac{1}{2}w & v \end{bmatrix},$$
$$\mathbf{C} = \begin{bmatrix} 0 & -\left(f + \frac{u}{a}\tan\phi\right) & 0 \\ f + \frac{u}{a}\tan\phi & 0 & 0 \\ 0 & -\frac{1}{2}\frac{w}{a}\tan\phi & 0 \end{bmatrix}.$$

Now let $\mathcal{H} = L^2(S)$, the Hilbert space of real square-integrable 3-vectors on the sphere of radius $a$, with inner product $(\cdot,\cdot)$ and corresponding norm $\|\cdot\|$. Appendix 2b establishes a Sobolev-type lemma for the family of Hilbert spaces $\{\Phi_p = \Phi_p(S), p \geq 0\}$,

$$\Phi_p = \{\mathbf{h} \in L^2(S) : ||(I - \Delta)^p\mathbf{h}|| < \infty\}\,,$$

where $\Delta$ is the Laplacian operator on the sphere, with inner product

$$(\mathbf{g}, \mathbf{h})_p = ((I - \Delta)^p \mathbf{g}, (I - \Delta)^p \mathbf{h})$$

for all $\mathbf{g}, \mathbf{h} \in \Phi_p$, and corresponding norm $\|\mathbf{h}\|_p = (\mathbf{h}, \mathbf{h})_p^{1/2}$ for all $\mathbf{h} \in \Phi_p$. Thus if $\mathbf{h} \in \Phi_p$ and $p$ is a positive integer or half-integer, then all partial derivatives of the components of $\mathbf{h}$ up to order $2p$ are square-integrable. The spaces $\Phi_p$ are convenient for spherical geometry since the Laplacian operator is coordinate-free. The existence and uniqueness theory of Sect. 4.2 based on Sobolev spaces and inequalities carries over to the sphere using the spaces $\Phi_p$ for integers and half-integers $p$.

Tentatively let

$$\mathcal{S} = \{\mathbf{h} \in \Phi_2 : ||\mathbf{h}||_2^2 < E\}$$

for some $E < \infty$. It follows from Appendix 2b that

$$\mathcal{S} \subset \Phi_2 \subset C^1(S)\,,$$

as expected from Sect. 4.2.1. It follows from Sect. 4.2.2 that for the symmetric hyperbolic system (Eq. 58) there is a time interval $\mathcal{T} = \mathcal{T}(\mathcal{S}) = [t_0, T]$ such that, corresponding to each $\mathbf{s}_{t_0} \in \mathcal{S}$, there exists a unique classical solution $\mathbf{s}_t \in \Phi_2$ for all $t \in \mathcal{T}$, and further that $\mathbf{s} \in C^1(S \times \mathcal{T})$.

However, since $w = 2\sqrt{\Phi}$, this solution does not solve the original shallow-water system unless $w \geq 0$ on $S \times \mathcal{T}$. The differential equation for $w$ is

$$\frac{\partial w}{\partial t} + \mathbf{V} \cdot \nabla w + \tfrac{1}{2} w \nabla \cdot \mathbf{V} = 0\,,$$

and therefore along the curves $\mathbf{x} = \mathbf{x}(t) = (\lambda(t), \phi(t))$ defined by

$$\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} = \mathbf{V}(\mathbf{x}, t)\,, \tag{59}$$

the solution $w$ satisfies the ordinary differential equation

$$\frac{\mathrm{d}w}{\mathrm{d}t} + \tfrac{1}{2} w \nabla \cdot \mathbf{V} = 0\,.$$

This guarantees that if $w > 0$ initially, then $w > 0$ for all $t \in \mathcal{T}$. Thus redefine $\mathcal{S}$ as

$$\mathcal{S} = \{\mathbf{h} \in \Phi_2 : ||\mathbf{h}||_2^2 < E\} \cap \{(u, v, w) \in \Phi_2 : w > 0\}\,.$$

Note that the latter set is open in $L^2(S)$ since it is open in $\Phi_2$, and therefore $\mathcal{S} \in \mathcal{B}(L^2(S))$ since the intersection of two open sets is open. Also note that the initial-value problem for Eq. (59) is well-posed since $\mathbf{V} \in C^1(S \times \mathcal{T})$.

The classical solutions of the shallow-water equations satisfy the energy equation

$$\frac{\partial}{\partial t}\left[\Phi(u^2+v^2)+\Phi^2\right]+\nabla\cdot\left\{\left[\Phi(u^2+v^2)+2\Phi^2\right]\mathbf{V}\right\}=0\,.$$

This suggests introducing a new set of dependent variables, the energy variables $\mathbf{s}=(\alpha,\beta,\Phi)^T$ with $\alpha=u\Phi^{1/2}$ and $\beta=v\Phi^{1/2}$. In the energy variables, the physical total energy is just $\frac{1}{2}\|\mathbf{s}\|^2$, and it is conserved. It can be verified that in terms of the energy variables, the shallow-water system can be written as

$$\frac{\partial\mathbf{s}}{\partial t}+\mathbf{G}\mathbf{s}=0\,, \tag{60}$$

where $\mathbf{G}=\mathbf{G}(\mathbf{s})$ has the form

$$\mathbf{G}=\mathbf{A}\frac{1}{a\cos\phi}\frac{\partial}{\partial\lambda}+\mathbf{B}\frac{1}{a}\frac{\partial}{\partial\phi}+\frac{1}{2}\frac{1}{a\cos\phi}\left(\frac{\partial\mathbf{A}}{\partial\lambda}+\frac{\partial\mathbf{B}\cos\phi}{\partial\phi}\right)+\mathbf{C}\,,$$

with

$$\mathbf{A}=\begin{bmatrix}\alpha\Phi^{-1/2} & 0 & \frac{4}{5}\Phi^{1/2}\\ 0 & \alpha\Phi^{-1/2} & 0\\ \frac{4}{5}\Phi^{1/2} & 0 & \frac{2}{5}\alpha\Phi^{-1/2}\end{bmatrix},$$

$$\mathbf{B}=\begin{bmatrix}\beta\Phi^{-1/2} & 0 & 0\\ 0 & \beta\Phi^{-1/2} & \frac{4}{5}\Phi^{1/2}\\ 0 & \frac{4}{5}\Phi^{1/2} & \frac{2}{5}\beta\Phi^{-1/2}\end{bmatrix},$$

$$\mathbf{C}=\begin{bmatrix}0 & -(f+\frac{1}{a}\alpha\Phi^{-1/2}\tan\phi) & 0\\ f+\frac{1}{a}\alpha\Phi^{-1/2}\tan\phi & 0 & \frac{2}{5}\frac{1}{a}\Phi^{1/2}\tan\phi\\ 0 & -\frac{2}{5}\frac{1}{a}\Phi^{1/2}\tan\phi & 0\end{bmatrix}.$$

For the system given by Eq. (60) to yield the solution of the original shallow-water system requires being able to recover $u=\alpha\Phi^{-1/2}$ and $v=\beta\Phi^{-1/2}$ from $\alpha$, $\beta$ and $\Phi$. Now, products of scalars in $\Phi_2$ are also scalars in $\Phi_2$ since the elements of $\Phi_2$ are all bounded, continuous functions. But $\Phi^{-1/2}$ is not in $\Phi_2$ unless $\Phi$ is bounded from below by a positive constant. Thus for the energy variables, the initial space $\mathcal{S}$ is defined as $\mathcal{S}=\mathcal{S}_\gamma$, where

$$\mathcal{S}_\gamma=\{\mathbf{h}\in\Phi_2:\|\mathbf{h}\|_2^2<E\}\cap\{(\alpha,\beta,\Phi)\in\Phi_2:\Phi>\gamma\}$$

for some constant $\gamma>0$. It follows for the energy variables that for all $t\in\mathcal{T}$, the unique solution $\mathbf{s}_t$ corresponding to each $\mathbf{s}_{t_0}\in\mathcal{S}_\gamma$ is in $\mathcal{S}_\delta$ for some constant $\delta>0$. The symmetry of the matrices $\mathbf{A}$ and $\mathbf{B}$, the skew-symmetry of the matrix $\mathbf{C}$, and the form of the differential operator $\mathbf{G}$ imply immediately that, for all $\delta>0$, $\mathbf{G}$ is a skew-symmetric operator on $\mathcal{S}_\delta$:

$$(\mathbf{s},\mathbf{G}(\mathbf{r})\mathbf{s})=0$$

for all $\mathbf{r}, \mathbf{s} \in \mathcal{S}_\delta$. This of course implies energy conservation, as noted in Sect. 4.2.2.

Now suppose for the energy variables that $\mathbf{s}_{t_0}$ is an $\mathcal{S}_\gamma$-valued random variable. Thus $\mathbf{s}_{t_0}$ is also second-order, and it follows from Theorem 1 that $\mathbf{s}_t = \mathbf{N}_{t,t_0}(\mathbf{s}_{t_0})$ is a second-order $L^2(S)$-valued random variable, with mean $\bar{\mathbf{s}}_t \in L^2(S)$ and covariance operator $\mathcal{P}_t : L^2(S) \to L^2(S)$, which are related by

$$||\bar{\mathbf{s}}_t||^2 + \operatorname{tr} \mathcal{P}_t = ||\bar{\mathbf{s}}_{t_0}||^2 + \operatorname{tr} \mathcal{P}_{t_0} < E$$

for all $t \in \mathcal{T}$. The results of Sect. 4.2.3 show further that, for all $t \in \mathcal{T}$, $\bar{\mathbf{s}}_t \in \Phi_2 \subset C^1(S)$,

$$\mathcal{E}||\mathbf{s}_t||^2 = ||\bar{\mathbf{s}}_t||^2 + \operatorname{tr} \mathcal{P}_t \leq \mathcal{E}||\mathbf{s}_t||_2^2 < E\,,$$

and

$$\operatorname{tr} \mathcal{P}_t = \int_S \operatorname{tr} \mathbf{P}_t(\mathbf{x}, \mathbf{x})\, a^2 \cos\phi \, \mathrm{d}\phi \, \mathrm{d}\lambda\,,$$

where $\mathbf{P}_t$ is the covariance matrix of $\mathbf{s}_t$.

The discussion at the end of Sect. 4.2.3 gives the general form of $\Phi_2$-valued random variables with bounded $\Phi_2$ norm. The Sobolev-type inequality (Eq. 73) of Appendix 2b then suggests a way of ensuring that such a random variable $\mathbf{s}_{t_0}$ is an $\mathcal{S}_\gamma$-valued random variable. Let

$$\Phi_{t_0}(\omega) = \overline{\Phi}_{t_0} + \Phi'_{t_0}(\omega)$$

for all $\omega \in \Omega$. The series expansions in Appendix 2b give the form of every scalar in $\Phi_2$. Suppose that $\overline{\Phi}_{t_0} \in \Phi_2$ with $\overline{\Phi}_{t_0} > \mu > \gamma > 0$. Equation (73) shows how to ensure that $|\Phi'_{t_0}(\omega)| < \mu - \gamma$ for all $\omega \in \Omega$, and therefore that $\Phi_{t_0}(\omega) > \gamma$ for all $\omega \in \Omega$.

## 6 Concluding Remarks

This chapter has formulated the principle of energetic consistency (PEC), demonstrated its validity for a wide range of non-linear dynamical systems, and illustrated its application to distinguishing between artificial and genuine uncertainties in ensemble Kalman filter (EnKF) methods. It has been argued that because EnKF methods rely at least tacitly on the minimum variance optimality criterion, it is natural to choose the state variables in EnKF schemes to be energy variables for the dynamical system being observed. This requires only that the observation operators be expressed in terms of energy variables. Once the state variables are chosen to be energy variables, the PEC can be applied.

The PEC has been used to show that some of the assumptions and approximations made in EnKF schemes give rise to artificial energetic sources or sinks of uncertainty, while others are energetically neutral. It has also been shown that the PEC can be implemented numerically to determine the magnitude of artificial sources and sinks, which is problem-dependent.

The PEC was used to show, in particular, that the spurious numerical dissipation typical of discrete dynamical models generally gives rise to an artificial energetic sink of uncertainty, which can easily result in exponential decay of total variance – ensemble collapse – if left untreated. The simple hypotheses under which this behaviour was shown to occur indicates that this is a generic problem, not only for ensemble filters, but for filtering schemes in general. That this has not yet been obvious for EnKF schemes is perhaps because such artificial sinks of uncertainty cannot be clearly identified and distinguished from genuine ones unless the state variables are chosen to be energy variables. Since the state variables have not traditionally been chosen to be energy variables, such sinks have usually been compensated for by an artificially large "model error" term or by "covariance inflation."

The general remedy suggested in Sect. 2.4.4 for this spurious loss of variance is simply to pre- and post-multiply the ensemble conditional covariance matrix by appropriate operators that directly counteract the spurious numerical dissipation, according to Eq. (37). Weak-constraint, long-window 4D-Var methods approximate the estimation error covariance evolution of the extended Kalman filter (Cox 1964, Eqs. 40, 41, 42, and 43; Fisher et al. 2005). To the extent that the covariance matrix of the extended Kalman filter approximates the conditional covariance matrix, Eq. (37) then applies to long-window 4D-Var methods, and could be implemented in 4D-Var simply by pre-multiplying the tangent linear model by the matrix $\mathbf{I} - \Delta t_k \mathbf{D}$, where the matrix $\mathbf{D}$ is obtained by linearizing about the 4D-Var trajectory instead of the ensemble mean state as in Eq. (31). Such a resolution of the variance loss problem may be necessary for weak-constraint 4D-Var methods (Trémolet 2006, 2007) to function properly as filters when used with a long time window. It should be noted, however, that as an approximation to the extended Kalman filter rather than the full second-moment closure dynamics, 4D-Var methods omit the so-called nonlinear bias term in the second-moment closure equation for the mean state (Cohn 1993, pp. 3131–3132), and therefore lack energetic consistency (Cohn 2009).

Among the results of the more theoretical sections of this chapter that may have important practical implications for data assimilation is the breakdown of the Gaussian hypothesis. It has been shown that the stochastic initial-value problem for symmetric hyperbolic systems is well-posed under natural hypotheses, but that in general the state cannot be Gaussian-distributed at any time. This result is due not to the fact that there are often state variables that are restricted to be positive, and which therefore can only be Gaussian-distributed as an approximation, sometimes a good one, but is rather a consequence of well-posedness and of boundedness of the solution operator. Also, the usual Kalman-type observation update formula, which for its probabilistic interpretation is based on an assumption that the conditional mean state is Gaussian-distributed, lacks energetic consistency except in the sense of expectation. Similarly, the probabilistic interpretation of 4D-Var is based on an

assumption that the initial state is Gaussian-distributed. Thus it will be worthwhile to try to formulate generalized observation updates for use in EnKF methods, and also generalized versions of 4D-Var methods. Data assimilation is still a young field, and it is clear that much work lies ahead.

**Acknowledgments** The framework of this chapter was motivated in large part by the tracer assimilation work of Ménard et al. (2000) and Ménard and Chang (2000), which identified the problem of spurious variance loss and also made clear the usefulness of having a conserved scalar quantity to work with in data assimilation. The author would like to thank the editors of this book for their diligence and superb work. The generous support of NASA's Modeling, Analysis and Prediction program is also gratefully acknowledged.

## Appendix 1: Random Variables Taking Values in Hilbert Space

Appendix 1a defines Hilbert space-valued random variables and gives some of their main properties. Appendices 1b–1d give the definition, main properties and general construction, respectively, of Hilbert space-valued random variables of second order. Definitions of basic terms used in this appendix are provided in Appendix 3. Further treatment of Hilbert space-valued random variables, and of random variables taking values in more general spaces, can be found in the books of Itô (1984) and Kallianpur and Xiong (1995).

Hilbert space-valued random variables, like scalar random variables, are defined with reference to some probability space $(\Omega, \mathcal{F}, P)$, with $\Omega$ the sample space, $\mathcal{F}$ the event space and $P$ the probability measure. Thus throughout this appendix, a probability space $(\Omega, \mathcal{F}, P)$ is considered to be given. The expectation operator is denoted by $\mathcal{E}$. It is assumed that the given probability space is complete.

A real, separable Hilbert space $\mathcal{H}$ is also considered to be given. The inner product and corresponding norm on $\mathcal{H}$ are denoted by $(\cdot, \cdot)$ and $\| \cdot \|$, respectively. The Borel field generated by the open sets in $\mathcal{H}$ is denoted by $\mathcal{B}(\mathcal{H})$, i.e., $\mathcal{B}(\mathcal{H})$ is the smallest $\sigma$-algebra of sets in $\mathcal{H}$ that contains all the open sets in $\mathcal{H}$. Recall that every separable Hilbert space has a countable orthonormal basis, and that every orthonormal basis of a separable Hilbert space has the same number of elements $N \leq \infty$, the dimension of the space. For notational convenience it is assumed in this appendix that $\mathcal{H}$ is infinite-dimensional, with $\{\mathbf{h}_i\}_{i=1}^{\infty}$ denoting an orthonormal basis for $\mathcal{H}$. The results of this appendix hold just as well in the finite-dimensional case, by taking $\{\mathbf{h}_i\}_{i=1}^{N}$, $N < \infty$, as an orthonormal basis for $\mathcal{H}$, and by replacing infinite sums by finite ones.

### *1a $\mathcal{H}$-Valued Random Variables*

Recall that if $X$ and $Y$ are sets, $\mathbf{f}$ is a map from $X$ into $Y$, and $B$ is a subset of $Y$, then the set

$$\mathbf{f}^{-1}[B] = \{\mathbf{x} \in X : \mathbf{f}(\mathbf{x}) \in B\}$$

is called the inverse image of $B$ (under $\mathbf{f}$). Recall also that the event space $\mathcal{F}$ of the probability space $(\Omega, \mathcal{F}, P)$ consists of the measurable subsets of $\Omega$, which are called events.

Let $(Y, \mathcal{C})$ be a measurable space, i.e., $Y$ is a set and $\mathcal{C}$ is a $\sigma$-algebra of subsets of $Y$. A map $\mathbf{f} : \Omega \to Y$ is called a $(Y, \mathcal{C})$-*valued random variable* if the inverse image of every set $C$ in the collection $\mathcal{C}$ is an event, i.e., if $\mathbf{f}^{-1}[C] \in \mathcal{F}$ for every set $C \in \mathcal{C}$ (e.g. Itô 1984, p. 18; Kallianpur and Xiong 1995, p. 86; see also Reed and Simon 1972, p. 24).

Thus an $(\mathcal{H}, \mathcal{B}(\mathcal{H}))$-valued random variable is a map $\mathbf{s} : \Omega \to \mathcal{H}$ such that

$$\{\omega \in \Omega : \mathbf{s}(\omega) \in B\} \in \mathcal{F}$$

for every set $B \in \mathcal{B}(\mathcal{H})$. Hereafter, an $(\mathcal{H}, \mathcal{B}(\mathcal{H}))$-valued random variable is called simply an $\mathcal{H}$-valued random variable, with the understanding that this always means an $(\mathcal{H}, \mathcal{B}(\mathcal{H}))$-valued random variable. An equivalent definition of $\mathcal{H}$-valued random variables, expressed in terms of scalar random variables, is given in Appendix 1b.

Let $\mathcal{S}$ be a non-empty set in $\mathcal{B}(\mathcal{H})$. It follows that the collection $\mathcal{B}_{\mathcal{S}}(\mathcal{H})$ of all sets in $\mathcal{B}(\mathcal{H})$ that are subsets of $\mathcal{S}$,

$$\mathcal{B}_{\mathcal{S}}(\mathcal{H}) = \{B \in \mathcal{B}(\mathcal{H}) : B \subset \mathcal{S}\},$$

is a $\sigma$-algebra of subsets of $\mathcal{S}$, namely, the collection of all sets $C$ of the form $C = B \cap \mathcal{S}$ with $B \in \mathcal{B}(\mathcal{H})$. Hence $(\mathcal{S}, \mathcal{B}_{\mathcal{S}}(\mathcal{H}))$ is a measurable space, and an $(\mathcal{S}, \mathcal{B}_{\mathcal{S}}(\mathcal{H}))$-valued random variable is a map $\mathbf{s} : \Omega \to \mathcal{S}$ such that

$$\{\omega \in \Omega : \mathbf{s}(\omega) \in C\} \in \mathcal{F}$$

for every set $C \in \mathcal{B}_{\mathcal{S}}(\mathcal{H})$. Hereafter, an $(\mathcal{S}, \mathcal{B}_{\mathcal{S}}(\mathcal{H}))$-valued random variable is called simply an $\mathcal{S}$-valued random variable, with the understanding that this always means an $(\mathcal{S}, \mathcal{B}_{\mathcal{S}}(\mathcal{H}))$-valued random variable.

It follows by definition that every $\mathcal{S}$-valued random variable is an $\mathcal{H}$-valued random variable, for if $\mathbf{s} : \Omega \to \mathcal{S}$ and $\mathbf{s}^{-1}[C] \in \mathcal{F}$ for every set $C \in \mathcal{B}_{\mathcal{S}}(\mathcal{H})$, then $\mathbf{s}^{-1}[B] = \mathbf{s}^{-1}[B \cap \mathcal{S}] \in \mathcal{F}$ for every set $B \in \mathcal{B}(\mathcal{H})$. Also, every $\mathcal{H}$-valued random variable taking values only in $\mathcal{S}$ is an $\mathcal{S}$-valued random variable, for if $\mathbf{s} : \Omega \to \mathcal{S}$ and $\mathbf{s}^{-1}[B] \in \mathcal{F}$ for every set $B \in \mathcal{B}(\mathcal{H})$, then in particular $\mathbf{s}^{-1}[C] \in \mathcal{F}$ for every set $C \in \mathcal{B}_{\mathcal{S}}(\mathcal{H})$.

Finally, let $\mathbf{N}$ be a continuous map from $\mathcal{S}$ into $\mathcal{H}$. It follows that if $\mathbf{s}$ is an $\mathcal{S}$-valued random variable, then $\mathbf{N}(\mathbf{s})$ is an $\mathcal{H}$-valued random variable, i.e. that

$$\{\omega \in \Omega : \mathbf{N}(\mathbf{s}(\omega)) \in B\} \in \mathcal{F}$$

for every set $B \in \mathcal{B}(\mathcal{H})$. To see this, note first that

$$\{\omega \in \Omega : \mathbf{N}(\mathbf{s}(\omega)) \in B\} = \mathbf{s}^{-1}[\mathbf{N}^{-1}[B]]\,,$$

and consider the class of sets $E$ in $\mathcal{H}$ such that $\mathbf{N}^{-1}[E] \in \mathcal{B}_{\mathcal{S}}(\mathcal{H})$. It can be checked that this class of sets is a $\sigma$-algebra. Moreover, this class contains all the open sets in $\mathcal{H}$, because if $O$ is an open set in $\mathcal{H}$ then $\mathbf{N}^{-1}[O]$ is also an open set in $\mathcal{H}$ by the continuity of $\mathbf{N}$ (e.g. Reed and Simon 1972, p. 8) and so

$$C = \mathbf{N}^{-1}[O] = \mathbf{N}^{-1}[O] \cap \mathcal{S} \in \mathcal{B}_{\mathcal{S}}(\mathcal{H})\,.$$

But $\mathcal{B}(\mathcal{H})$ is the smallest $\sigma$-algebra containing all the open sets in $\mathcal{H}$, hence this class includes $\mathcal{B}(\mathcal{H})$, i.e., $\mathbf{N}^{-1}[B] \in \mathcal{B}_{\mathcal{S}}(\mathcal{H})$ for every set $B \in \mathcal{B}(\mathcal{H})$. If $\mathbf{s}$ is an $\mathcal{S}$-valued random variable then $\mathbf{s}^{-1}[C] \in \mathcal{F}$ for every set $C \in \mathcal{B}_{\mathcal{S}}(\mathcal{H})$, and therefore $\mathbf{s}^{-1}[\mathbf{N}^{-1}[B]] \in \mathcal{F}$ for every set $B \in \mathcal{B}(\mathcal{H})$, i.e., $\mathbf{N}(\mathbf{s})$ is an $\mathcal{H}$-valued random variable.

## *1b Second-Order $\mathcal{H}$-Valued Random Variables*

If $\mathbf{s}$ is an $\mathcal{H}$-valued random variable and $\mathbf{h} \in \mathcal{H}$, then by the Schwarz inequality,

$$|(\mathbf{h}, \mathbf{s}(\omega))| \leq \|\mathbf{h}\|\,\|\mathbf{s}(\omega)\| < \infty \tag{61}$$

for all $\omega \in \Omega$, so for each fixed $\mathbf{h} \in \mathcal{H}$, the inner product ($\mathbf{h}$,$\mathbf{s}$) is a map from $\Omega$ into $\mathbb{R}$. In fact, it can be shown (e.g. Kallianpur and Xiong 1995, Corollary 3.1.1(b), p. 87) that a map $\mathbf{s} : \Omega \to \mathcal{H}$ is an $\mathcal{H}$-valued random variable if, and only if, ($\mathbf{h}$,$\mathbf{s}$) is a scalar random variable for every $\mathbf{h} \in \mathcal{H}$. That is, a map $\mathbf{s} : \Omega \to \mathcal{H}$ is an $\mathcal{H}$-valued random variable if, and only if,

$$\{\omega \in \Omega : (\mathbf{h}, \mathbf{s}(\omega)) \leq \alpha\} \in \mathcal{F}$$

for every $\mathbf{h} \in \mathcal{H}$ and every $\alpha \in \mathbb{R}$.

It follows that if $\mathbf{s}$ is an $\mathcal{H}$-valued random variable, then $\|\mathbf{s}\|^2$ is a scalar random variable, that is,

$$\{\omega \in \Omega : ||\mathbf{s}(\omega)||^2 \leq \alpha\} \in \mathcal{F}$$

for every $\alpha \in \mathbb{R}$. To see this, observe that if $\mathbf{s}$ is an $\mathcal{H}$-valued random variable, then $(\mathbf{h}_i, \mathbf{s})$ for $i = 1, 2, \ldots$ are scalar random variables, hence

$$s_n = \sum_{i=1}^{n} (\mathbf{h}_i, \mathbf{s})^2$$

are scalar random variables with $0 \leq s_n \leq s_{n+1}$ for $n = 1, 2, \ldots$, and by Parseval's relation,

$$||\mathbf{s}(\omega)||^2 = \sum_{i=1}^{\infty} (\mathbf{h}_i, \mathbf{s}(\omega))^2 = \lim_{n\to\infty} s_n(\omega)$$

for all $\omega \in \Omega$. Thus $\|\mathbf{s}\|^2$ is the limit of an increasing sequence of non-negative scalar random variables, and is therefore a (non-negative) scalar random variable.

If a map $\mathbf{s} : \Omega \to \mathcal{H}$ is an $\mathcal{H}$-valued random variable, then since $\|\mathbf{s}\|^2 \geq 0$ is a scalar random variable, it follows that $\mathcal{E}\|\mathbf{s}\|^2$ is defined and either $\mathcal{E}\|\mathbf{s}\|^2 = \infty$ or $\mathcal{E}\|\mathbf{s}\|^2 < \infty$. An $\mathcal{H}$-valued random variable $\mathbf{s}$ is called *second-order* if $\mathcal{E}\|\mathbf{s}\|^2 < \infty$.

## *1c Properties of Second-Order $\mathcal{H}$-Valued Random Variables*

In this subsection let $\mathbf{s} : \Omega \to \mathcal{H}$ be a second-order $\mathcal{H}$-valued random variable. Since $\mathcal{E}\|\mathbf{s}\|^2 < \infty$, it follows from Eq. (61) that

$$\mathcal{E}(\mathbf{h}, \mathbf{s})^2 \leq \|\mathbf{h}\|^2 \, \mathcal{E}\|\mathbf{s}\|^2 < \infty \tag{62}$$

for each $\mathbf{h} \in \mathcal{H}$. Thus, for each $\mathbf{h} \in \mathcal{H}$, ($\mathbf{h}$,$\mathbf{s}$) is a second-order scalar random variable, and therefore its mean is defined and finite. The mean of ($\mathbf{h}$,$\mathbf{s}$) will be denoted by

$$m[\mathbf{h}] = \mathcal{E}(\mathbf{h}, \mathbf{s}),$$

for each $\mathbf{h} \in \mathcal{H}$. Since $\mathcal{E}\|\mathbf{s}\|^2 < \infty$, $\|\mathbf{s}\|$ is a second-order scalar random variable, and its mean $M = \mathcal{E}\|\mathbf{s}\|$ satisfies $0 \leq M \leq \left(\mathcal{E}\|\mathbf{s}\|^2\right)^{1/2} < \infty$. Now

$$|m[\mathbf{h}]| = |\mathcal{E}(\mathbf{h}, \mathbf{s})| \leq \mathcal{E}\,|(\mathbf{h}, \mathbf{s})| \leq M||\mathbf{h}||$$

for each $\mathbf{h} \in \mathcal{H}$, by Eq. (61), and also

$$m[\alpha\mathbf{g} + \beta\mathbf{h}] = \alpha m[\mathbf{g}] + \beta m[\mathbf{h}]$$

for each $\mathbf{g}, \mathbf{h} \in \mathcal{H}$ and $\alpha, \beta \in \mathbb{R}$. Thus $m[\cdot]$ is a bounded linear functional on $\mathcal{H}$, and by the Riesz representation theorem for Hilbert space (e.g. Royden 1968, p. 213; Reed and Simon 1972, p. 43) this implies that there exists a unique element $\bar{\mathbf{s}} \in \mathcal{H}$, called the *mean* of $\mathbf{s}$, such that

$$m[\mathbf{h}] = (\mathbf{h}, \bar{\mathbf{s}})$$

for each $\mathbf{h} \in \mathcal{H}$. Thus the mean $\bar{\mathbf{s}}$ of $\mathbf{s}$ is defined uniquely in $\mathcal{H}$, and satisfies $(\mathbf{h}, \bar{\mathbf{s}}) = \mathcal{E}(\mathbf{h}, \mathbf{s})$ for every $\mathbf{h} \in \mathcal{H}$.

Now let $\mathbf{s}'(\omega) = \mathbf{s}(\omega) - \bar{\mathbf{s}}$ for each $\omega \in \Omega$. Since

$$||\mathbf{s}'(\omega)|| \leq ||\mathbf{s}(\omega)|| + ||\bar{\mathbf{s}}|| < \infty$$

for each $\omega \in \Omega$, $\mathbf{s}' = \mathbf{s} - \bar{\mathbf{s}}$ is a map from $\Omega$ into $\mathcal{H}$. Furthermore, for every $\mathbf{h} \in \mathcal{H}$, $(\mathbf{h},\mathbf{s})$ is a scalar random variable, $|(\mathbf{h}, \bar{\mathbf{s}})| < \infty$, and $|(\mathbf{h}, \mathbf{s}(\omega))| < \infty$ for each $\omega \in \Omega$. Therefore $(\mathbf{h}, \mathbf{s}') = (\mathbf{h}, \mathbf{s}) - (\mathbf{h}, \bar{\mathbf{s}})$ is a scalar random variable for every $\mathbf{h} \in \mathcal{H}$, and hence $\mathbf{s}'$ is an $\mathcal{H}$-valued random variable. Also,

$$\mathcal{E}(\mathbf{h}, \mathbf{s}') = \mathcal{E}(\mathbf{h}, \mathbf{s}) - (\mathbf{h}, \bar{\mathbf{s}}) = 0$$

for every $\mathbf{h} \in \mathcal{H}$, so the mean of $\mathbf{s}'$ is $\mathbf{0} \in \mathcal{H}$. Thus

$$\mathcal{E}\|\mathbf{s}\|^2 = \mathcal{E}(\bar{\mathbf{s}} + \mathbf{s}', \bar{\mathbf{s}} + \mathbf{s}') = \|\bar{\mathbf{s}}\|^2 + \mathcal{E}\|\mathbf{s}'\|^2 \tag{63}$$

and, in particular, $\mathcal{E}\|\mathbf{s}'\|^2 \leq \mathcal{E}\|\mathbf{s}\|^2 < \infty$. Therefore $\mathbf{s}' : \Omega \to \mathcal{H}$ is a second-order $\mathcal{H}$-valued random variable, and $\|\mathbf{s}'\|$ is a second-order scalar random variable.

Since $\mathbf{s}'$ is a second-order $\mathcal{H}$-valued random variable, $(\mathbf{g}, \mathbf{s}')$ and $(\mathbf{h}, \mathbf{s}')$ are second-order scalar random variables, for each $\mathbf{g}, \mathbf{h} \in \mathcal{H}$. Therefore the expectation

$$C[\mathbf{g}, \mathbf{h}] = \mathcal{E}(\mathbf{g}, \mathbf{s}')(\mathbf{h}, \mathbf{s}')$$

is defined for all $\mathbf{g}, \mathbf{h} \in \mathcal{H}$, and in fact

$$|C[\mathbf{g}, \mathbf{h}]| \leq \mathcal{E}\left|(\mathbf{g}, \mathbf{s}')(\mathbf{h}, \mathbf{s}')\right| \leq \left[\mathcal{E}(\mathbf{g}, \mathbf{s}')^2\right]^{1/2} \left[\mathcal{E}(\mathbf{h}, \mathbf{s}')^2\right]^{1/2} \leq \|\mathbf{g}\|\,\|\mathbf{h}\|\,\mathcal{E}\|\mathbf{s}'\|^2 . \tag{64}$$

The functional $C$, called the *covariance functional* of $\mathbf{s}$, is also linear in its two arguments. Thus $C[\cdot, \cdot]$ is a bounded bilinear functional on $\mathcal{H} \times \mathcal{H}$. It follows (e.g. Rudin 1991, Theorem 12.8, p. 310) that there exists a unique bounded linear operator $\mathcal{P} : \mathcal{H} \to \mathcal{H}$, called the *covariance operator* of $\mathbf{s}$, such that

$$C[\mathbf{g}, \mathbf{h}] = (\mathbf{g}, \mathcal{P}\mathbf{h})$$

for each $\mathbf{g}, \mathbf{h} \in \mathcal{H}$. The covariance operator $\mathcal{P}$ is *self-adjoint*, i.e., $(\mathcal{P}\mathbf{g}, \mathbf{h}) = (\mathbf{g}, \mathcal{P}\mathbf{h})$ for all $\mathbf{g}, \mathbf{h} \in \mathcal{H}$, since the covariance functional is symmetric, $C[\mathbf{h}, \mathbf{g}] = C[\mathbf{g}, \mathbf{h}]$ for all $\mathbf{g}, \mathbf{h} \in \mathcal{H}$. The covariance operator is also *positive semidefinite*, i.e., $(\mathbf{h}, \mathcal{P}\mathbf{h}) \geq 0$ for all $\mathbf{h} \in \mathcal{H}$, since

$$(\mathbf{h}, \mathcal{P}\mathbf{h}) = C[\mathbf{h}, \mathbf{h}] = \mathcal{E}(\mathbf{h}, \mathbf{s}')^2 \geq 0$$

for all $\mathbf{h} \in \mathcal{H}$.

Now consider the second-order scalar random variable $\|\mathbf{s}'\|$. By Parseval's relation,

$$||\mathbf{s}'(\omega)||^2 = \sum_{i=1}^{\infty} (\mathbf{h}_i, \mathbf{s}'(\omega))^2$$

for all $\omega \in \Omega$, and therefore

$$\mathcal{E}||\mathbf{s}'||^2 = \sum_{i=1}^{\infty} \mathcal{E}(\mathbf{h}_i, \mathbf{s}')^2 \,,$$

because $\{(\mathbf{h}_i, \mathbf{s}')^2\}_{i=1}^{\infty}$ is a sequence of non-negative random variables. Furthermore,

$$\mathcal{E}(\mathbf{h}_i, \mathbf{s}')^2 = (\mathbf{h}_i, \mathcal{P}\mathbf{h}_i) \tag{65}$$

for $i = 1, 2, \ldots$, by definition of the covariance operator $\mathcal{P}$, and therefore

$$\mathcal{E}||\mathbf{s}'||^2 = \sum_{i=1}^{\infty} (\mathbf{h}_i, \mathcal{P}\mathbf{h}_i) \,.$$

The summation on the right-hand side, called the *trace* of $\mathcal{P}$ and written tr $\mathcal{P}$, is independent of the choice of orthonormal basis $\{\mathbf{h}_i\}_{i=1}^{\infty}$ for $\mathcal{H}$, for any positive, semidefinite bounded linear operator from $\mathcal{H}$ into $\mathcal{H}$ (e.g. Reed and Simon 1972, Theorem VI.18, p. 206). Thus

$$\text{tr}\,\mathcal{P} = \sum_{i=1}^{\infty} (\mathbf{h}_i, \mathcal{P}\mathbf{h}_i) = \mathcal{E}||\mathbf{s}'||^2 < \infty \,,$$

and Eq. (63) can be written as

$$\mathcal{E}\|\mathbf{s}\|^2 = \|\bar{\mathbf{s}}\|^2 + \text{tr}\,\mathcal{P} \,, \tag{66}$$

which is a generalization of Eq. (80) to second-order $\mathcal{H}$-valued random variables.

Since tr $\mathcal{P} < \infty$, $\mathcal{P}$ is a *trace class* operator, and therefore also a *compact* operator (e.g. Reed and Simon 1972, Theorem VI.21, p. 209). Since $\mathcal{P}$ is self-adjoint in addition to being compact, it follows from the Hilbert-Schmidt theorem (e.g. Reed and Simon 1972, Theorem VI.16, p. 203) that there exists an orthonormal basis for $\mathcal{H}$ which consists of *eigenvectors* $\{\tilde{\mathbf{h}}_i\}_{i=1}^{\infty}$ of $\mathcal{P}$,

$$\mathcal{P}\tilde{\mathbf{h}}_i = \lambda_i \tilde{\mathbf{h}}_i$$

for $i = 1, 2, \ldots$, where the corresponding *eigenvalues* $\lambda_i = (\tilde{\mathbf{h}}_i, \mathcal{P}\tilde{\mathbf{h}}_i)$ for $i = 1, 2, \ldots$ are all real numbers and satisfy $\lambda_i \to 0$ as $i \to \infty$. In fact, the eigenvalues are all non-negative since $\mathcal{P}$ is positive semidefinite, and therefore $\lambda_i = \|\mathcal{P}\tilde{\mathbf{h}}_i\|$ for $i = 1, 2, \ldots$. Further, it follows from Eq. (65) that

$$\lambda_i = (\tilde{\mathbf{h}}_i, \mathcal{P}\tilde{\mathbf{h}}_i) = \mathcal{E}(\tilde{\mathbf{h}}_i, \mathbf{s}')^2 = \sigma_i^2 \,,$$

where $\sigma_i^2$ is the variance of the scalar random variable $(\tilde{\mathbf{h}}_i, \mathbf{s})$, for $i = 1, 2, \ldots$. By the definition of tr $\mathcal{P}$,

$$\mathcal{E}||\mathbf{s}'||^2 = \text{tr}\,\mathcal{P} = \sum_{i=1}^{\infty} (\tilde{\mathbf{h}}_i, \mathcal{P}\tilde{\mathbf{h}}_i) = \sum_{i=1}^{\infty} \lambda_i < \infty \,.$$

Thus the eigenvalues $\{\lambda_i\}_{i=1}^{\infty}$ of $\mathcal{P}$ are the variances $\{\sigma_i^2\}_{i=1}^{\infty}$ and have finite sum tr $\mathcal{P}$. Equation (66) can then be rewritten as

$$\mathcal{E}\|\mathbf{s}\|^2 = \|\bar{\mathbf{s}}\|^2 + \sum_{i=1}^{\infty} \sigma_i^2 \,, \tag{67}$$

which is another generalization of Eq. (80).

Since every $\mathbf{h} \in \mathcal{H}$ has the representation $\mathbf{h} = \sum_{i=1}^{\infty} (\mathbf{h}_i, \mathbf{h})\mathbf{h}_i$, and since $\mathcal{P}\mathbf{h} \in \mathcal{H}$ for every $\mathbf{h} \in \mathcal{H}$, taking $\mathbf{h}_i = \tilde{\mathbf{h}}_i$ and using the fact that

$$(\tilde{\mathbf{h}}_i, \mathcal{P}\mathbf{h}) = (\mathcal{P}\tilde{\mathbf{h}}_i, \mathbf{h}) = \lambda_i(\tilde{\mathbf{h}}_i, \mathbf{h}) = \sigma_i^2(\tilde{\mathbf{h}}_i, \mathbf{h})$$

for $i = 1, 2, \ldots$, gives the following representation for $\mathcal{P}$:

$$\mathcal{P}\mathbf{h} = \sum_{i=1}^{\infty} \sigma_i^2 (\tilde{\mathbf{h}}_i, \mathbf{h})\tilde{\mathbf{h}}_i \tag{68}$$

for every $\mathbf{h} \in \mathcal{H}$. Thus the expectation $\mathcal{E}(\mathbf{g}, \mathbf{s}')(\mathbf{h}, \mathbf{s}')$ is given by the convergent series

$$\mathcal{E}(\mathbf{g}, \mathbf{s}')(\mathbf{h}, \mathbf{s}') = C[\mathbf{g}, \mathbf{h}] = (\mathbf{g}, \mathcal{P}\mathbf{h}) = \sum_{i=1}^{\infty} \sigma_i^2 (\tilde{\mathbf{h}}_i, \mathbf{g})(\tilde{\mathbf{h}}_i, \mathbf{h}) \,,$$

for every $\mathbf{g}, \mathbf{h} \in \mathcal{H}$.

Finally, since $\mathcal{P}$ is a positive semidefinite bounded linear operator from $\mathcal{H}$ into $\mathcal{H}$, there exists a unique positive semidefinite bounded linear operator $\mathcal{P}^{1/2} : \mathcal{H} \to \mathcal{H}$, called the *square root* of $\mathcal{P}$, that satisfies $\left(\mathcal{P}^{1/2}\right)^2 = \mathcal{P}$ (e.g. Reed and Simon 1972, Theorem VI.9, p. 196). Since $\mathcal{P}$ is also self-adjoint and trace class, $\mathcal{P}^{1/2}$ is self-adjoint and *Hilbert-Schmidt* (e.g. Reed and Simon 1972, p. 210), with the same eigenvectors as $\mathcal{P}$ and with eigenvalues that are the non-negative square roots of the corresponding eigenvalues of $\mathcal{P}$. That is,

$$\mathcal{P}^{1/2}\tilde{\mathbf{h}}_i = \sigma_i \tilde{\mathbf{h}}_i \,,$$

where $\sigma_i = \lambda_i^{1/2} = [\mathcal{E}(\tilde{\mathbf{h}}_i, \mathbf{s}')^2]^{1/2}$, for $i = 1, 2, \ldots$. Therefore $\sigma_i = (\tilde{\mathbf{h}}_i, \mathcal{P}^{1/2}\tilde{\mathbf{h}}_i) = \|\mathcal{P}^{1/2}\tilde{\mathbf{h}}_i\|$ for $i = 1, 2, \ldots$, and $\mathcal{P}^{1/2}$ has the representation

$$\mathcal{P}^{1/2}\mathbf{h} = \sum_{i=1}^{\infty} \sigma_i(\tilde{\mathbf{h}}_i, \mathbf{h})\tilde{\mathbf{h}}_i \tag{69}$$

for every $\mathbf{h} \in \mathcal{H}$.

## *1d Construction of Second-Order $\mathcal{H}$-Valued Random Variables*

It will now be shown how essentially all second-order $\mathcal{H}$-valued random variables can be constructed. This will be accomplished by first reconsidering, in a suggestive notation, the defining properties of every second-order $\mathcal{H}$-valued random variable. The construction given here is by Itô's regularization theorem (Itô 1984, Theorem 2.3.3, p. 27; Kallianpur and Xiong 1995, Theorem 3.1.2, p. 87) applied to $\mathcal{H}$, and amounts to formalizing on $\mathcal{H}$ the usual construction of infinite-dimensional random variables through random Fourier series.

For the moment, fix a second-order $\mathcal{H}$-valued random variable $\mathbf{s}$, and consider the behaviour of

$$s[\mathbf{h}] = (\mathbf{h}, \mathbf{s})$$

as a functional of $\mathbf{h} \in \mathcal{H}$, that is, as $\mathbf{h}$ varies throughout $\mathcal{H}$. The functional $s[\cdot]$ has three important properties. First, on evaluation at any $\mathbf{h} \in \mathcal{H}$, it is a scalar random variable, with

$$s[\mathbf{h}](\omega) = (\mathbf{h}, \mathbf{s}(\omega))$$

for each $\omega \in \Omega$, since $\mathbf{s} : \Omega \to \mathcal{H}$ is an $\mathcal{H}$-valued random variable. Thus $s[\cdot]$ is a map from $\mathcal{H}$ into the set of scalar random variables on $(\Omega, \mathcal{F}, P)$. Second, this map is linear,

$$s[\alpha\mathbf{g} + \beta\mathbf{h}] = \alpha s[\mathbf{g}] + \beta s[\mathbf{h}]$$

for all $\mathbf{g}, \mathbf{h} \in \mathcal{H}$ and $\alpha, \beta \in \mathbb{R}$, by linearity of the inner product. Third, according to Eq. (62),

$$(\mathcal{E}s^2[\mathbf{h}])^{1/2} \leq \gamma\|\mathbf{h}\|\,, \tag{70}$$

where

$$\gamma = (\mathcal{E}||\mathbf{s}||^2)^{1/2} < \infty\,,$$

since the $\mathcal{H}$-valued random variable $\mathbf{s}$ is second-order. Thus $s[\cdot]$ is a linear map from $\mathcal{H}$ into the set of second-order scalar random variables on $(\Omega, \mathcal{F}, P)$.

Now recall the space $L^2(\Omega, \mathcal{F}, P)$, whose elements are the equivalence classes of second-order scalar random variables, where two scalar random variables are called equivalent if they are equal wp1 (with probability one; see Appendix 3c). The space $L^2(\Omega, \mathcal{F}, P)$ is a Hilbert space, with the inner product of any two elements $\tilde{r}, \tilde{s} \in L^2(\Omega, \mathcal{F}, P)$ given by $\mathcal{E}\tilde{r}\tilde{s}$ and the corresponding norm of any element $\tilde{s} \in L^2(\Omega, \mathcal{F}, P)$ given by $(\mathcal{E}\tilde{s}^2)^{1/2}$. The inequality given by Eq. (70) states that the functional $s[\cdot]$ is bounded, when viewed as a map from $\mathcal{H}$ into $L^2(\Omega, \mathcal{F}, P)$.

A map $s[\cdot]$ from $\mathcal{H}$ into the set of scalar random variables on $(\Omega, \mathcal{F}, P)$, which is linear in the sense that if $\mathbf{g}, \mathbf{h} \in \mathcal{H}$ and $\alpha, \beta \in \mathbb{R}$ then

$$s[\alpha\mathbf{g} + \beta\mathbf{h}] = \alpha s[\mathbf{g}] + \beta s[\mathbf{h}] \text{ wp1} ,$$

is called a *random linear functional* (e.g. Itô 1984, p. 22; Omatu and Seinfeld 1989, p. 48). Observe that the set of $\omega \in \Omega$ of probability measure zero where linearity fails to hold can depend on $\alpha, \beta, \mathbf{g}$ and $\mathbf{h}$. If linearity holds for all $\omega \in \Omega$, for all $\mathbf{g}, \mathbf{h} \in \mathcal{H}$ and $\alpha, \beta \in \mathbb{R}$, then the random linear functional is called *perfect*. If $s[\cdot]$ is a random linear functional and there is a constant $\gamma \in \mathbb{R}$ such that Eq. (70) holds for all $\mathbf{h} \in \mathcal{H}$, then the random linear functional is called *second-order*. Thus, given any particular $\mathcal{H}$-valued random variable $\mathbf{s}$, the map $s[\cdot]$ defined for all $\mathbf{h} \in \mathcal{H}$ by $s[\mathbf{h}] = (\mathbf{h}, \mathbf{s})$ is a perfect random linear functional, and if $\mathbf{s}$ is second-order then so is $s[\cdot]$.

Now it will be shown that a random linear functional $s[\cdot]$ is second-order if, and only if,

$$\sum_{i=1}^{\infty} \mathcal{E}s^2[\mathbf{h}_i] < \infty . \tag{71}$$

In particular, a collection $\{s_i\}_{i=1}^{\infty}$ of scalar random variables with $\sum_{i=1}^{\infty} \mathcal{E}s_i^2 < \infty$ can be used to define a second-order random linear functional, by setting $s[\mathbf{h}_i] = s_i$ for $i = 1, 2, \ldots$. It will then be shown how to construct, from any given second-order random linear functional $s[\cdot]$, a second-order $\mathcal{H}$-valued random variable $\mathbf{s}$ such that, for all $\mathbf{h} \in \mathcal{H}$,

$$(\mathbf{h}, \mathbf{s}) = s[\mathbf{h}] \text{ wp1} .$$

Such an $\mathcal{H}$-valued random variable $\mathbf{s}$ is called a *regularized version* of the random linear functional $s[\cdot]$ (Itô 1984, Definition 2.3.2, p. 23).

Let $s[\cdot]$ be a second-order random linear functional. Given any $\mathbf{h} \in \mathcal{H}$ and positive integer $n$, it follows from the linearity of $s[\cdot]$ that

$$s\left[\sum_{i=1}^{n} (\mathbf{h}_i, \mathbf{h})\mathbf{h}_i\right] = \sum_{i=1}^{n} (\mathbf{h}_i, \mathbf{h}) s[\mathbf{h}_i] \text{ wp1} ,$$

where the set of probability measure zero on which equality does not hold may depend on $\mathbf{h}$ and on the orthonormal basis elements $\{\mathbf{h}_i\}_{i=1}^n$. By the boundedness of $s[\cdot]$ it follows that

$$\mathcal{E}\left(\sum_{i=1}^{n}(\mathbf{h}_i,\mathbf{h})s[\mathbf{h}_i]\right)^2 \leq \gamma^2\left\|\sum_{i=1}^{n}(\mathbf{h}_i,\mathbf{h})\mathbf{h}_i\right\|^2,$$

for some constant $\gamma \in \mathbb{R}$ which is independent of $\mathbf{h}$ and $n$. Taking the limit as $n \to \infty$ gives

$$\mathcal{E}\left(\sum_{i=1}^{\infty}(\mathbf{h}_i,\mathbf{h})s[\mathbf{h}_i]\right)^2 \leq \gamma^2||\mathbf{h}||^2 < \infty,$$

for all $\mathbf{h} \in \mathcal{H}$. Thus the series $\sum_{i=1}^{\infty}(\mathbf{h}_i,\mathbf{h})s[\mathbf{h}_i]$ converges in $L^2(\Omega,\mathcal{F},P)$, i.e., there exists a unique element $\tilde{s}[\mathbf{h}] \in L^2(\Omega,\mathcal{F},P)$ such that

$$\lim_{n\to\infty}\mathcal{E}\left(\tilde{s}[\mathbf{h}]-\sum_{i=1}^{n}(\mathbf{h}_i,\mathbf{h})s[\mathbf{h}_i]\right)^2 = 0,$$

for all $\mathbf{h} \in \mathcal{H}$. Equivalently, since a series converges in a Hilbert space if, and only if, it converges in norm,

$$\sum_{i=1}^{\infty}\left(\mathcal{E}\left\{(\mathbf{h}_i,\mathbf{h})s[\mathbf{h}_i]\right\}^2\right)^{1/2} = \sum_{i=1}^{\infty}|(\mathbf{h}_i,\mathbf{h})|\left(\mathcal{E}s^2[\mathbf{h}_i]\right)^{1/2} < \infty,$$

for all $\mathbf{h} \in \mathcal{H}$. By the Riesz representation theorem applied to the Hilbert space of square-summable sequences of real numbers, and since

$$\sum_{i=1}^{\infty}(\mathbf{h}_i,\mathbf{h})^2 = ||\mathbf{h}||^2$$

by Parseval's relation, the series $\sum_{i=1}^{\infty}(\mathbf{h}_i,\mathbf{h})s[\mathbf{h}_i]$ therefore converges in $L^2(\Omega,\mathcal{F},P)$, for all $\mathbf{h} \in \mathcal{H}$, if, and only if, Eq. (71) holds, in which case

$$\sum_{i=1}^{\infty}|(\mathbf{h}_i,\mathbf{h})|\left(\mathcal{E}s^2[\mathbf{h}_i]\right)^{1/2} \leq ||\mathbf{h}||\left[\sum_{i=1}^{\infty}\mathcal{E}s^2[\mathbf{h}_i]\right]^{1/2} < \infty,$$

by the Schwarz inequality. Thus, if $s[\cdot]$ is a second-order random linear functional, then Eq. (71) holds, for every orthonormal basis $\{\mathbf{h}_i\}_{i=1}^{\infty}$ of $\mathcal{H}$.

Conversely, suppose that Eq. (71) holds for a random linear functional $s[\cdot]$, for some orthonormal basis $\{\mathbf{h}_i\}_{i=1}^{\infty}$ of $\mathcal{H}$. Since every $\mathbf{h} \in \mathcal{H}$ has the representation $\mathbf{h} = \sum_{i=1}^{\infty}(\mathbf{h}_i,\mathbf{h})\mathbf{h}_i$, it follows from the linearity of $s[\cdot]$ that if $\mathbf{h} \in \mathcal{H}$ then

$$s[\mathbf{h}] = \sum_{i=1}^{\infty} (\mathbf{h}_i, \mathbf{h}) s[\mathbf{h}_i] \text{ wp1} ,$$

and therefore

$$s^2[\mathbf{h}] \leq ||\mathbf{h}||^2 \sum_{i=1}^{\infty} s^2[\mathbf{h}_i] \text{ wp1} ,$$

by the Schwarz inequality and Parseval's relation. Thus

$$\mathcal{E} s^2[\mathbf{h}] \leq ||\mathbf{h}||^2 \sum_{i=1}^{\infty} \mathcal{E} s^2[\mathbf{h}_i] ,$$

for every $\mathbf{h} \in \mathcal{H}$, i.e., Eq. (70) holds with

$$\gamma^2 = \sum_{i=1}^{\infty} \mathcal{E} s^2[\mathbf{h}_i] < \infty ,$$

by Eq. (71), and therefore $s[\cdot]$ is a second-order random linear functional. Furthermore, since $s[\cdot]$ is a second-order random linear functional, Eq. (71) holds for every orthonormal basis $\{\mathbf{h}_i\}_{i=1}^{\infty}$ of $\mathcal{H}$.

Now let $s[\cdot]$ be a given second-order random linear functional. Since

$$\mathcal{E} \sum_{i=1}^{\infty} s^2[\mathbf{h}_i] = \sum_{i=1}^{\infty} \mathcal{E} s^2[\mathbf{h}_i] < \infty ,$$

the sum $\sum_{i=1}^{\infty} s^2[\mathbf{h}_i]$ must be finite wp1, i.e., if

$$E = \left\{ \omega \in \Omega : \sum_{i=1}^{\infty} s^2[\mathbf{h}_i](\omega) < \infty \right\}$$

then $E \in \mathcal{F}$ and $P(E) = 1$, where the set $E$ may depend on $\{\mathbf{h}_i\}_{i=1}^{\infty}$. Define $\mathbf{s}(\omega)$ for each $\omega \in \Omega$ by

$$\mathbf{s}(\omega) = \begin{cases} \sum_{i=1}^{\infty} \mathbf{h}_i s[\mathbf{h}_i](\omega) & \text{if } \omega \in E \\ 0 & \text{if } \omega \notin E \end{cases} .$$

By Parseval's relation it follows that

$$||\mathbf{s}(\omega)||^2 = \begin{cases} \sum_{i=1}^{\infty} s^2[\mathbf{h}_i](\omega) & \text{if } \omega \in E \\ 0 & \text{if } \omega \notin E \end{cases} ,$$

and therefore $\|\mathbf{s}(\omega)\|^2 < \infty$ for all $\omega \in \Omega$. Thus $\mathbf{s}$ is a map from $\Omega$ into $\mathcal{H}$, and for any $\mathbf{h} \in \mathcal{H}$,

$$(\mathbf{h}, \mathbf{s}(\omega)) = \sum_{i=1}^{\infty} (\mathbf{h}_i, \mathbf{h}) s[\mathbf{h}_i](\omega) \tag{72}$$

for each $\omega \in E$. Now, if $\mathbf{h} \in \mathcal{H}$ then

$$s[\mathbf{h}] = \sum_{i=1}^{\infty} (\mathbf{h}_i, \mathbf{h}) s[\mathbf{h}_i] \,\mathrm{wp1}\,,$$

and so there is a set $E_{\mathbf{h}} \in \mathcal{F}$ with $P(E_{\mathbf{h}}) = 1$, that may depend on $\{\mathbf{h}_i\}_{i=1}^{\infty}$ as well as on $\mathbf{h}$, such that

$$s[\mathbf{h}](\omega) = \sum_{i=1}^{\infty} (\mathbf{h}_i, \mathbf{h}) s[\mathbf{h}_i](\omega)$$

for each $\omega \in E_{\mathbf{h}}$. Therefore, for all $\mathbf{h} \in \mathcal{H}$,

$$(\mathbf{h}, \mathbf{s}(\omega)) = s[\mathbf{h}](\omega)$$

for each $\omega \in E \cap E_{\mathbf{h}}$, and $P(E \cap E_{\mathbf{h}}) = 1$. Since the probability space $(\Omega, \mathcal{F}, P)$ was assumed to be complete, and since $s[\mathbf{h}]$ is a scalar random variable for each $\mathbf{h} \in \mathcal{H}$, it follows that (**h**,**s**) is a scalar random variable for each $\mathbf{h} \in \mathcal{H}$. Therefore the map $\mathbf{s} : \Omega \to \mathcal{H}$ is an $\mathcal{H}$-valued random variable. Since

$$\mathcal{E}||\mathbf{s}||^2 = \mathcal{E} \sum_{i=1}^{\infty} s^2[\mathbf{h}_i] = \sum_{i=1}^{\infty} \mathcal{E} s^2[\mathbf{h}_i] < \infty\,,$$

**s** is a second-order $\mathcal{H}$-valued random variable. Since $s[\cdot]$ is bounded as a map from $\mathcal{H}$ into $L^2(\Omega, \mathcal{F}, P)$,

$$\lim_{n \to \infty} \mathcal{E} \left( s[\mathbf{h}] - \sum_{i=1}^{n} (\mathbf{h}_i, \mathbf{h}) s[\mathbf{h}_i] \right)^2 = 0$$

for all $\mathbf{h} \in \mathcal{H}$, and since Eq. (72) holds for all $\mathbf{h} \in \mathcal{H}$ and $\omega \in E$, it follows that

$$\mathcal{E}\,(s[\mathbf{h}] - (\mathbf{h}, \mathbf{s}))^2 = 0$$

for all $\mathbf{h} \in \mathcal{H}$. Therefore, for all $\mathbf{h} \in \mathcal{H}$, $(\mathbf{h}, \mathbf{s}) = s[\mathbf{h}]$wp1.

## Appendix 2: The Hilbert Spaces $\boldsymbol{\Phi}_p$

Let $\mathcal{H}$ be a real, separable Hilbert space, with inner product and corresponding norm denoted by $(\cdot,\cdot)$ and $\|\cdot\|$, respectively. Denote by $\mathcal{B}(\mathcal{H})$ the Borel field generated by the open sets in $\mathcal{H}$. For convenience it will be assumed in this appendix that $\mathcal{H}$ is infinite-dimensional.

Appendix 2a uses a self-adjoint linear operator on $\mathcal{H}$ to construct a special family of Hilbert spaces $\{\Phi_p, p \geq 0\}$. The inner product and corresponding norm on $\Phi_p$ are denoted by $(\cdot,\cdot)_p$ and $\|\cdot\|_p$, respectively, for each $p \geq 0$. These Hilbert spaces have the following properties: (i) $\Phi_0 = \mathcal{H}$; (ii) for each $p > 0$, $\Phi_p \subset \mathcal{H}$, and therefore $\Phi_p$ is real and separable; (iii) for each $p > 0$, $\Phi_p$ is dense in $\mathcal{H}$, and therefore $\Phi_p$ is infinite-dimensional; and (iv) if $0 \leq q \leq r$, then $\|\mathbf{h}\| = \|\mathbf{h}\|_0 \leq \|\mathbf{h}\|_q \leq \|\mathbf{h}\|_r$ for all $\mathbf{h} \in \Phi_r$, and therefore $\mathcal{H} = \Phi_0 \supset \Phi_q \supset \Phi_r$. In view of property (iv), the family $\{\Phi_p, p \geq 0\}$ is called a *decreasing family* of Hilbert spaces. The construction given here follows closely that of Kallianpur and Xiong (1995, Example 1.3.2, pp. 40–42). For various concrete examples and classical applications of decreasing families of Hilbert spaces constructed in this way, see Reed and Simon (1972, pp. 141–145), Itô (1984, pp. 1–12), Kallianpur and Xiong (1995, pp. 29–40), and Lax (2006, pp. 61–67).

Appendix 2b discusses the spaces $\Phi_p$ in case $\mathcal{H} = L^2(S)$, the space of square-integrable vector or scalar fields on the sphere $S$, when the operator $\mathbf{L}$ used in the construction of the spaces $\Phi_p$ is taken to be $\mathbf{L} = -\Delta$, where $\Delta$ is the Laplacian operator on the sphere.

### *2a Construction of the Hilbert Spaces $\boldsymbol{\Phi}_p$*

Let $\mathbf{L}$ be a densely defined, positive semidefinite, self-adjoint linear operator on $\mathcal{H}$, and let $\mathbf{I}$ denote the identity operator on $\mathcal{H}$. It follows from elementary arguments (e.g. Riesz and Sz.-Nagy 1955, p. 324) that the inverse operator $(\mathbf{I} + \mathbf{L})^{-1}$ is a *bounded*, positive semidefinite, self-adjoint linear operator defined on *all* of $\mathcal{H}$, in fact with

$$||(\mathbf{I} + \mathbf{L})^{-1}\mathbf{h}|| \leq ||\mathbf{h}||$$

for all $\mathbf{h} \in \mathcal{H}$. Assume that some power $p_1 > 0$ of $(\mathbf{I}+\mathbf{L})^{-1}$ is a compact operator on $\mathcal{H}$. Then it follows from the Hilbert-Schmidt theorem (e.g. Reed and Simon 1972, Theorem VI.16, p. 203) that there exists a countable orthonormal basis for $\mathcal{H}$ which consists of eigenvectors $\{\mathbf{g}_i\}_{i=1}^{\infty}$ of $(\mathbf{I} + \mathbf{L})^{-p_1}$,

$$(\mathbf{I} + \mathbf{L})^{-p_1}\mathbf{g}_i = \mu_i \mathbf{g}_i$$

for $i = 1, 2, \ldots$, where the corresponding eigenvalues $\{\mu_i\}_{i=1}^{\infty}$ satisfy $1 \geq \mu_1 \geq \mu_2 \geq \cdots$, with $\mu_i \to 0$ as $i \to \infty$. Moreover, $\mu_i > 0$ for $i = 1, 2, \ldots$, for suppose otherwise. Then there is a first zero eigenvalue, call it $\mu_{M+1}$, since

the eigenvalues decrease monotonically toward zero. Therefore $(\mathbf{I}+\mathbf{L})^{-p_1}$ has finite rank $M$, hence $\mathbf{I}+\mathbf{L}$ is defined everywhere in $\mathcal{H}$ and also has rank $M$. But $\operatorname{rank}(\mathbf{I}+\mathbf{L}) \geq \operatorname{rank}\mathbf{I} = \infty$ since $\mathbf{L}$ is positive semidefinite and $\mathcal{H}$ was assumed infinite-dimensional, a contradiction.

Now define $\{\lambda_i\}_{i=1}^{\infty}$ by $(1+\lambda_i)^{-p_1} = \mu_i$. Then $0 \leq \lambda_1 \leq \lambda_2 \leq \cdots$, with $\lambda_i \to \infty$ as $i \to \infty$, and $\lambda_i < \infty$ for $i = 1, 2, \ldots$ since $\mu_i > 0$ for $i = 1, 2, \ldots$. Since the function $\lambda(\mu) = \mu^{-1/p_1} - 1$ is measurable and finite for $\mu \in (0, 1]$, it follows from the functional calculus for self-adjoint operators (e.g. Riesz and Sz.-Nagy 1955, pp. 343–346; Reed and Simon 1972, pp. 259–264) that

$$\mathbf{L}\mathbf{g}_i = \lambda_i \mathbf{g}_i$$

for $i = 1, 2, \ldots$, and similarly for all $p \geq 0$ that

$$(\mathbf{I}+\mathbf{L})^p \mathbf{g}_i = (1+\lambda_i)^p \mathbf{g}_i$$

for $i = 1, 2, \ldots$, with $(\mathbf{I}+\mathbf{L})^p$ densely defined and self-adjoint in $\mathcal{H}$ for all $p \geq 0$.

For each $p \geq 0$, denote by $\Phi_p$ the domain of definition of $(\mathbf{I}+\mathbf{L})^p$, i.e.,

$$\Phi_p = \{\mathbf{h} \in \mathcal{H} : ||(\mathbf{I}+\mathbf{L})^p \mathbf{h}|| < \infty\}.$$

In particular, $\Phi_0 = \mathcal{H}$. Now

$$\begin{aligned} ||(\mathbf{I}+\mathbf{L})^p \mathbf{h}||^2 &= \sum_{i=1}^{\infty} ((\mathbf{I}+\mathbf{L})^p \mathbf{h}, \mathbf{g}_i)^2 = \sum_{i=1}^{\infty} (\mathbf{h}, (\mathbf{I}+\mathbf{L})^p \mathbf{g}_i)^2 \\ &= \sum_{i=1}^{\infty} (\mathbf{h}, (1+\lambda_i)^p \mathbf{g}_i)^2 = \sum_{i=1}^{\infty} (1+\lambda_i)^{2p} (\mathbf{h}, \mathbf{g}_i)^2 \end{aligned}$$

for each $p \geq 0$, where the first equality is Parseval's relation and the second one is due to the fact that $(\mathbf{I}+\mathbf{L})^p$ is self-adjoint. Thus for each $p \geq 0$, $\Phi_p$ is given explicitly by

$$\Phi_p = \left\{ \mathbf{h} \in \mathcal{H} : \sum_{i=1}^{\infty} (1+\lambda_i)^{2p} (\mathbf{h}, \mathbf{g}_i)^2 < \infty \right\}.$$

Using this formula, it can be checked that for each $p \geq 0$, $\Phi_p$ is an inner product space, with inner product $(\cdot, \cdot)_p$ defined by

$$(\mathbf{g}, \mathbf{h})_p = \sum_{i=1}^{\infty} (1+\lambda_i)^{2p} (\mathbf{g}, \mathbf{g}_i)(\mathbf{h}, \mathbf{g}_i) = ((\mathbf{I}+\mathbf{L})^p \mathbf{g}, (\mathbf{I}+\mathbf{L})^p \mathbf{h})$$

for all $\mathbf{g}, \mathbf{h} \in \Phi_p$, and corresponding norm $\|\cdot\|_p$ defined by

$$||\mathbf{h}||_p^2 = (\mathbf{h}, \mathbf{h})_p = ||(\mathbf{I} + \mathbf{L})^p \mathbf{h}||^2$$

for all $\mathbf{h} \in \Phi_p$. It follows also that if $0 \leq q \leq r$, then $\|\mathbf{h}\| = \|\mathbf{h}\|_0 \leq \|\mathbf{h}\|_q \leq \|\mathbf{h}\|_r$ for all $\mathbf{h} \in \Phi_r$, and therefore that $\Phi_r \subset \Phi_q \subset \mathcal{H}$.

Each inner product space $\Phi_p$, $p > 0$, is in fact a Hilbert space, i.e., is already complete in the norm $\|\cdot\|_p$. To see this, suppose that $\{\mathbf{h}_n\}_{n=1}^{\infty}$ is a Cauchy sequence in $\Phi_p$ for some fixed $p > 0$, i.e. that $\|\mathbf{h}_n - \mathbf{h}_m\|_p \to 0$ as $n, m \to \infty$. Since $\|\mathbf{h}_n - \mathbf{h}_m\| \leq \|\mathbf{h}_n - \mathbf{h}_m\|_p$ for all $n, m \geq 1$, it follows that $\{\mathbf{h}_n\}_{n=1}^{\infty}$ is also a Cauchy sequence in $\mathcal{H}$, and since $\mathcal{H}$ is complete, the sequence converges to a unique element $\mathbf{h}_\infty \in \mathcal{H}$. It remains to show that in fact $\mathbf{h}_n \to \mathbf{h}_\infty \in \Phi_p$ as $n \to \infty$.

Now

$$||\mathbf{h}_n - \mathbf{h}_m||_p^2 = ||(\mathbf{I} + \mathbf{L})^p (\mathbf{h}_n - \mathbf{h}_m)||^2 = \sum_{i=1}^{\infty} (1 + \lambda_i)^{2p} (\mathbf{h}_n - \mathbf{h}_m, \mathbf{g}_i)^2 \,.$$

Thus, that $\{\mathbf{h}_n\}_{n=1}^{\infty}$ is a Cauchy sequence in $\Phi_p$ means that, given any $\varepsilon > 0$, there exists an $M = M(\varepsilon)$ such that, for all $n, m \geq M$,

$$\sum_{i=1}^{I} (1 + \lambda_i)^{2p} (\mathbf{h}_n - \mathbf{h}_m, \mathbf{g}_i)^2 < \varepsilon$$

for any $I \geq 1$. But for each $i = 1, 2, \ldots,$

$$|(\mathbf{h}_m - \mathbf{h}_\infty, \mathbf{g}_i)| \leq ||\mathbf{h}_m - \mathbf{h}_\infty||\, ||\mathbf{g}_i|| = ||\mathbf{h}_m - \mathbf{h}_\infty|| \to 0 \text{ as } m \to \infty \,,$$

hence $(\mathbf{h}_m, \mathbf{g}_i) \to (\mathbf{h}_\infty, \mathbf{g}_i)$ as $m \to \infty$, and therefore

$$\sum_{i=1}^{I} (1 + \lambda_i)^{2p} (\mathbf{h}_n - \mathbf{h}_\infty, \mathbf{g}_i)^2 < \varepsilon$$

for all $n \geq M$ and $I \geq 1$. Letting $I \to \infty$ then gives

$$||\mathbf{h}_n - \mathbf{h}_\infty||_p^2 = \sum_{i=1}^{\infty} (1 + \lambda_i)^{2p} (\mathbf{h}_n - \mathbf{h}_\infty, \mathbf{g}_i)^2 < \varepsilon$$

for all $n \geq M$, and therefore $\mathbf{h}_n \to \mathbf{h}_\infty \in \Phi_p$ as $n \to \infty$.

Thus, for each $p > 0$, $\Phi_p$ is a Hilbert space, with inner product $(\cdot, \cdot)_p$ and corresponding norm $\|\cdot\|_p$. It can be checked that $\{(1 + \lambda_i)^{-p} \mathbf{g}_i\}_{i=1}^{\infty}$ is an orthonormal basis for $\Phi_p$, for each $p > 0$.[8]

---

[8]It follows that, for any sequence $\{r_n\}_{n=0}^{\infty}$ with $0 \leq r_0 < r_1 < r_2 < \cdots \to \infty$, $\Phi = \cap_{n=0}^{\infty} \Phi_{r_n}$ is a separable Frèchet space, and since the norms $\|\cdot\|_{r_n}$ are Hilbertian seminorms on $\Phi$, also a countably Hilbertian space. If $(\mathbf{I} + \mathbf{L})^{-p_1}$ is not just compact but in fact Hilbert-Schmidt, and if, for

## *2b The Case $\mathcal{H} = L^2(S)$ with $\mathbf{L} = -\Delta$*

Now let $\mathcal{H} = L^2(S)$, the Hilbert space of real, Lebesgue square-integrable scalars on the unit 2-sphere $S$, with inner product

$$(\phi, \psi) = \int_S \phi(\mathbf{x})\psi(\mathbf{x})\, \mathrm{d}\mathbf{x}$$

for all $\phi, \psi \in L^2(S)$, where $\mathbf{x} = (x_1, x_2)$ denotes spherical coordinates on $S$ and $\mathrm{d}\mathbf{x}$ denotes the surface area element, and with corresponding norm $\|\phi\| = (\phi, \phi)^{1/2}$ for all $\phi \in L^2(S)$. Let $\mathbf{L} = -\Delta$, where $\Delta$ is the Laplacian operator on $L^2(S)$. Thus $\mathbf{L}$ is a densely defined, positive semidefinite, self-adjoint linear operator on $L^2(S)$. Denote by $I$ the identity operator on $L^2(S)$.

It will be shown first that for all $p_1 > 1/2$, $(I - \Delta)^{-p_1}$ is a Hilbert-Schmidt operator on $L^2(S)$, hence a compact operator on $L^2(S)$. By Appendix 2a, this allows construction of the decreasing family of Hilbert spaces $\{\Phi_p = \Phi_p(S), p \geq 0\}$,

$$\Phi_p = \{\phi \in L^2(S) : ||(I - \Delta)^p \phi|| < \infty\},$$

with inner product

$$(\phi, \psi)_p = ((I - \Delta)^p \phi, (I - \Delta)^p \psi)$$

for all $\phi, \psi \in \Phi_p$, and corresponding norm $\|\phi\|_p = (\phi, \phi)_p^{1/2}$ for all $\phi \in \Phi_p$. Thus if $\phi \in \Phi_p$ and $p$ is a positive integer or half-integer, then all partial (directional) derivatives of $\phi$ up to order $2p$ are Lebesgue square-integrable.

Second, a Sobolev-type lemma for the sphere will be established, showing that if $\phi \in \Phi_{1/2+q}$ with $q > 0$, then $\phi$ is a bounded function on $S$, with bound

$$\max_{\mathbf{x} \in D} |\phi(\mathbf{x})|^2 < \tfrac{1}{4\pi} \left(1 + \tfrac{1}{2q}\right) \|\phi\|^2_{1/2+q}\,. \tag{73}$$

It follows that if $\phi \in \Phi_{1+q}$ with $q > 0$, then the first partial derivatives of $\phi$ are bounded functions on the sphere, and in particular that $\Phi_{1+q} \subset C^0(S)$, the space of continuous functions on the sphere. It will be shown that, in fact, if $\phi \in \Phi_{1+q}$ with $q > 0$, then $\phi$ is Lipschitz continuous on $S$. Thus, for any $q > 0$ and any non-negative integer $l$, $\Phi_{1+l/2+q} \subset C^l(S)$, the space of functions with $l$ continuous partial derivatives on the sphere, and in fact all of the partial derivatives up to order $l$ of a function $\phi \in \Phi_{1+l/2+q}$ are Lipschitz continuous.

---

instance, $p_n = np_1$, then $\Phi = \cap_{n=0}^{\infty} \Phi_{p_n}$ is a countably Hilbertian nuclear space, and it is possible to define $\Phi'$-valued random variables, where $\Phi'$ is the dual space of $\Phi$. Such random variables are useful for stochastic differential equations in infinite-dimensional spaces (see the books of Itô 1984 and Kallianpur and Xiong 1995), but are not immediately important for the principle of energetic consistency developed in this chapter.

These results carry over to vectors in the usual way. Thus denoting by $L^2(S)$ also the Hilbert space of real, Lebesgue square-integrable $n$-vectors on $S$, the inner product is

$$(\mathbf{g}, \mathbf{h}) = \int_S \mathbf{g}^T(\mathbf{x})\mathbf{h}(\mathbf{x})\,\mathrm{d}\mathbf{x} = \sum_{i=1}^{n}(g_i, h_i)$$

for all $\mathbf{g}, \mathbf{h} \in L^2(S)$, and the corresponding norm is $\|\mathbf{h}\| = (\mathbf{h}, \mathbf{h})^{1/2}$ for all $\mathbf{h} \in L^2(S)$. Thus for $n$-vectors on $S$, the Hilbert spaces $\Phi_p$, $p \geq 0$, are defined by

$$\Phi_p = \{\mathbf{h} \in L^2(S) : ||(I - \Delta)^p \mathbf{h}|| < \infty\},$$

with inner product

$$(\mathbf{g}, \mathbf{h})_p = ((I - \Delta)^p \mathbf{g}, (I - \Delta)^p \mathbf{h}) = \sum_{i=1}^{n}(g_i, h_i)_p$$

for all $\mathbf{g}, \mathbf{h} \in \Phi_p$, and corresponding norm $\|\mathbf{h}\|_p = (\mathbf{h}, \mathbf{h})_p^{1/2}$ for all $\mathbf{h} \in \Phi_p$.

To establish that $(I - \Delta)^{-p}$ is a Hilbert-Schmidt operator on $L^2(S)$ if $p > 1/2$, note first that

$$\sum_{l=0}^{\infty} \frac{2l+1}{[1 + l(l+1)]^{1+2\varepsilon}} < 1 + \frac{1}{2\varepsilon} \tag{74}$$

if $\varepsilon > 0$. To obtain this inequality, let

$$f(x) = \frac{2x+1}{[1 + x(x+1)]^{1+2\varepsilon}}$$

for $x \geq 0$ and $\varepsilon > 0$. Then $f$ is monotone decreasing for $x \geq 1/2$, and $f(0) > f(1)$, and so

$$\sum_{l=0}^{\infty} \frac{2l+1}{[1 + l(l+1)]^{1+2\varepsilon}} = f(0) + \sum_{l=1}^{\infty} f(l) < f(0) + \int_0^{\infty} f(x)\,\mathrm{d}x = 1 + \frac{1}{2\varepsilon}.$$

The sum in Eq. (74) diverges logarithmically for $\varepsilon = 0$.

Now let $C = (I - \Delta)^{-p}$ with $p > 0$. Thus $C$ is a bounded operator from $L^2(S)$ into $L^2(S)$, with $\|C\phi\| \leq \|\phi\|$ for all $\phi \in L^2(S)$. The real and imaginary parts of the spherical harmonics $Y_l^m$ form an orthonormal basis for $L^2(S)$, and

$$\Delta Y_l^m = -l(l+1)Y_l^m$$

for $l \geq 0$ and $|m| \leq l$. Thus

$$CY_l^m = \lambda_l^m Y_l^m ,$$

with eigenvalues $\lambda_l^m = (Y_l^m, CY_l^m) = [1 + l(l+1)]^{-p}$ for $l \geq 0$ and $|m| \leq l$. But

$$\sum_{l=0}^{\infty} \sum_{m=-l}^{l} (\lambda_l^m)^2 = \sum_{l=0}^{\infty} \frac{2l+1}{[1 + l(l+1)]^{2p}} ,$$

and so this sum is finite for $p > 1/2$ by Eq. (74). Hence $C$ is Hilbert-Schmidt for $p > 1/2$.

To establish the bound of Eq. (73), suppose that $\phi \in \Phi_{1/2+q}$ with $q > 0$. Thus $(I - \Delta)^{1/2+q}\phi \in L^2(S)$ and has a spherical harmonic expansion

$$(I - \Delta)^{1/2+q}\phi = \sum_{l=0}^{\infty} \sum_{m=-l}^{l} \beta_l^m Y_l^m ,$$

where the convergence is in $L^2(S)$, with

$$\|\phi\|_{1/2+q}^2 = \|(I - \Delta)^{1/2+q}\phi\|^2 = \sum_{l=0}^{\infty} \sum_{m=-l}^{l} |\beta_l^m|^2 < \infty . \tag{75}$$

Therefore

$$\phi = \sum_{l=0}^{\infty} \sum_{m=-l}^{l} [1 + l(l+1)]^{-1/2-q} \beta_l^m Y_l^m , \tag{76}$$

where the convergence is in $\Phi_{1/2+q}$. It will be shown that this series converges absolutely, hence pointwise, so that

$$\phi(\mathbf{x}) = \sum_{l=0}^{\infty} \sum_{m=-l}^{l} [1 + l(l+1)]^{-1/2-q} \beta_l^m Y_l^m(\mathbf{x})$$

for each $\mathbf{x} \in S$. This will also give Eq. (73).

Now,

$$|\phi| \leq \sum_{l=0}^{\infty} [1 + l(l+1)]^{-1/2-q} \sum_{m=-l}^{l} |\beta_l^m| \, |Y_l^m| ,$$

and so

$$|\phi| \leq \sum_{l=0}^{\infty} [1 + l(l+1)]^{-1/2-q} \left\{ \sum_{m=-l}^{l} |\beta_l^m|^2 \right\}^{1/2} \left\{ \sum_{m=-l}^{l} |Y_l^m|^2 \right\}^{1/2}$$

by the Schwarz inequality. The spherical harmonic addition theorem says that

$$P_l(\cos\gamma) = \frac{4\pi}{2l+1} \sum_{m=-l}^{l} Y_l^m(\mathbf{x})\overline{Y}_l^m(\mathbf{y})$$

for $l \geq 0$, where $P_l$ is the $l$th Legendre polynomial and $\gamma$ is the angle between $\mathbf{x}$ and $\mathbf{y}$. This implies that

$$\sum_{m=-l}^{l} |Y_l^m(\mathbf{x})|^2 = \frac{2l+1}{4\pi}$$

for all $\mathbf{x} \in S$, and so

$$|\phi| \leq \frac{1}{\sqrt{4\pi}} \sum_{l=0}^{\infty} [2l+1]^{1/2}[1+l(l+1)]^{-1/2-q} \left\{ \sum_{m=-l}^{l} |\beta_l^m|^2 \right\}^{1/2} .$$

Another application of the Schwarz inequality then gives

$$|\phi| \leq \frac{1}{\sqrt{4\pi}} \left\{ \sum_{l=0}^{\infty} [2l+1][1+l(l+1)]^{-1-2q} \right\}^{1/2} \left\{ \sum_{l=0}^{\infty} \sum_{m=-l}^{l} |\beta_l^m|^2 \right\}^{1/2} ,$$

or, using Eq. (75),

$$|\phi|^2 \leq \frac{1}{4\pi} ||\phi||_{1/2+q}^2 \sum_{l=0}^{\infty} \frac{2l+1}{[1+l(l+1)]^{1+2q}} .$$

Therefore, by Eq. (74), the sum in Eq. (76) converges absolutely, and Eq. (73) holds.

Now suppose that $\phi \in \Phi_{1+q}$ with $q > 0$. To establish that $\phi$ is Lipschitz continuous on $S$, note first that by the previous result,

$$\phi(\mathbf{x}) = \sum_{l=0}^{\infty} \sum_{m=-l}^{l} [1+l(l+1)]^{-1-q} \beta_l^m Y_l^m(\mathbf{x})$$

for each $\mathbf{x} \in S$, where

$$\|\phi\|_{1+q}^2 = \|(I-\Delta)^{1+q}\phi\|^2 = \sum_{l=0}^{\infty} \sum_{m=-l}^{l} |\beta_l^m|^2 < \infty . \tag{77}$$

Therefore,

$$|\phi(\mathbf{x}) - \phi(\mathbf{y})| \leq \sum_{l=0}^{\infty} \sum_{m=-l}^{l} [1 + l(l+1)]^{-1-q} |\beta_l^m| \, |Y_l^m(\mathbf{x}) - Y_l^m(\mathbf{y})|$$

for each $\mathbf{x}, \mathbf{y} \in S$, and so by the Schwarz inequality,

$$|\phi(\mathbf{x}) - \phi(\mathbf{y})| \leq \sum_{l=0}^{\infty} [1 + l(l+1)]^{-1-q} \left\{ \sum_{m=-l}^{l} |\beta_l^m|^2 \right\}^{1/2} \left\{ \sum_{m=-l}^{l} |Y_l^m(\mathbf{x}) - Y_l^m(\mathbf{y})|^2 \right\}^{1/2}.$$

By the spherical harmonic addition theorem,

$$\begin{aligned} \sum_{m=-l}^{l} |Y_l^m(\mathbf{x}) - Y_l^m(\mathbf{y})|^2 &= \sum_{m=-l}^{l} \left[ |Y_l^m(\mathbf{x})|^2 - 2\mathrm{Re} Y_l^m(\mathbf{x}) \overline{Y}_l^m(\mathbf{y}) + |Y_l^m(\mathbf{y})|^2 \right] \\ &= \frac{2l+1}{2\pi} \left[ 1 - P_l(\cos\gamma) \right], \end{aligned}$$

where $\gamma = \gamma(\mathbf{x}, \mathbf{y})$ is the angle between $\mathbf{x}$ and $\mathbf{y}$. Therefore,

$$|\phi(\mathbf{x}) - \phi(\mathbf{y})| \leq \sum_{l=0}^{\infty} [1 + l(l+1)]^{-1-q} \left( \frac{2l+1}{2\pi} \right)^{1/2} \left[ 1 - P_l(\cos\gamma) \right]^{1/2} \left\{ \sum_{m=-l}^{l} |\beta_l^m|^2 \right\}^{1/2},$$

and so by Eq. (77) and the Schwarz inequality,

$$|\phi(\mathbf{x}) - \phi(\mathbf{y})| \leq \frac{1}{\sqrt{2\pi}} \left\{ \sum_{l=0}^{\infty} [1 + l(l+1)]^{-2-2q} (2l+1) [1 - P_l(\cos\gamma)] \right\}^{1/2} \|\phi\|_{1+q}.$$

Now, $P_l(1) = 1$, $P_l'(1) = l(l+1)/2$, and $P_l''(1) = [l(l+1) - 2] P_l'(1)/4 \geq 0$ for $l \geq 0$. It follows that for $\gamma$ sufficiently small,

$$1 - P_l(\cos\gamma) \leq (1 - \cos\gamma) P_l'(1) = l(l+1) \sin^2 \tfrac{\gamma}{2},$$

and so

$$|\phi(\mathbf{x}) - \phi(\mathbf{y})| \leq \frac{K}{\sqrt{2\pi}} \|\phi\|_{1+q} \left| \sin \frac{\gamma(\mathbf{x}, \mathbf{y})}{2} \right|, \tag{78}$$

where

$$K^2 = \sum_{l=0}^{\infty} [1 + l(l+1)]^{-2-2q} (2l+1) l(l+1).$$

This series converges for $q > 0$ since the terms decay like $l^{-1-4q}$, and Eq. (78) shows that $\phi$ is Lipschitz continuous.

## Appendix 3: Some Basic Concepts and Definitions

This appendix summarizes background material used elsewhere in this chapter. For further treatment see, for instance, Doob (1953), Royden (1968), and Reed and Simon (1972).

### *3a Measure Spaces*

Let $X$ be a set. A collection $\mathcal{C}$ of subsets of $X$ is called a $\sigma$-*algebra*, or *Borel field*, if (i) the empty set $\emptyset$ is in $\mathcal{C}$, (ii) for every set $A \in \mathcal{C}$, the complement $\tilde{A} = \{x \in X : x \notin A\}$ of $A$ is in $\mathcal{C}$, and (iii) for every countable collection $\{A_i\}_{i=1}^{\infty}$ of sets $A_i \in \mathcal{C}$, the union $\cup_{i=1}^{\infty} A_i$ of the sets is in $\mathcal{C}$. Given any collection $\mathcal{A}$ of subsets of $X$, there is a smallest $\sigma$-algebra which contains $\mathcal{A}$, i.e., there is a $\sigma$-algebra $\mathcal{C}$ such that (i) $\mathcal{A} \subset \mathcal{C}$, and (ii) if $\mathcal{B}$ is a $\sigma$-algebra and $\mathcal{A} \subset \mathcal{B}$ then $\mathcal{C} \subset \mathcal{B}$. The smallest $\sigma$-algebra containing a given collection $\mathcal{A}$ of subsets of $X$ is called the *Borel field of X generated by* $\mathcal{A}$. A *measurable space* is a couple $(X, \mathcal{C})$ consisting of a set $X$ and a $\sigma$-algebra $\mathcal{C}$ of subsets of $X$. If $(X, \mathcal{C})$ is a measurable space and $Y \in \mathcal{C}$, then $(Y, \mathcal{C}_Y)$ is a measurable space, where

$$\mathcal{C}_Y = \{A \in \mathcal{C} : A \subset Y\},$$

i.e., $\mathcal{C}_Y$ consists of all the sets in $\mathcal{C}$ that are subsets of $Y$.

The set $\mathbb{R}^e$ of *extended real numbers* is the union of the set $\mathbb{R}$ of real numbers and the sets $\{\infty\}$ and $\{-\infty\}$. Multiplication of any two extended real numbers is defined as usual, with the convention that $0 \cdot \infty = 0$. Addition and subtraction of any two extended real numbers is also defined, except that $\infty - \infty$ is undefined, as usual.

Let $Y$ and $Z$ be two sets. A function $g$ is called a *map* from $Y$ into $Z$, written $g : Y \to Z$, if $g(y)$ is defined for all $y \in Y$ and $g(y) \in Z$ for all $y \in Y$. Thus a map $g : \mathbb{R} \to \mathbb{R}$ is a real-valued function defined on all of the real line, a map $g : Y \to \mathbb{R}$ is a real-valued function defined on all of $Y$, and a map $g : Y \to \mathbb{R}^e$ is an extended real-valued function defined on all of $Y$.

Let $(X, \mathcal{C})$ be a measurable space. A subset $A$ of $X$ is called *measurable* if $A \in \mathcal{C}$. A map $g : X \to \mathbb{R}^e$ is called *measurable* (with respect to $\mathcal{C}$) if

$$\{x \in X : g(x) \leq \alpha\} \in \mathcal{C},$$

for every $\alpha \in \mathbb{R}$. If $g : X \to \mathbb{R}^e$ is measurable then $|g|$ is measurable, and if $h : X \to \mathbb{R}^e$ is another measurable map then $gh$ is measurable. A *measure* $\mu$ on $(X, \mathcal{C})$ is a map $\mu : \mathcal{C} \to \mathbb{R}^e$ that satisfies (i) $\mu(A) \geq 0$ for every measurable set $A$, (ii) $\mu(\emptyset) = 0$, and (iii)

$$\mu\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} \mu(E_i),$$

for every countable collection $\{E_i\}_{i=1}^{\infty}$ of disjoint measurable sets, i.e., for every countable collection of sets $E_i \in \mathcal{C}$ with $\cap_{i=1}^{\infty} E_i = \emptyset$. A *measure space* $(X, \mathcal{C}, \mu)$ is a measurable space $(X, \mathcal{C})$ together with a measure $\mu$ on $(X, \mathcal{C})$.

Let $(X, \mathcal{C}, \mu)$ be a measure space. A condition $C(x)$ defined for all $x \in X$ is said to hold *almost everywhere* (a.e.) (with respect to $\mu$) if the set $E = \{x \in X : C(x) \text{ is false}\}$ on which it fails to hold is a measurable set of measure zero, i.e., $E \in \mathcal{C}$ and $\mu(E) = 0$. In particular, two maps $g : X \to \mathbb{R}^e$ and $h : X \to \mathbb{R}^e$ are said to be equal almost everywhere, written $g = h$ a.e., if the subset of $X$ on which they are not equal is a measurable set of measure zero.

A measure space $(X, \mathcal{C}, \mu)$ is called *complete* if $\mathcal{C}$ contains all subsets of measurable sets of measure zero, i.e., if $B \in \mathcal{C}$, $\mu(B) = 0$, and $A \subset B$ together imply that $A \in \mathcal{C}$. If $(X, \mathcal{C}, \mu)$ is a complete measure space and $A$ is a subset of a measurable set of measure zero, then $\mu(A) = 0$. If $(X, \mathcal{C}, \mu)$ is a measure space then there is a complete measure space $(X, \mathcal{C}_0, \mu_0)$, called the *completion* of $(X, \mathcal{C}, \mu)$, which is determined uniquely by the conditions that (i) $\mathcal{C} \subset \mathcal{C}_0$, (ii) if $D \in \mathcal{C}$ then $\mu(D) = \mu_0(D)$, and (iii) $D \in \mathcal{C}_0$ if and only if $D = A \cup B$ where $B \in \mathcal{C}$ and $A \subset C \in \mathcal{C}$ with $\mu(C) = 0$. Thus a measure space can always be completed by enlarging its $\sigma$-algebra to include the subsets of measurable sets of measure zero and extending its measure so that the domain of definition of the extended measure includes the enlarged $\sigma$-algebra.

An *open interval* on the real number line $\mathbb{R}$ is a set $(\alpha, \beta) = \{x \in \mathbb{R} : \alpha < x < \beta\}$ with $\alpha, \beta \in \mathbb{R}^e$ and $\alpha < \beta$. Denote by $\mathcal{B}(\mathbb{R})$ the Borel field of $\mathbb{R}$ generated by the open intervals, and denote by $\mathcal{I}(\mathbb{R}) \subset \mathcal{B}(\mathbb{R})$ the sets that are countable unions of disjoint open intervals. For each set $I = \cup_{i=1}^{\infty}(\alpha_i, \beta_i) \in \mathcal{I}(\mathbb{R})$, define

$$m^*(I) = \sum_{i=1}^{\infty} (\beta_i - \alpha_i),$$

and for each set $B \in \mathcal{B}(\mathbb{R})$ define

$$m^*(B) = \inf m^*(I),$$

where the infimum (greatest lower bound) is taken over all those $I \in \mathcal{I}(\mathbb{R})$ such that $B \subset I$. Then $m^*$ is a measure on the measurable space $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. The completion of the measure space $(\mathbb{R}, \mathcal{B}(\mathbb{R}), m^*)$ is denoted by $(\mathbb{R}, \mathcal{M}, m)$. The sets in $\mathcal{M}$ are called the *Lebesgue measurable sets* on $\mathbb{R}$, and $m$ is called *Lebesgue measure* on $\mathbb{R}$.

Let $(X, \mathcal{C}, \mu)$ be a complete measure space, and let $g : X \to \mathbb{R}^e$ and $h : X \to \mathbb{R}^e$ be two maps. If $g$ is measurable and $g = h$ a.e., then $h$ is measurable.

## *3b Integration*

In this subsection let $(X, \mathcal{C}, \mu)$ be a measure space. The *characteristic function* $\chi_A$ of a subset $A$ of $X$ is the map $\chi_A : X \to \{0, 1\}$ defined for each $x \in X$ by

$$\chi_A(x) = \begin{cases} 1 \text{ if } x \in \mathrm{A} \\ 0 \text{ if } x \notin \mathrm{A} \end{cases}.$$

A characteristic function $\chi_A$ is a measurable map if, and only if, $A$ is a measurable set. A map $\phi : X \to \mathbb{R}^e$ is called *simple* if it is measurable and takes on only a finite number of values. Thus the characteristic function of a measurable set is simple, and if $\phi$ is simple and takes on the values $\alpha_1, \ldots, \alpha_n$ then $\phi = \sum_{i=1}^n \alpha_i \chi_{E_i}$, where $E_i = \{x \in X : \phi(x) = \alpha_i\} \in \mathcal{C}$ for $i = 1, \ldots, n$. If $\phi$ is simple and the values $\alpha_1, \ldots, \alpha_n$ it takes on are all non-negative, the integral of $\phi$ over a measurable set $E$ with respect to measure $\mu$ is defined as

$$\int_E \phi \, \mathrm{d}\mu = \sum_{i=1}^n \alpha_i \mu(E_i \cap E),$$

where $E_i = \{x \in X : \phi(x) = \alpha_i\}$ for $i = 1, \ldots, n$. It is possible that $\int_E \phi \, \mathrm{d}\mu = \infty$, for instance if $\alpha_1 \neq 0$ and $\mu(E_1 \cap E) = \infty$, or if $\alpha_1 = \infty$ and $\mu(E_1 \cap E) \neq 0$.

Let $E$ be a measurable set and let $g : X \to \mathbb{R}^e$ be a map which is non-negative, i.e., $g(x) \geq 0$ for all $x \in X$. If $g$ is measurable, the integral of $g$ over $E$ with respect to $\mu$ is defined as

$$\int_E g \, \mathrm{d}\mu = \sup \int_E \phi \, \mathrm{d}\mu,$$

where the supremum (least upper bound) is taken over all simple maps $\phi$ with $0 \leq \phi \leq g$. Function $g$ is called *integrable* (over $E$, with respect to $\mu$) if $g$ is measurable and

$$\int_E g \, \mathrm{d}\mu < \infty.$$

If $\{h_i\}_{i=1}^\infty$ is a collection of non-negative measurable maps from $X$ into $\mathbb{R}^e$, then $h = \sum_{i=1}^\infty h_i$ is a non-negative measurable map from $X$ into $\mathbb{R}^e$ and

$$\int_E h \, \mathrm{d}\mu = \sum_{i=1}^\infty \int_E h_i \, \mathrm{d}\mu,$$

and in particular, $h$ is integrable if and only if $\sum_{i=1}^\infty \int_E h_i \, \mathrm{d}\mu < \infty$.

Let $E$ be a measurable set and let $g : X \to \mathbb{R}^e$ be a map. The *positive part* $g^+$ of $g$ is the non-negative map $g^+ = g \vee 0$, i.e., $g^+(x) = \max\{g(x), 0\}$ for each $x \in X$, and the *negative part* $g^-$ is the non-negative map $g^- = (-g) \vee 0$. Thus $g = g^+ - g^-$ and $|g| = g^+ + g^-$. If $g$ is measurable, so are $g^+$ and $g^-$, as well as $|g|$. Function $g$ is called *integrable* (over $E$, with respect to $\mu$) if both $g^+$ and $g^-$ are integrable, in which case the integral of $g$ is defined as

$$\int_E g \, \mathrm{d}\mu = \int_E g^+ \, \mathrm{d}\mu - \int_E g^- \, \mathrm{d}\mu.$$

Thus $g$ is integrable over $E$ if, and only if, $|g|$ is integrable over $E$, in which case

$$\left| \int_E g \, \mathrm{d}\mu \right| \leq \int_E |g| \, \mathrm{d}\mu < \infty \, .$$

If $g$ is integrable over $X$, then $|g| < \infty$ a.e., $g$ is integrable over $E$, and

$$\int_E |g| \, \mathrm{d}\mu \leq \int_X |g| \, \mathrm{d}\mu < \infty \, .$$

If $g$ is measurable, then

$$\int_X |g| \, \mathrm{d}\mu = 0$$

if, and only if, $g = 0$ a.e.

Let $E$ be a measurable set and let $g : X \to \mathbb{R}^e$ and $h : X \to \mathbb{R}^e$ be two maps. If $g^2$ and $h^2$ are integrable over $E$ then $gh$ is integrable over $E$, and

$$\left| \int_E gh \, \mathrm{d}\mu \right| \leq \int_E |gh| \, \mathrm{d}\mu \leq \left( \int_E g^2 \, \mathrm{d}\mu \right)^{1/2} \left( \int_E h^2 \, \mathrm{d}\mu \right)^{1/2} < \infty \, . \tag{79}$$

If $g$ and $h$ are integrable over $E$ and $g = h$ a.e., then

$$\int_E g \, \mathrm{d}\mu = \int_E h \, \mathrm{d}\mu \, .$$

If the measure space is complete, and if $g$ is integrable over $E$ and $g = h$ a.e., then $h$ is integrable over $E$ and

$$\int_E g \, \mathrm{d}\mu = \int_E h \, \mathrm{d}\mu \, .$$

Now consider the complete measure space $(\mathbb{R}, \mathcal{M}, m)$, where $\mathcal{M}$ is the $\sigma$-algebra of Lebesgue measurable sets on $\mathbb{R}$ and $m$ is Lebesgue measure on $\mathbb{R}$. If $g : \mathbb{R} \to \mathbb{R}^e$ is measurable with respect to $\mathcal{M}$, and is either non-negative or integrable over $\mathbb{R}$ with respect to $m$, the integral of $g$ over a Lebesgue measurable set $E$ is called the *Lebesgue integral* of $g$ over $E$, and is often written as

$$\int_E g \, \mathrm{d}m = \int_E g(x) \, \mathrm{d}x \, .$$

A *Borel measure* on $\mathbb{R}$ is a measure defined on the Lebesgue measurable sets $\mathcal{M}$ that is finite for bounded sets. If $F$ is a monotone increasing function on $\mathbb{R}$ that is continuous on the right, i.e., if $F(\beta) \geq F(\alpha)$ and $\lim_{\beta \to \alpha} F(\beta) = F(\alpha)$ for all $\alpha, \beta \in \mathbb{R}$ with $\alpha < \beta$, then there exists a unique Borel measure $\mu$ on $\mathbb{R}$ such that

$$\mu((\alpha, \beta]) = F(\beta) - F(\alpha)$$

for all $\alpha, \beta \in \mathbb{R}$ with $\alpha < \beta$, where $(\alpha, \beta] = \{x \in \mathbb{R} : \alpha < x \leq \beta\}$. Let $F$ be a monotone increasing function that is continuous on the right, and let $\mu$ be the corresponding Borel measure. If $g : \mathbb{R} \to \mathbb{R}^e$ is measurable with respect to $\mathcal{M}$, and is either non-negative or integrable over $\mathbb{R}$ with respect to the Borel measure $\mu$, the *Lebesgue-Stieltjes integral* of $g$ over a Lebesgue measurable set $E$ is defined as

$$\int_E g(x)\,\mathrm{d}F(x) = \int_E g\,\mathrm{d}\mu\,.$$

## *3c Probability*

A *probability space* is a measure space $(\Omega, \mathcal{F}, P)$ with $P(\Omega) = 1$. The set $\Omega$ is called the *sample space*, the $\sigma$-algebra $\mathcal{F}$ of measurable sets is called the *event space*, a measurable set is called an *event*, and $P$ is called the *probability measure*. For the rest of this subsection, let $(\Omega, \mathcal{F}, P)$ be a probability space.

A measurable map from $\Omega$ into $\mathbb{R}^e$ is called a (scalar) *random variable*. Thus a map $s : \Omega \to \mathbb{R}^e$ is a random variable if, and only if,

$$\{\omega \in \Omega : s(\omega) \leq x\} \in \mathcal{F}$$

for every $x \in \mathbb{R}$. In particular, if $s$ is a random variable then the function

$$F_s(x) = P(\{\omega \in \Omega : s(\omega) \leq x\})\,,$$

called the probability *distribution function* of $s$, is defined for all $x \in \mathbb{R}$. The distribution function of a random variable is monotone increasing and continuous on the right. If the distribution function $F_s$ of a random variable $s$ is an indefinite integral, i.e., if

$$F_s(x) = \int_{-\infty}^{x} f_s(y)\,\mathrm{d}y$$

for all $x \in \mathbb{R}$ and some Lebesgue integrable function $f_s$, then $f_s$ is called the probabilty *density function* of $s$, and $\mathrm{d}F_s/\mathrm{d}x = f_s$ a.e. (with respect to Lebesgue measure) in $\mathbb{R}$.

The *expectation operator* $\mathcal{E}$ is the integration operator over $\Omega$ with respect to probability measure. Thus if $s$ is a random variable then $\mathcal{E}|s|$ is defined, since $|s|$ is a random variable and $|s| \geq 0$, and

$$\mathcal{E}|s| = \int_\Omega |s|\,\mathrm{d}P \leq \infty\,,$$

while a random variable $s$ is integrable over $\Omega$ if, and only if, $\mathcal{E}|s| < \infty$, in which case

$$\mathcal{E}s = \int_\Omega s \, \mathrm{d}P$$

and $|\mathcal{E}s| \leq \mathcal{E}|s| < \infty$. If $s$ is a random variable with $\mathcal{E}|s| < \infty$, then $\overline{s} = \mathcal{E}s$ is called the *mean* of $s$, and the mean can be evaluated equivalently as the Lebesgue-Stieltjes integral

$$\mathcal{E}s = \int_{-\infty}^{\infty} x \, \mathrm{d}F_s(x) \,,$$

where $F_s$ is the distribution function of $s$, hence

$$\mathcal{E}s = \int_{-\infty}^{\infty} x f_s(x) \, \mathrm{d}x$$

if also $s$ has a density function $f_s$, where the integral is the Lebesgue integral.

If $s$ is a random variable then $\mathcal{E}s^2$ is defined, since $s^2 \geq 0$ is a random variable, and either $\mathcal{E}s^2 = \infty$ or $\mathcal{E}s^2 < \infty$. A random variable $s$ is called *second-order* if $\mathcal{E}s^2 < \infty$. If $r$ and $s$ are random variables then $\mathcal{E}|rs|$ is defined since $rs$ is a random variable, and $\mathcal{E}|rs| \leq \infty$. If $r$ and $s$ are second-order random variables, then

$$\mathcal{E}\,|rs| \leq \left(\mathcal{E}r^2\right)^{1/2} \left(\mathcal{E}s^2\right)^{1/2} < \infty$$

by Eq. (79), hence $\mathcal{E}rs$ is defined and $|\mathcal{E}rs| \leq \mathcal{E}\,|rs| < \infty$. In particular, on taking $r = 1$ and using the fact that

$$\mathcal{E}1 = \int_\Omega 1 \, \mathrm{d}P = P(\Omega) = 1 \,,$$

it follows that if $s$ is a second-order random variable then its mean $\overline{s} = \mathcal{E}s$ is defined, with

$$0 \leq |\overline{s}| = |\mathcal{E}s| \leq \mathcal{E}\,|s| \leq \left(\mathcal{E}s^2\right)^{1/2} < \infty \,.$$

The *variance* $\sigma^2 = \mathcal{E}(s - \overline{s})^2$ of a second-order random variable $s$ is therefore also defined, and finite, with

$$0 \leq \sigma^2 = \mathcal{E}\left(s^2 - 2\overline{s}s + \overline{s}^2\right) = \mathcal{E}s^2 - \overline{s}^2 < \infty \,,$$

and

$$\mathcal{E}s^2 = \overline{s}^2 + \sigma^2 \,. \tag{80}$$

A condition $C(\omega)$ defined for all $\omega \in \Omega$ is said to hold *with probability one* (wp1), or *almost surely* (a.s.), if it holds a.e. with respect to probability measure. Thus if $s$ is a random variable, then $s = 0$ wp1 if, and only if, $\mathcal{E}|s| = 0$. If $s$ is a random variable with $\mathcal{E}|s| < \infty$, i.e., if the mean of $s$ is defined, then $|s| < \infty$ wp1. If $r$ and $s$ are two random variables with $\mathcal{E}|r| < \infty$ and $\mathcal{E}|s| < \infty$, and if $r = s$ wp1, then $r$ and $s$ have the same distribution function and, in particular, $\mathcal{E}r = \mathcal{E}s$. If the probability space is complete, and if $s$ is a random variable, $r : \Omega \to \mathbb{R}^e$ and $r = s$ wp1, then $r$ is a random variable and has the same distribution function as $s$, and if, in addition, $\mathcal{E}|s| < \infty$, then $\mathcal{E}|r| < \infty$ and $\mathcal{E}r = \mathcal{E}s$.

## *3d Hilbert Space*

A non-empty set $V$ is called a *linear space* or *vector space* (over the reals) if $\alpha\mathbf{g} + \beta\mathbf{h} \in V$ for all $\mathbf{g}, \mathbf{h} \in V$ and $\alpha, \beta \in \mathbb{R}$. A *norm* on a linear space $V$ is a real-valued function $\|\cdot\|$ such that, for all $\mathbf{g}, \mathbf{h} \in V$ and $\alpha \in \mathbb{R}$, (i) $\|\mathbf{h}\| \geq 0$, (ii) $\|\mathbf{h}\| = 0$ if, and only if, $\mathbf{h} = \mathbf{0}$, (iii) $\|\alpha\mathbf{h}\| = |\alpha|\,\|\mathbf{h}\|$, and (iv) $\|\mathbf{g}+\mathbf{h}\| \leq \|\mathbf{g}\|+\|\mathbf{h}\|$. An *inner product* on a linear space $V$ is a real-valued function $(\cdot,\cdot)$ such that, for all $\mathbf{f}, \mathbf{g}, \mathbf{h} \in V$ and $\alpha \in \mathbb{R}$, (i) $(\mathbf{h}, \mathbf{h}) \geq 0$, (ii) $(\mathbf{h}, \mathbf{h}) = 0$ if, and only if, $\mathbf{h} = \mathbf{0}$, (iii) $(\mathbf{g}, \alpha\mathbf{h}) = \alpha(\mathbf{g}, \mathbf{h})$, (iv) $(\mathbf{f}, \mathbf{g} + \mathbf{h}) = (\mathbf{f}, \mathbf{g}) + (\mathbf{f}, \mathbf{h})$, and (v) $(\mathbf{g}, \mathbf{h}) = (\mathbf{h}, \mathbf{g})$. A *normed linear space* is a linear space equipped with a norm, and an *inner product space* is a linear space equipped with an inner product. Every inner product space $V$ is a normed linear space, with norm $\|\cdot\|$ given by $\|\mathbf{h}\| = (\mathbf{h}, \mathbf{h})^{1/2}$ for all $\mathbf{h} \in V$, where $(\cdot,\cdot)$ is the inner product on $V$. A normed linear space $V$ is an inner product space if, and only if, its norm $\|\cdot\|$ satisfies the *parallelogram law*

$$||\mathbf{g} + \mathbf{h}||^2 + ||\mathbf{g} - \mathbf{h}||^2 = 2(||\mathbf{g}||^2 + ||\mathbf{h}||^2)\,,$$

for all $\mathbf{g}, \mathbf{h} \in V$. On every inner product space $V$, the inner product $(\cdot,\cdot)$ is given by the *polarization identity*

$$(\mathbf{g}, \mathbf{h}) = \tfrac{1}{4}(||\mathbf{g} + \mathbf{h}||^2 - ||\mathbf{g} - \mathbf{h}||^2)\,,$$

for all $\mathbf{g}, \mathbf{h} \in V$, where $\|\cdot\|$ is the norm corresponding to the inner product, i.e., $\|\mathbf{h}\| = (\mathbf{h}, \mathbf{h})^{1/2}$ for all $\mathbf{h} \in V$. The *Schwarz inequality*

$$|(\mathbf{g}, \mathbf{h})| \leq ||\mathbf{g}||\,||\mathbf{h}|| < \infty\,,$$

for all $\mathbf{g}, \mathbf{h} \in V$, holds on every inner product space $V$, where $(\cdot,\cdot)$ is the inner product on $V$ and $\|\cdot\|$ is the corresponding norm.

A subset $O$ of a normed linear space $V$ is called *open* in $V$ if for every $\mathbf{g} \in O$, there exists an $\varepsilon > 0$ such that if $\mathbf{h} \in V$ and $\|\mathbf{g}-\mathbf{h}\| < \varepsilon$ then $\mathbf{h} \in O$. A subset $B$ of a normed linear space $V$ is called *dense* in $V$ if for every $\mathbf{h} \in V$ and $\varepsilon > 0$, there exists an element $\mathbf{g} \in B$ such that $\|\mathbf{g} - \mathbf{h}\| < \varepsilon$. A normed linear space is called *separable* if it has a dense subset that contains countably many elements.

A sequence of elements $\mathbf{h}_1, \mathbf{h}_2, \ldots$ in a normed linear space $V$ is called a *Cauchy sequence* if $\|\mathbf{h}_m - \mathbf{h}_n\| \to 0$ as $m, n \to \infty$. A sequence of elements $\mathbf{h}_1, \mathbf{h}_2, \ldots$ in a normed linear space $V$ is said to *converge* in $V$ if there exists an element $\mathbf{h} \in V$ such that $\|\mathbf{h} - \mathbf{h}_n\| \to 0$ as $n \to \infty$, in which case one writes $\mathbf{h} = \lim_{n\to\infty} \mathbf{h}_n$. A normed linear space $V$ is called *complete* if every Cauchy sequence of elements in $V$ converges in $V$. A complete normed linear space is called a *Banach space*. A Banach space on which the norm is defined by an inner product is called a *Hilbert space*. That is, a Hilbert space is an inner product space which is complete in the norm defined by the inner product.

Let $\mathcal{H}$ be a Hilbert space, with inner product $(\cdot, \cdot)$ and corresponding norm $\|\cdot\|$. A subset $S$ of $\mathcal{H}$ is called an *orthogonal system* if $\mathbf{g} \neq \mathbf{0}$, $\mathbf{h} \neq \mathbf{0}$ and $(\mathbf{g}, \mathbf{h}) = 0$, for every $\mathbf{g}, \mathbf{h} \in \mathcal{H}$. An orthogonal system $S$ is called an *orthogonal basis* (or *complete orthogonal system*) if no other orthogonal system contains $S$ as a proper subset. An orthogonal basis $S$ is called an *orthonormal basis* if $\|\mathbf{h}\| = 1$ for every $\mathbf{h} \in S$. There exists an orthonormal basis which has countably many elements if, and only if, $\mathcal{H}$ is separable. If $\mathcal{H}$ is a separable Hilbert space then every orthonormal basis for $\mathcal{H}$ has the same number of elements $N \leq \infty$, and $N$ is called the *dimension* of $\mathcal{H}$.

Let $\mathcal{H}$ be a separable Hilbert space, with inner product $(\cdot, \cdot)$, corresponding norm $\|\cdot\|$, and orthonormal basis $S = \{\mathbf{h}_i\}_{i=1}^N$, $N \leq \infty$. If $\mathbf{h} \in \mathcal{H}$ then the sequence of partial sums $\sum_{i=1}^n (\mathbf{h}_i, \mathbf{h})\mathbf{h}_i$ converges to $\mathbf{h}$, i.e.,

$$\lim_{n\to N} ||\mathbf{h} - \sum_{i=1}^{n} (\mathbf{h}_i, \mathbf{h})\mathbf{h}_i|| = 0\,,$$

and so every $\mathbf{h} \in \mathcal{H}$ has the representation

$$\mathbf{h} = \sum_{i=1}^{N} (\mathbf{h}_i, \mathbf{h})\mathbf{h}_i\,.$$

Furthermore,

$$(\mathbf{g}, \mathbf{h}) = \sum_{i=1}^{N} (\mathbf{h}_i, \mathbf{g})(\mathbf{h}_i, \mathbf{h})\,,$$

for every $\mathbf{g}, \mathbf{h} \in \mathcal{H}$. Therefore, for every $\mathbf{h} \in \mathcal{H}$,

$$||\mathbf{h}||^2 = \sum_{i=1}^{N} (\mathbf{h}_i, \mathbf{h})^2\,,$$

which is called *Parseval's relation*.

An example of a separable Hilbert space of dimension $N \leq \infty$ is the space $\ell_N^2$ of square-summable sequences of $N$ real numbers, with inner product $(\mathbf{g}, \mathbf{h}) = \sum_{i=1}^N g_i h_i$, where $g_i$ and $h_i$ denote element $i$ of $\mathbf{g} \in \ell_N^2$ and $\mathbf{h} \in \ell_N^2$, respectively. An

orthonormal basis for $\ell_N^2$ is the set of unit vectors $\{\mathbf{e}_j\}_{j=1}^N$, where element $i$ of $\mathbf{e}_j$ is 1 if $i = j$ and 0 if $i \neq j$. In case $N < \infty$, the elements of $\ell_N^2$ are usually written as (column) $N$-vectors $\mathbf{g} = (g_1, \ldots, g_N)^T$, the inner product is then $(\mathbf{g}, \mathbf{h}) = \mathbf{g}^T\mathbf{h}$, and the columns of the $N \times N$ identity matrix constitute an orthonormal basis.

Let $(X, \mathcal{C}, \mu)$ be a measure space. Denote by $\mathcal{L}^1(X, \mathcal{C}, \mu)$ the set of integrable maps from $X$ into $\mathbb{R}^e$, and consider the function $\|\cdot\|$ defined for all $g \in \mathcal{L}^1(X, \mathcal{C}, \mu)$ by

$$\|g\| = \int_X |g| \, \mathrm{d}\mu \, .$$

The set $\mathcal{L}^1(X, \mathcal{C}, \mu)$ is a linear space, and the function $\|\cdot\|$ is by definition real-valued, i.e., $\|g\| < \infty$ for all $g \in \mathcal{L}^1(X, \mathcal{C}, \mu)$. The function $\|\cdot\|$ also satisfies all of the properties of a norm, except that $\|g\| = 0$ does not imply $g = 0$. However, $\|g\| = 0$ does imply that $g = 0$ a.e., and $g = 0$ a.e. implies that $\|g\| = 0$, for all $g \in \mathcal{L}^1(X, \mathcal{C}, \mu)$. Two maps $g$ and $h$ from $X$ into $\mathbb{R}^e$ are called *equivalent*, or are said to belong to the same *equivalence class*, if $g = h$ a.e. If $g$ and $h$ are equivalent, and if $g, h \in \mathcal{L}^1(X, \mathcal{C}, \mu)$, then $\|g\| = \|h\|$. That is, $\|\cdot\|$ assigns the same real number to each member of a given equivalence class of elements of $\mathcal{L}^1(X, \mathcal{C}, \mu)$, and thereby the domain of definition of the function $\|\cdot\|$ is extended from the elements of $\mathcal{L}^1(X, \mathcal{C}, \mu)$ to the equivalence classes of elements of $\mathcal{L}^1(X, \mathcal{C}, \mu)$. The set $L^1(X, \mathcal{C}, \mu)$ of equivalence classes of elements of $\mathcal{L}^1(X, \mathcal{C}, \mu)$ is a linear space, and $\|\cdot\|$ is a norm on this space. The Riesz-Fischer theorem states that $L^1(X, \mathcal{C}, \mu)$ is complete in this norm, i.e., that $L^1(X, \mathcal{C}, \mu)$ is a Banach space under the norm $\|\cdot\|$. The elements of $L^1(X, \mathcal{C}, \mu)$, unlike those of $\mathcal{L}^1(X, \mathcal{C}, \mu)$, are not defined pointwise in $X$, and therefore are not maps.

Denote by $\mathcal{L}^2(X, \mathcal{C}, \mu)$ the set of square-integrable maps from $X$ into $\mathbb{R}^e$, and consider the function $\|\cdot\|$ defined for all $g \in \mathcal{L}^2(X, \mathcal{C}, \mu)$ by

$$\|g\| = \left( \int_X g^2 \, \mathrm{d}\mu \right)^{1/2} \, .$$

Again, the function $\|\cdot\|$ assigns the same real number to each member of any given equivalence class of elements of $\mathcal{L}^2(X, \mathcal{C}, \mu)$, i.e., to each $g, h \in \mathcal{L}^2(X, \mathcal{C}, \mu)$ such that $g = h$ a.e., and in particular, $\|g\| = 0$ if and only if $g = 0$ a.e. Thus the domain of definition of the function $\|\cdot\|$ can be extended to the equivalence classes. The set $L^2(X, \mathcal{C}, \mu)$ of equivalence classes of elements of $\mathcal{L}^2(X, \mathcal{C}, \mu)$ is a linear space, $\|\cdot\|$ is a norm on this space, and $L^2(X, \mathcal{C}, \mu)$ is complete in this norm. Therefore $L^2(X, \mathcal{C}, \mu)$ is a Banach space under the norm $\|\cdot\|$. Moreover, this norm satisfies the parallelogram law, and therefore $L^2(X, \mathcal{C}, \mu)$ is a Hilbert space. The polarization identity yields the inner product $(\cdot, \cdot)$ on $L^2(X, \mathcal{C}, \mu)$, viz.,

$$(g, h) = \int_X gh \, \mathrm{d}\mu \, ,$$

for all $g, h \in L^2(X, \mathcal{C}, \mu)$. Again, the elements of $L^2(X, \mathcal{C}, \mu)$ are not defined pointwise and are not maps. The Schwarz inequality holds on $L^2(X, \mathcal{C}, \mu)$ since $L^2(X, \mathcal{C}, \mu)$ is an inner product space, and gives Eq. (79) when restricted to the elements of $\mathcal{L}^2(X, \mathcal{C}, \mu)$.

Let $V_1$ and $V_2$ be two normed linear spaces, with inner products $\|\cdot\|_1$ and $\|\cdot\|_2$, respectively, and let $\mathcal{H}$ be a Hilbert space, with inner product $(\cdot, \cdot)$. A *bounded linear operator* from $V_1$ into $V_2$ is a map $\mathcal{T} : V_1 \to V_2$ such that (i) $\mathcal{T}(\alpha\mathbf{g} + \beta\mathbf{h}) = \alpha\mathcal{T}\mathbf{g} + \beta\mathcal{T}\mathbf{h}$ for all $\mathbf{g}, \mathbf{h} \in V_1$ and $\alpha, \beta \in \mathbb{R}$, and (ii) there exists a constant $\gamma \in \mathbb{R}$ such that $\|\mathcal{T}\mathbf{h}\|_2 \leq \gamma\|\mathbf{h}\|_1$ for all $\mathbf{h} \in V_1$. A bounded linear operator $\mathcal{T} : \mathcal{H} \to \mathcal{H}$ is called *self-adjoint* if $(\mathcal{T}\mathbf{g}, \mathbf{h}) = (\mathbf{g}, \mathcal{T}\mathbf{h})$ for all $\mathbf{g}, \mathbf{h} \in \mathcal{H}$, and is called *positive semidefinite* if $(\mathbf{h}, \mathcal{T}\mathbf{h}) \geq 0$ for all $\mathbf{h} \in \mathcal{H}$.

At the beginning of this subsection, the field of scalars for linear spaces $V$ was taken to be the real numbers, and inner products were therefore defined to be real-valued. Thus the Hilbert spaces defined here are real Hilbert spaces. It is also possible, of course, to define complex Hilbert spaces. One property that is lost by restricting attention in this chapter to real Hilbert spaces is that, while every positive semidefinite operator on a complex Hilbert space is self-adjoint, a positive semidefinite operator on a real Hilbert space need not be self-adjoint (e.g. Reed and Simon 1972, p. 195). Covariance operators on a real Hilbert space are necessarily self-adjoint as well as positive semidefinite, however, as discussed in Appendix 1c.

## References


Anderson, J.L. and S.L. Anderson, 1999. A Monte Carlo implementation of the nonlinear filtering problem to produce ensemble assimilations and forecasts. *Mon. Weather Rev.*, **127**, 2741–2758.

Coddington, E.A., and N. Levinson, 1955. *Theory of Ordinary Differential Equations*, McGraw-Hill, NewYork.

Cohn, S.E., 1993. Dynamics of short-term univariate forecast error covariances. *Mon. Weather Rev.*, **121**, 3123–3149.

Cohn, S.E., 1997. An introduction to estimation theory. *J. Meteor. Soc. Jpn.,* **75**, 257–288.

Cohn, S.E., 2009. Energetic consistency and coupling of the mean and covariance dynamics. In *Handbook of Numerical Analysis*, vol. XIV, P.G. Ciarlet, (ed.), pp. 443–478. *Special Volume: Computational Methods for the Atmosphere and the Oceans*, Temam, R.M. and J.J. Tribbia (guest eds.), Elsevier, Amsterdam.

Courant, R. and D. Hilbert, 1962. *Methods of Mathematical Physics, vol. II: Partial Differential Equations*, Wiley, NewYork.

Cox, H., 1964. On the estimation of state variables and parameters for noisy dynamic systems. *IEEE Trans. Automat. Contr.*, **9**, 5–12.

Dee, D.P., and A.M. da Silva, 2003. The choice of variable for atmospheric moisture analysis. *Mon. Weather Rev.,* **131**, 155–171.

Doob, J.L., 1953. *Stochastic Processes*, Wiley, NewYork.

Epstein, E.S., 1969. Stochastic dynamic prediction. *Tellus*, **21**, 739–759.

Fisher, M., M. Leutbecher, and G.A. Kelly, 2005. On the equivalence between Kalman smoothing and weak-constraint four-dimensional variational data assimilation. *Q. J. R. Meteorol. Soc.*, **131**, 3235–3246.

Fleming, R.J., 1971. On stochastic dynamic prediction I. The energetics of uncertainty and the question of closure. *Mon. Weather Rev.*, **99**, 851–872.

Houtekamer, P.L., and H.L. Mitchell, 2001. A sequential ensemble Kalman filter for atmospheric data assimilation. *Mon. Weather Rev.*, **129**, 123–137.

Houtekamer, P.L., and H.L. Mitchell, 2005. Ensemble Kalman filtering. *Q. J. R. Meteorol. Soc.*, **131**, 3269–3289.

Houtekamer, P.L., and Co-authors, 2005. Atmospheric data assimilation with an ensemble Kalman filter: Results with real observations. *Mon. Weather Rev*., **133**, 604–620.

Itô, K., 1984. Foundations of Stochastic Differential Equations in Infinite Dimensional Spaces. *CBMS-NSF Regional Conference Series in Applied Mathematics*, vol. 47, Society for Industrial and Applied Mathematics, Philadelphia, PA.

Janjić, T. and S.E. Cohn, 2006. Treatment of observation error due to unresolved scales in atmospheric data assimilation. *Mon. Weather Rev.*, **134**, 2900–2915.

Jazwinski, A.H., 1970. *Stochastic Processes and Filtering Theory*, Academic Press, NewYork.

Kallianpur, G., and J. Xiong, 1995. *Stochastic Differential Equations in Infinite Dimensional Spaces*, Lecture Notes-Monograph Series, vol. 26, Institute of Mathematical Statistics, Hayward, CA.

Kasahara, A., 1974. Various vertical coordinate systems used for numerical weather prediction. *Mon. Weather Rev.*, **102**, 509–522.

Kraichnan, R.H., 1961. Dynamics of nonlinear stochastic systems. *J. Math. Phys.*, **2**, 124–148.

Kreiss, H.-O. and J. Lorenz, 1989. *Initial-Boundary Value Problems and the Navier-Stokes Equations*, Academic Press, NewYork.

Lax, P.D., 1973. Hyperbolic Systems of Conservation Laws and the Mathematical Theory of Shock Waves. *CBMS-NSF Regional Conference Series in Applied Mathematics*, vol. 11, Society for Industrial and Applied Mathematics, Philadelphia, PA.

Lax, P.D., 2006. *Hyperbolic Partial Differential Equations*, Courant Lecture Notes in Mathematics, Vol. 14, American Mathematical Society, NewYork.

Lin, S.-J., 2004. A "vertically Lagrangian" finite-volume dynamical core for global models. *Mon. Weather Rev.*, **132**, 2293–2307.

Lin, S.-J., and R.B. Rood, 1997. An explicit flux-form semi-Lagrangian shallow-water model on the sphere. *Q. J. R. Meteorol. Soc.*, **123**, 2477–2498.

Ménard, R., and L.-P. Chang, 2000. Assimilation of stratospheric chemical tracer observations using a Kalman filter. Part II: $\chi^2$-validated results and analysis of variance and correlation dynamics. *Mon. Weather Rev.*, **128**, 2672–2686.

Ménard, R. and Co-authors, 2000. Assimilation of stratospheric chemical tracer observations using a Kalman filter. Part I: Formulation. *Mon. Weather Rev.*, **128**, 2654–2671.

Mitchell, H.L., P.L. Houtekamer and G. Pellerin, 2002. Ensemble size, balance, and model-error representation in an ensemble Kalman filter. *Mon. Weather Rev.*, **130**, 2791–2808.

Omatu, S. and J.H. Seinfeld, 1989. *Distributed Parameter Systems: Theory and Applications*, Oxford University Press, NewYork.

Ott, E. and Co-authors, 2004. A local ensemble Kalman filter for atmospheric data assimilation. *Tellus*, **56A**, 415–428.

Reed, M. and B. Simon, 1972. *Methods of Modern Mathematical Physics, vol. I: Functional Analysis*, Academic Press, NewYork.

Riesz, F. and B. Sz.-Nagy, 1955. *Functional Analysis*, Frederick Ungar, NewYork.

Royden, H.L., 1968. *Real Analysis*, 2nd ed., Macmillan, NewYork.

Rudin, W., 1991. *Functional Analysis*, 2nd ed., McGraw-Hill, NewYork.

Staniforth, A., N. Wood and C. Girard, 2003. Energy and energy-like invariants for deep non-hydrostatic atmospheres. *Q. J. R. Meteorol. Soc.*, **129**, 3495–3499.

Trémolet, Y., 2006. Accounting for an imperfect model in 4D-Var. *Q. J. R. Meteorol. Soc.*, **132**, 2483–2504.

Trémolet, Y., 2007. Model-error estimation in 4D-Var. *Q. J. R. Meteorol. Soc.*, **133**, 1267–1280.

von Storch, H. and F.W. Zwiers, 1999. *Statistical Analysis in Climate Research*, Cambridge University Press, NewYork.

# Evaluation of Assimilation Algorithms

**Olivier Talagrand**

## 1 Introduction

The theory of statistical linear estimation (*Best Linear Unbiased Estimate*, or *BLUE* – the term *Best Linear Unbiased Estimator* is also used), upon which a large number of presently existing assimilation algorithms are based, has been described in chapter *Variational Assimilation* (Talagrand). On the face of Eq. (7) of that chapter (which is the same as Eq. (2) below), determination of the *BLUE* requires the a priori specification of the expectation $\boldsymbol{\mu}$ and the covariance matrix $\mathbf{S}$ of the errors affecting the data. A number of questions naturally arise in this context:

Q1. How is it possible to objectively evaluate the quality of an assimilation algorithm?
Q2. Is it possible to objectively determine the expectation $\boldsymbol{\mu}$ and covariance $\mathbf{S}$, whose explicit specification is, at least apparently, required for determining the *BLUE*?
Q3. Is it possible to objectively verify if an assimilation algorithm is optimal in a given precise sense, for instance in the sense of least error variance?

These questions are discussed in this chapter. Answers, at least partial, are given. It is stressed that any procedure for achieving any of the above goals requires hypotheses that cannot be objectively validated on the basis of the data only. Section 2 summarizes the main elements of the theory of the *BLUE*, as already described in chapter *Variational Assimilation*, and gives additional elements. The three questions above are dealt with in Sects. 3, 4, 5 and 6, with Sect. 5 being more specifically devoted to objective evaluation of internal consistency of an assimilation algorithm. Conclusions and comments are given in Sect. 7.

The notations are the same as in chapter *Variational Assimilation*, with the exception that the notation $\mathcal{E}$ will be used, as will be explained below, for denoting a

O. Talagrand (✉)
Laboratoire de Météorologie Dynamique/CNRS, École Normale Supérieure, Paris, France
e-mail: Talagrand@lmd.ens.fr

W. Lahoz et al. (eds.), *Data Assimilation*, DOI 10.1007/978-3-540-74703-1_8,

particular type of statistical expectation. For any integer $q$, $\mathbf{I}_q$ will denote the unit matrix of order $q$.

## 2 Reminder on Statistical Linear Estimation

An unknown *true state vector* $\mathbf{x}^t$, belonging to *state space* $\mathcal{S}$, with dimension $n$, is to be determined from a known *data vector* $\mathbf{z}$, belonging to *data space* $\mathcal{D}$, with dimension $m$, of the form

$$\mathbf{z} = \boldsymbol{\Gamma}\, \mathbf{x}^t + \boldsymbol{\varepsilon}. \tag{1}$$

In this expression, $\boldsymbol{\Gamma}$ is a known operator from $\mathcal{S}$ into $\mathcal{D}$, called the *data operator*, represented by an $m \times n$-matrix, while $\boldsymbol{\varepsilon}$ is an unknown $m$–dimensional error, assumed to be a realization of a vector random variable in data space. We look for an estimate of $\mathbf{x}^t$ (the "analysis") of the form

$$\mathbf{x}^a = \mathbf{a} + \mathbf{A}\mathbf{z},$$

where the $n$–vector $\mathbf{a}$ and the $n \times m$-matrix $\mathbf{A}$ are to be determined under the following two conditions:

(1) The estimate $\mathbf{x}^a$ is independent of the choice of the origin in state space;
(2) For any component of $\mathbf{x}$, the statistical expectation of the squared estimation error is minimum.

The solution to this problem is

$$\mathbf{x}^a = (\boldsymbol{\Gamma}^T\, \mathbf{S}^{-1}\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}^T\, \mathbf{S}^{-1}(\mathbf{z} - \boldsymbol{\mu}) \tag{2}$$

i.e.,

$$\mathbf{A} = (\boldsymbol{\Gamma}^T\mathbf{S}^{-1}\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}^T\mathbf{S}^{-1} \tag{3a}$$

$$\mathbf{a} = -\mathbf{A}\boldsymbol{\mu}, \tag{3b}$$

where $\boldsymbol{\mu} \equiv \mathcal{E}[\boldsymbol{\varepsilon}]$ and $\mathbf{S} \equiv \mathcal{E}[(\boldsymbol{\varepsilon} - \boldsymbol{\mu})(\boldsymbol{\varepsilon} - \boldsymbol{\mu})^T]$ are, respectively, the expectation and covariance matrix of the data error $\boldsymbol{\varepsilon}$.

The corresponding estimation error $\mathbf{x}^a$–$\mathbf{x}^t$ has zero expectation

$$\mathcal{E}[\mathbf{x}^a - \mathbf{x}^t] = 0, \tag{4}$$

and has covariance matrix

$$\mathbf{P}^a \equiv \mathcal{E}[(\mathbf{x}^a - \mathbf{x}^t)(\mathbf{x}^a - \mathbf{x}^t)^T] = (\boldsymbol{\Gamma}^T\mathbf{S}^{-1}\boldsymbol{\Gamma})^{-1}. \tag{5}$$

The estimate $\mathbf{x}^a$ is the *Best Linear Unbiased Estimate* (*BLUE*) of $\mathbf{x}$ from $\mathbf{z}$. Its explicit determination requires, at least in view of Eq. (2), the explicit specification of both the expectation $\boldsymbol{\mu}$ and the covariance matrix $\mathbf{S}$ of the error $\boldsymbol{\varepsilon}$.

The *BLUE* is unambiguously defined if, and only if, the matrix $\boldsymbol{\Gamma}$ is of rank $n$, i.e., if, and only if, the condition $\boldsymbol{\Gamma}\mathbf{x} = 0$ implies $\mathbf{x} = 0$. This is called the *determinacy* condition. It implies that $m \geq n$. We set $m = n + p$, $p \geq 0$. The determinacy condition depends only on the data matrix $\boldsymbol{\Gamma}$, and says nothing as to the accuracy of the estimate, which depends on the covariance matrix $\mathbf{S}$ (Eq. 5).

The *BLUE* $\mathbf{x}^a$ can also be obtained as the minimizer of the following scalar objective function, defined on state space

$$\mathbf{x} \in \boldsymbol{\mathcal{D}} \rightarrow J(\mathbf{x}) \equiv \frac{1}{2}[\boldsymbol{\Gamma}\mathbf{x} - (\mathbf{z} - \boldsymbol{\mu})]^T \mathbf{S}^{-1}[\boldsymbol{\Gamma}\mathbf{x} - (\mathbf{z} - \boldsymbol{\mu})]. \tag{6}$$

Variational assimilation is based on explicit minimization of objective functions of form given by Eq. (6). The Kalman filter, although being of a totally different algorithmic form, also amounts to minimizing an objective function of form given by Eq. (6).

Equations (2), (5) and (6) are invariant in a change of origin, as well as in any invertible linear change of coordinates, in either state or data space. In particular, the product ${\mathbf{x}_1}^T \mathbf{S}^{-1} \mathbf{x}_2$, being invariant in a linear change of coordinates, defines a proper scalar product for any two vectors $\mathbf{x}_1$ and $\mathbf{x}_2$ in data space. That scalar product is called the *Mahalanobis scalar product* associated with the covariance matrix $\mathbf{S}$. Expression (6) shows that the image $\boldsymbol{\Gamma}\mathbf{x}^a$ of the *BLUE* $\mathbf{x}^a$ through $\boldsymbol{\Gamma}$ is the point in the image space $\boldsymbol{\Gamma}(\boldsymbol{\mathcal{S}})$ that lies closest, in the sense of the $\mathbf{S}$-Mahalanobis norm, to the unbiased data vector $\mathbf{z}$-$\boldsymbol{\mu}$. The *BLUE* is thus seen to be the output of the three following operations:

(1) Remove the bias in the data vector $\mathbf{z}$ by subtracting the mean error $\boldsymbol{\mu}$;
(2) Project the unbiased data vector onto the subspace $\boldsymbol{\Gamma}(\boldsymbol{\mathcal{S}})$ orthogonally with respect to the $\mathbf{S}$-Mahalanobis scalar product;
(3) Take the inverse of the projection through $\boldsymbol{\Gamma}$. The determinacy condition $\text{rank}(\boldsymbol{\Gamma}) = n$ ensures that the inverse is uniquely defined.

It is seen that the component of $\mathbf{z}$ that is $\mathbf{S}$–Mahalanobis orthogonal to $\boldsymbol{\Gamma}(\boldsymbol{\mathcal{S}})$ has no impact on the result of the estimation process. More precisely, project the data space $\boldsymbol{\mathcal{D}}$ onto the subspace $\boldsymbol{\Gamma}(\boldsymbol{\mathcal{S}})$, and the subspace $\perp\boldsymbol{\Gamma}(\boldsymbol{\mathcal{S}})$ that is $\mathbf{S}$–Mahalanobis orthogonal to $\boldsymbol{\Gamma}(\boldsymbol{\mathcal{S}})$, and denote $\mathbf{w}_1$ and $\mathbf{w}_2$ the corresponding respective components of a vector $\mathbf{w}$ in $\boldsymbol{\mathcal{D}}$. The data operator $\boldsymbol{\Gamma}$ now becomes

$$\boldsymbol{\Gamma} = (\boldsymbol{\Gamma}_1, \mathbf{0})^T$$

where $\boldsymbol{\Gamma}_1$ is an $n \times n$ invertible operator from $\boldsymbol{\mathcal{S}}$ onto $\boldsymbol{\Gamma}(\boldsymbol{\mathcal{S}})$. The data vector $\mathbf{z}$ decomposes into

$$\mathbf{z}_1 = \boldsymbol{\Gamma}_1 \mathbf{x} + \boldsymbol{\varepsilon}_1 \tag{7a}$$

$$\mathbf{z}_2 = \boldsymbol{\varepsilon}_2 \tag{7b}$$

It is now seen that the analysed estimate is equal to

$$\mathbf{x}^a = \boldsymbol{\Gamma}_1^{-1}(\mathbf{z}_1 - \boldsymbol{\mu}_1) = \mathbf{x} + \boldsymbol{\Gamma}_1^{-1}(\boldsymbol{\varepsilon}_1 - \boldsymbol{\mu}_1).$$

The determination of $\mathbf{x}^a$ therefore requires the knowledge of only the orthogonal subspace $\perp \boldsymbol{\Gamma}(\boldsymbol{\mathcal{S}})$ and of the component $\boldsymbol{\mu}_1$ of the mean error $\boldsymbol{\mu}$.

As for the covariance matrix $\mathbf{S}$, it decomposes into the block diagonal matrix

$$\mathbf{S} = \mathrm{diag}(\mathbf{S}_1, \mathbf{S}_2),$$

where $\mathbf{S}_1$ and $\mathbf{S}_2$ are symmetric definite positive matrices. It is seen from Eq. (5) that the analysis error covariance matrix is equal to

$$\mathbf{P}^a \equiv \boldsymbol{\Gamma}_1^{-1} \mathbf{S}_1 \boldsymbol{\Gamma}_1^{-T},$$

where ${\boldsymbol{\Gamma}_1}^{-T}$ denotes the transpose of ${\boldsymbol{\Gamma}_1}^{-1}$.

It results from the above that, contrary to what Eqs. (2), (5) or (6) apparently suggest, the *BLUE* $\mathbf{x}^a$ and associated estimation error covariance matrix $\mathbf{P}^a$ do not depend on the full expectation $\boldsymbol{\mu}$ and covariance matrix $\mathbf{S}$ of the error $\boldsymbol{\varepsilon}$. They depend only on the orthogonal subspace $\perp \boldsymbol{\Gamma}(\boldsymbol{\mathcal{S}})$ and of the components $\boldsymbol{\mu}_1$ and $\mathbf{S}_1$ of $\boldsymbol{\mu}$ and $\mathbf{S}$ along $\boldsymbol{\Gamma}(\boldsymbol{\mathcal{S}})$. Both $\mathbf{x}^a$ and $\mathbf{P}^a$ are independent of the components $\boldsymbol{\mu}_2$ and $\mathbf{S}_2$ along $\perp \boldsymbol{\Gamma}(\boldsymbol{\mathcal{S}})$.

We will consider assimilation systems of the general form given by Eq. (2), for given, not necessarily exact, bias $\boldsymbol{\mu}$ and covariance $\mathbf{S}$. Such systems also provide an estimate of the corresponding error covariance matrix, in the form given by Eq. (5). Since the assumed $\boldsymbol{\mu}$ and $\mathbf{S}$ are not necessarily exact, the corresponding estimate is not necessarily the *BLUE*. One major point of this chapter is precisely to discuss the possibility of identifying possible misspecifications in either the expectation or covariance of the data error, and of determining exactly those quantities (see chapters *Error Statistics in Data Assimilation: Estimation and Modelling*, Buehner; *Bias Estimation*, Ménard).

For convenience, as well as for consistency with usual notations, we will asume that the mean error $\boldsymbol{\mu}$ (or more precisely what is thought to be the mean error) has been substracted from the data vector $\mathbf{z}$. That mean error will not therefore appear explicitly in the equations any more. But the possibility exists that it was not correctly specified in the first place.

The matrix $\mathbf{A} = (\boldsymbol{\Gamma}^T \mathbf{S}^{-1} \boldsymbol{\Gamma})^{-1} \boldsymbol{\Gamma}^T \mathbf{S}^{-1}$ is a left-inverse of $\boldsymbol{\Gamma}$. Conversely, any left-inverse $\boldsymbol{\Delta}$ of $\boldsymbol{\Gamma}$ is of the form $(\boldsymbol{\Gamma}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\Gamma})^{-1} \boldsymbol{\Gamma}^T \boldsymbol{\Sigma}^{-1}$, with an appropriately chosen $m \times m$ definite-positive symmetric matrix $\boldsymbol{\Sigma}$. To see that, let us first note that, if the state and data spaces have the same dimension ($m = n$), $\boldsymbol{\Gamma}$, which has rank $n$, is exactly invertible, with inverse ${\boldsymbol{\Gamma}}^{-1}$. $(\boldsymbol{\Gamma}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\Gamma})^{-1} \boldsymbol{\Gamma}^T \boldsymbol{\Sigma}^{-1}$ is then equal to ${\boldsymbol{\Gamma}}^{-1}$ for

any $\boldsymbol{\Sigma}$. If $m > n$, $\boldsymbol{\Delta}$ has a null-space Ker($\boldsymbol{\Delta}$) with dimension $p = m - n$. In a way similar to what has been done above, project the data space onto the image space $\boldsymbol{\Gamma}(\boldsymbol{\mathcal{S}})$ and the kernel Ker($\boldsymbol{\Delta}$). Any definite-positive matrix $\boldsymbol{\Sigma}$ that decomposes in that projection into

$$\boldsymbol{\Sigma} = \mathrm{diag}(\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2),$$

defines an operator $(\boldsymbol{\Gamma}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}^T\boldsymbol{\Sigma}^{-1}$ which, in addition to being a left-inverse of $\boldsymbol{\Gamma}$, has null space Ker($\boldsymbol{\Delta}$). That operator is therefore identical with $\boldsymbol{\Delta}$. It is seen that $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ can be arbitrary. In particular, if multiplication by $\boldsymbol{\Delta}$ is used for obtaining the *BLUE* of $\mathbf{x}$, the corresponding estimation error covariance matrix will be equal to

$$\mathbf{P}^a = \boldsymbol{\Gamma}_1^{-1}\boldsymbol{\Sigma}_1\boldsymbol{\Gamma}_1^{-T}.$$

Since $\boldsymbol{\Sigma}_1$ can be arbitrary, so can $\mathbf{P}^a$. The knowledge of the left inverse operator that defines the *BLUE* does not bring any information on the associated estimation error. Contrary to what one might be tempted to think, the knowledge of the matrix $(\boldsymbol{\Gamma}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}^T\boldsymbol{\Sigma}^{-1}$ does not bring, even for known $\boldsymbol{\Gamma}$, any information on the matrix $(\boldsymbol{\Gamma}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Gamma})^{-1}$. Any left-inverse of $\boldsymbol{\Gamma}$ can coexist with any estimation error covariance matrix $\mathbf{P}^a$.

We will, therefore, consider estimation schemes of the form

$$\mathbf{x}^e = \mathbf{A}^e\mathbf{z}, \tag{8}$$

where $\mathbf{A}^e$ is a left-inverse of $\boldsymbol{\Gamma}$ (the superscript $e$ stresses the fact that the estimate $\mathbf{x}^e$ may not be optimal). The scheme will be associated with an estimated error covariance matrix $\mathbf{P}^e$. As mentioned above, one particular purpose of this chapter is to determine whether the possible optimality of the scheme given by Eq. (8) can be established on objective grounds. In agreement with the fact that there exists no link in the optimal case between the quantities $\mathbf{A}$ and $\mathbf{P}^a$, no link will be assumed here between $\mathbf{A}^e$ and $\mathbf{P}^e$.

As discussed in the chapter *Variational Assimilation* (Talagrand) it is always possible, when the determinacy condition is verified, to transform the data vector $\mathbf{z}$, through linear invertible operations, into two components of the form

$$\mathbf{x}^b = \mathbf{x}^t + \boldsymbol{\varepsilon}^b \tag{9a}$$

$$\mathbf{y} = \mathbf{H}\mathbf{x}^t + \boldsymbol{\varepsilon}^o. \tag{9b}$$

The vector $\mathbf{x}^b$, which has dimension $n$, is an explicit estimate of the unknown state vector $\mathbf{x}$, called the *background estimate* of $\mathbf{x}$. The vector $\mathbf{y}$, which has dimension $p$, is an additional set of data, linked to the real state vector $\mathbf{x}$ through the (linear) observation operator $\mathbf{H}$, represented by a $p \times n$-matrix.

In the format of Eq. (9), the data operator is $\mathbf{\Gamma} = (\mathbf{I}_n, \mathbf{H}^T)^T$. It is in the format of Eq. (9) that data are usually available in meteorological and oceanographical applications. The expressions given by Eqs. (2), (3) and (5) for the *BLUE* $\mathbf{x}^a$ and the estimation error covariance matrix $\mathbf{P}^a$ then assume the form

$$\mathbf{x}^a = \mathbf{x}^b - \mathcal{E}[\boldsymbol{\varepsilon}^b \mathbf{d}^T]\{\mathcal{E}[\mathbf{d}\mathbf{d}^T]\}^{-1}(\mathbf{y} - \mathbf{H}\mathbf{x}^b), \tag{10a}$$

$$\mathbf{P}^a = \mathcal{E}[\boldsymbol{\varepsilon}^b(\boldsymbol{\varepsilon}^b)^T] - \mathcal{E}[\boldsymbol{\varepsilon}^b \mathbf{d}^T]\{\mathcal{E}[\mathbf{d}\mathbf{d}^T]\}^{-1}\mathcal{E}[\mathbf{d}(\boldsymbol{\varepsilon}^b)^T], \tag{10b}$$

where $\mathbf{d} \equiv \mathbf{y} - \mathbf{H}\mathbf{x}^b$ is the *innovation vector*. Contrary to what was done in the chapter *Variational Assimilation*, we do not assume for the time being that the background and observation errors $\boldsymbol{\varepsilon}^b$ and $\boldsymbol{\varepsilon}^o$ are uncorrelated, so that Eqs. (10) above are more general than Eqs. (15) of chapter *Variational Assimilation*.

In the format of Eq. (7), the left-inverses of $\mathbf{\Gamma} = (\mathbf{I}_n, \mathbf{H}^T)^T$ are of the form

$$(\mathbf{x}^b, \mathbf{y}) \rightarrow \mathbf{x}^e = \mathbf{x}^b + \mathbf{K}(\mathbf{y} - \mathbf{H}\mathbf{x}^b), \tag{11}$$

where the *gain matrix* $\mathbf{K}$ can be any $n \times p$ matrix. A given gain matrix $\mathbf{K}$ can coexist with any estimated error covariance matrix $\mathbf{P}^a$.

Forms given by Eqs. (2) and (11) are exactly equivalent. At any point below, we will use the form that is most convenient for our purpose.

## 3 Objective Evaluation of Assimilation Algorithms

The purpose of assimilation is to estimate as accurately as possible the state of the atmospheric or oceanic flow (see chapters *Numerical Weather Prediction*, Swinbank; *Ocean Data Assimilation*, Haines). The ultimate validation criterion is, therefore, the accuracy with which the flow is estimated, and is naturally quantified by the statistical difference between the estimated values and the corresponding real values of the various physical parameters that define the state of the flow. It is precisely the expectation of the square of that difference that the *BLUE* is intended to minimize. In most situations, the real values of the quantities to be evaluated will, however, not be available, even a posteriori. The validation can, therefore, be performed, at best, against observations or estimates that are themselves affected by errors. Consider the simple case when a scalar quantity $x$ ($n = 1$) is to evaluated from two scalar data ($m = 2$) of the form

$$z_1 = x^t + \varepsilon_1 \tag{12a}$$

$$z_2 = x^t + \varepsilon_2. \tag{12b}$$

This is of form given by Eq. (1), with $\mathbf{z} = (z_1, z_2)^T$, $\mathbf{\Gamma} = (1, 1)^T$, and $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2)^T$. We assume the errors to be unbiased ($\mathcal{E}[\varepsilon_1] = \mathcal{E}[\varepsilon_2] = 0$), mutually uncorrelated ($\mathcal{E}[\varepsilon_1\varepsilon_2] = 0$), and to have the same variance $s$ ($\mathcal{E}[\varepsilon_1^2] = \mathcal{E}[\varepsilon_2^2] = s$), so that $\mathbf{S} = s\mathbf{I}_2$.

The *BLUE* is then $x^a = (1/2)\,(z_1 + z_2)$, and the estimation error covariance matrix $\mathbf{P}^a$ reduces to the scalar $s/2$. Consider a linear estimate $x^e$ of the form

$$x^e = a_1 z_1 + a_2 z_2,$$

with $a_1 + a_2 = 1$ (that is the condition that $\mathbf{A}^e$ is in Eq. (8) a left-inverse of $\mathbf{\Gamma}$). We want to check the possible optimality of $x^e$ by comparison to an observation (or a different estimate) $x^o$ of the form

$$x^o = x^t + \eta,$$

where $\eta$ is a random error. In the logic of least-square statistical minimization taken here, the quantity $x^o$ can legitimately be used for validation if the mean quadratic error $\mathcal{E}[(x^e - x^0)^2]$, considered as a function of the estimate, $x^e$, is minimum when $x^e$ is equal to the *BLUE* $x^a$. This requires that $x^o$ be unbiased (if it had a non-zero bias, $\mathcal{E}[\eta]$, the estimate $x^a + \mathcal{E}[\eta]$ would achieve a better quadratic fit to $x^o$ than the *BLUE* $x^a$). It also requires the error $\eta$ to be uncorrelated with the data error $\boldsymbol{\varepsilon}$. It is obvious that if $\eta$ has variance $s$, but is for instance strongly and positively correlated with $\varepsilon_1$, but not with $\varepsilon_2$, a better fit to $x^o$ will be obtained if $a_1 > a_2$ than if $a_1 = a_2$. More precisely, it can be shown that the linear function of $\mathbf{z}$ which optimally estimates, in the sense of minimum statistical quadratic error, the quantity $x^o$ is equal to

$$x^{oa} = x^a + \mathcal{E}[\eta \boldsymbol{\varepsilon}^T]\mathbf{S}^{-1}(\mathbf{S} - \mathbf{\Gamma}\mathbf{P}^a\mathbf{\Gamma}^T)\mathbf{S}^{-1}\mathbf{z} = x^a + \mathcal{E}[\eta(\varepsilon_1 - \varepsilon_2)]\frac{z_1 - z_2}{2s}.$$

It is different from $x^a$ when the errors $\eta$ and $\boldsymbol{\varepsilon}$ are mutually correlated, with $\mathcal{E}[\eta(\varepsilon_1 - \varepsilon_2)] \neq 0$.

If the conditions of unbiasedness and decorrelation from data error are verified for the validating observation $x^o$, the mean quadratic difference between $x^e$ and $x^o$ is equal to

$$\mathcal{E}[(x^e - x^o)^2] = \mathcal{E}[(x^e - x)^2] + \mathcal{E}[\eta^2],$$

and it is minimum for $x^e = x^a$.

This shows that an estimate $\mathbf{x}^e$ can be usefully validated only against observations (or other estimates) that are unbiased and affected by errors that are themselves uncorrelated with the errors affecting the data used for producing $\mathbf{x}^e$. In particular, the fit of the analysed fields to the data that has been used in the analysis cannot be a proper diagnostic of the quality of the analysis. It can actually be shown that the fit of the analysed fields to any particular piece of data can be made arbitrarily small by simply decreasing the assumed variance for the error affecting that piece of data.

As a consequence, objective comparison between the results of two different assimilation systems can be performed only against observations or estimates that

are uncorrelated with data used in either one of the two systems. A practical difficulty is, of course, that the decorrelation between the data used in the assimilation and in the validation can never be objectively verified, and has to be hypothesized on the basis of, at best, physical knowledge, experience and good judgment.

## 4 Estimation of the Statistics of Data Errors

How is it possible to objectively determine the expectation $\boldsymbol{\mu} = \mathcal{E}[\boldsymbol{\varepsilon}]$ and the covariance matrix $\mathbf{S} = \mathcal{E}[(\boldsymbol{\varepsilon}-\boldsymbol{\mu})(\boldsymbol{\varepsilon}-\boldsymbol{\mu})^T]$? One way could be to proceed by trial and error experimentation, namely to vary $\boldsymbol{\mu}$ and $\mathbf{S}$, and to determine, through comparison against unbiased and independent data, which combination leads to the best statistical fit to those data. One could even envisage that an explicit statistical optimization process could be implemented for determining the optimal values of $\boldsymbol{\mu}$ and $\mathbf{S}$. The sheer numerical dimension of the meteorological or oceanographical problems clearly shows that there can be no hope to entirely and accurately determine $\boldsymbol{\mu}$ and $\mathbf{S}$ through such procedures. But empirical tuning of parameters has always been an important part of the development of assimilation systems. Empirical tuning can be systematized in the form of *cross validation*. A typical example is as follows. For a particular class of instrument, assumed to produce unbiased observations with the same constant error variance, the assumed variance is varied in a series of assimilation experiments in order to determine the value for which the fit to independent data is optimized. In spite of a number of instructive studies of cross validation, and of its extension called *generalized cross validation* (see, e.g., Wahba et al. 1995), this type of method has not been so far extensively used in meteorological or oceanographical applications.

New meteorological observations are available every day, and a natural question is whether it is possible to determine the quantities $\boldsymbol{\mu}$ and $\mathbf{S}$ through appropriate statistical processing of the observations. It is seen from the background-observation decomposition given by Eq. (9) that the only combination of the data that is independent of the unknown state vector $\mathbf{x}$ is the innovation

$$\mathbf{d} = \mathbf{y} - \mathbf{H}\mathbf{x}^b = -\mathbf{H}\boldsymbol{\varepsilon}^b + \boldsymbol{\varepsilon}^o. \tag{13}$$

Within the world of data and assimilation, the innovation is the only objective source of information on the errors affecting the data. The question we consider here is, therefore: Which knowledge on $\boldsymbol{\mu}$ and $\mathbf{S}$ can be obtained from statistical processing of the innovation vector?

Consider the *Data-minus-Analysis* (DmA) difference vector, viz.,

$$\boldsymbol{\delta} \equiv \mathbf{z} - \boldsymbol{\Gamma}\mathbf{x}^e. \tag{14}$$

It is the a posteriori misfit between the raw data and the estimated state vector $\mathbf{x}^e$. By the definition given by Eq. (8) of $\mathbf{x}^e$, $\boldsymbol{\delta}$ is $\boldsymbol{\Sigma}$-Mahalanobis orthogonal to the image subspace $\boldsymbol{\Gamma}(\boldsymbol{\mathcal{S}})$, where $\boldsymbol{\Sigma}$ is any one of the covariance matrices associated with the

left-inverse in Eq. (8). In background-observation format of Eq. (9), it decomposes into $\boldsymbol{\delta} = [(\mathbf{x}^b - \mathbf{x}^e)^T, (\mathbf{y} - \mathbf{H}\mathbf{x}^e)^T]^T$. Now, it is seen from Eq. (13) that

$$\mathbf{x}^b - \mathbf{x}^e = -\mathbf{K}\mathbf{d}$$

$$\mathbf{y} - \mathbf{H}\mathbf{x}^e = (\mathbf{I}_p - \mathbf{H}\mathbf{K})\mathbf{d}.$$

Given $\boldsymbol{\delta}$, these equations are invertible for $\mathbf{d}$, and show that, for any analysis scheme given by Eq. (11), the innovation and DmA vectors $\mathbf{d}$ and $\boldsymbol{\delta}$ are in one-to-one correspondence. As far as accumulating statistics is concerned, that can be done equivalently either a priori on the innovation vector or (after the analysis has been performed) on the DmA difference. Now, the DmA difference $\mathbf{z} - \boldsymbol{\Gamma}\mathbf{x}^e$ is the component of the data vector that has been seen to be "rejected" in the analysis, and to have no impact on the analysis, nor on the estimated analysis error. The conclusion is that no information on the data error that could be useful for the estimation process can be obtained by only statistical processing of the innovation. Prior information obtained from external knowledge of the process that produces the data, and from experience, good judgment or educated guess will always be necessary.

Now, appropriate external information always exists to some extent. To take a simple example, let us assume that (as has actually happened) the innovation corresponding to one type of observation in a particular meteorological station shows a specific statistical feature (a systematic bias, to fix ideas) that is not present in the innovations corresponding to similar observations performed with similar instruments in stations in the same region. It is obvious that the origin of the bias is to be looked for in the observation, and not in the numerical model that produces the innovation. But that conclusion, as obvious as it is, uses external knowledge relative to the observation and prediction system, and could not be obtained from only blind statistical processing of the innovation.

We will discuss at some length in Sect. 7 the implications of the conclusion that has been obtained above. We only mention at this stage that the existence of a one-to-one correspondence between the innovation and the DmA difference could have been inferred without computation by simply noting that both those quantities are obtained by eliminating the unknown $\mathbf{x}$ from the data $\mathbf{z}$. The result of the elimination must be independent of how the elimination is performed.

## 5 Diagnostics of Internal Consistency

The question arises, in view of the conclusion of the previous section, of what, if anything, can be done in terms of objective evaluation of the statistics of the data error. It is clear that, if some parameters of those statistics are known, other parameters can be obtained by differences from the accumulated statistics of the innovation. As an example, Daley (1993) considered the horizontal covariance function of the innovation for radiosonde geopotential observations, which he made homogeneous and isotropic through averaging over geographical location and direction. If the

observational error is spatially uncorrelated, and uncorrelated with the background error, it will appear in the covariance function as an additional Dirac term with correlation distance 0. Extrapolating the observed covariance to distance 0, Daley thus obtained an estimation of the variance of the observational error. Many similar diagnostics studies can be, and have been, performed. They are necessarily based on a priori assumptions as to a number of parameters of the probability distribution of the data errors. They will normally lead to new estimates of other parameters. This estimation process can be iterated.

A systematic approach is as follows. Any assimilation system of form given by Eq. (8) relies on a priori specification of the expectation $\boldsymbol{\mu}$ and the covariance matrix $\mathbf{S}$. These define in turn an expectation and a covariance matrix for the innovation $\mathbf{d}$. If $\boldsymbol{\mu}$ is for instance assumed to be zero, then necessarily $\mathcal{E}[\mathbf{d}] = 0$. In addition, if, as is usually done, the background and observation errors are assumed to be uncorrelated, with respective covariance matrices (see chapter *Variational Assimilation*, Talagrand)

$$\mathbf{P}^b \equiv \mathcal{E}[\boldsymbol{\varepsilon}^b(\boldsymbol{\varepsilon}^b)^T],\ \mathbf{R} \equiv \mathcal{E}[\boldsymbol{\varepsilon}^o(\boldsymbol{\varepsilon}^o)^T], \tag{15}$$

then

$$\mathcal{E}[\mathbf{d}\mathbf{d}^T] = \mathbf{H}\mathbf{P}^b\mathbf{H}^T + \mathbf{R}.$$

Comparison of the a posteriori observed statistics of the innovation with the a priori assumed statistics may reveal inconsistencies, which one may resolve by appropriately redefining the data error expectation and covariance matrix. In view of the one-to-one correspondence between the innovation and the DmA difference $\boldsymbol{\delta}$, the same diagnostics can be done alternatively on the latter. The information will be the same, and the choice is only a matter of convenience. But it must be stressed that, in view of the result proved in the previous section, consistency between the a priori assumed and the a posteriori observed statistics is neither a necessary nor a sufficient condition for optimality of the assimilation process. It is not a sufficient condition because the knowledge of the expectation and covariance of the innovation does not define the covariance matrices $\mathcal{E}[\boldsymbol{\varepsilon}^b\mathbf{d}^T]$ and $\mathcal{E}[\boldsymbol{\varepsilon}^b(\boldsymbol{\varepsilon}^b)^T]$ that are necessary for determining $\mathbf{x}^e$ and $\mathbf{P}^a$ (Eqs. 10). And it is not a necessary condition because a possible inconsistency can always be "explained out" by assuming that it entirely originates in the DmA difference, without modification of the orthogonal space $\perp\boldsymbol{\Gamma}(\boldsymbol{\mathcal{S}})$. As mentioned in the previous section, that will modify neither the estimate $\mathbf{x}^e$, nor the associated estimated estimation error covariance matrix $\mathbf{P}^e$. For a fully explicit example, consider again the case of data of form given by Eq. (12)

$$z_1 = x^t + \varepsilon_1,$$

$$z_2 = x^t + \varepsilon_2.$$

The estimation is performed under the hypothesis that the errors $\varepsilon_1$ and $\varepsilon_2$ are unbiased and mutually uncorrelated, and have same variance $s$. The corresponding estimate is

$$x^e = \frac{1}{2}(z_1 + z_2), \tag{16a}$$

and is expected to be associated with quadratic error

$$s^e = \frac{s}{2}. \tag{16b}$$

As for the innovation, which is here the difference $d = z_1 - z_2$, it is expected to have expectation 0 and variance $2s$. Assume statistics performed on the data show the innovation to have respective expectation and mean square

$$\mathcal{E}_e[d] = m, \tag{17a}$$

$$\mathcal{E}_e[d^2] = m^2 + 2\sigma, \tag{17b}$$

where the subscript $e$ means that $\mathcal{E}_e$ denotes a posteriori observed statistical means. Equations (17) are in contradiction with the hypotheses that have been made on $\varepsilon_1$ and $\varepsilon_2$ if $m \neq 0$ and/or $\sigma \neq s$. The image space $\mathbf{\Gamma}(\boldsymbol{\mathcal{S}})$ is in the present case the direction $z_1 = z_2$, while the space that is **S**–Mahalanobis orthogonal to $\mathbf{\Gamma}(\boldsymbol{\mathcal{S}})$ is the direction $z_1 + z_2 = 0$. Projecting the error vector $\boldsymbol{\varepsilon}$ onto those two directions, and concentrating as mentioned above the inconsistency on the latter, leads for the error components to the expectations

$$\mathcal{E}[\varepsilon_1] = -\mathcal{E}[\varepsilon_2] = -\frac{m}{2}, \tag{18a}$$

and covariance matrix

$$\mathbf{S} = \frac{1}{2}\begin{pmatrix} s+\sigma & s-\sigma \\ s-\sigma & s+\sigma \end{pmatrix}. \tag{18b}$$

It is easily verified that these expressions, while being compatible with Eqs. (17), lead to the estimate of Eq. (16a) and the corresponding estimation error, Eq. (16b). That is absolutely general, and it is always possible to specify the error expectation $\boldsymbol{\mu}$ and covariance matrix **S** so as to make them compatible with any expectation and covariance matrix for the innovation, as well as with any expressions for the *BLUE* $\mathbf{x}^e$ and associated estimation error covariance matrix $\mathbf{P}^e$. That may, on the other hand, require conditions that, in view of the available external knowledge on the data, may be very unlikely, if not impossible. In the above example, accommodation of a bias $m$ requires the biases in $\varepsilon_1$ and $\varepsilon_2$ to be exactly opposite of each other (Eq. 18a), and accommodation of an a posteriori observed variance $\sigma$ that is different from the a priori assumed variance $s$ requires correlation between $\varepsilon_1$ and $\varepsilon_2$

(Eq. 18b). That may be known from other sources to be very implausible, or simply impossible.

The reader may wonder at this stage what would change the analysis. That would be to modify the error covariance matrix $\mathbf{S}$ in such a way that the orthogonal space $\perp\boldsymbol{\Gamma}(\boldsymbol{\mathcal{S}})$ is changed. Keeping $\perp\boldsymbol{\Gamma}(\boldsymbol{\mathcal{S}})$ unchanged, but modifying the component $\mathbf{S}_1$ of $\mathbf{S}$ along $\boldsymbol{\Gamma}(\boldsymbol{\mathcal{S}})$ would not modify the analysis, but would modify the estimation error covariance matrix.

Keeping in mind that reliable interpretation of possible inconsistencies can only come from external knowledge, we describe below a number of diagnostics that can be, and have been, implemented for identifying possible inconsistencies between a priori assumed and a posteriori observed probability distributions of the innovation. Some of those diagnostics are implemented either on the innovation itself, others on the DmA difference, and still others on combinations of both.

A first obvious diagnostic is to test for the possible presence of a bias in either the innovation or the DmA difference. The presence of a statistically significant bias in either one of those two quantities is the signature of an improperly-taken-into-account bias in either the background or the observations (or both). One can argue that systematic check of a presence of a residual bias in either the innovation or the DmA difference is likely the first consistency diagnostic to be performed on an assimilation system. This problem is discussed in more detail in the chapters *Error Statistics in Data Assimilation: Estimation and Modelling* (Buehner) and *Bias Estimation* (Ménard). In these chapters algorithms are presented for evaluating, in particular, possible drifts in observational biases.

A second simple diagnostic bears on the covariance of the DmA difference $\boldsymbol{\delta}$. It is seen from Eqs. (2) and (14) that $\boldsymbol{\delta}$ is equal in a consistent system to

$$\boldsymbol{\delta} = (\mathbf{S} - \boldsymbol{\Gamma}\mathbf{P}^a\boldsymbol{\Gamma}^T)\mathbf{S}^{-1}\boldsymbol{\varepsilon},$$

and has covariance matrix

$$\mathcal{E}[\boldsymbol{\delta}\boldsymbol{\delta}^T] = \mathbf{S} - \boldsymbol{\Gamma}\mathbf{P}^a\boldsymbol{\Gamma}^T. \tag{19}$$

Noting that the second term on the right-hand side of Eq. (19) is the covariance matrix of the vector $\boldsymbol{\Gamma}(\mathbf{x}^a - \mathbf{x}^t)$, this equation can be written as

$$\mathcal{E}[(\mathbf{z}-\boldsymbol{\Gamma}\mathbf{x}^t)(\mathbf{z}-\boldsymbol{\Gamma}\mathbf{x}^t)^T] = \mathcal{E}[(\mathbf{z}-\boldsymbol{\Gamma}\mathbf{x}^a)(\mathbf{z}-\boldsymbol{\Gamma}\mathbf{x}^a)^T] + \mathcal{E}[(\boldsymbol{\Gamma}\mathbf{x}^a-\boldsymbol{\Gamma}\mathbf{x}^t)(\boldsymbol{\Gamma}\mathbf{x}^a-\boldsymbol{\Gamma}\mathbf{x}^t)^T]. \tag{20}$$

The Pythagorean form of this expression shows that the triangle with vertices $\{\mathbf{z}, \boldsymbol{\Gamma}\mathbf{x}^a, \boldsymbol{\Gamma}\mathbf{x}^t\}$ has a right angle (in the sense of orthogonality defined by statistical covariance) at point $\boldsymbol{\Gamma}\mathbf{x}^a$, or equivalently, that the difference $\boldsymbol{\Gamma}(\mathbf{x}^a - \mathbf{x}^t)$ is statistically uncorrelated with the DmA difference, $\mathbf{z} - \boldsymbol{\Gamma}\mathbf{x}^a$.

Equations (19) and (20) also show that the analysed fields must fit the data to within the accuracy assumed on the latter – Hollingsworth and Lönnberg (1989) have called *efficient* an assimilation system that possesses this particular property. This, with the check of unbiasedness of the innovation or DmA difference, is one

basic consistency check to perform on an assimilation system. Experience shows that systems that have been used in operations for a long time, and have been progressively improved through, mostly, comparison with independent observations, are consistent as concerns the particular diagnostic considered here. Newly developed systems, on the other hand, may be off by a factor as large as one order of magnitude. Such an inconsistency is the signature of a gross misspecification in the error covariance matrix **S**, although the check does not say where the misspecification lies, nor even, in all mathematical rigour, that the system is not optimal because of the misspecification.

Consider a sub-matrix of **S** that is diagonal, corresponding, for instance, to radiosonde observations that are at mutually sufficiently large distance for the corresponding representativeness errors to be uncorrelated. The analysis error will be spatially correlated, mainly because of an assumed correlation in the background error. The off-diagonal terms in **S** will be 0, and Eq. (19) then shows that the DmA difference will be negatively correlated at short distances. Hollingsworth and Lönnberg (1989) have described an example of a positive short-distance correlation of the DmA difference in the ECMWF (European Centre for Medium-Range Weather Forecasts) assimilation system. That was the signature of a misspecification somewhere in the matrix **S**. Later checks showed a negative correlation. The sign of the DmA difference spatial correlation does not seem to have been recently checked in operational assimilation systems.

The check defined by Eq. (19) does not, of course, provide a measure of the quality of the assimilation system. On the contrary, assume that, as a result for instance of an increase in the number of observations, the accuracy of the analysis increases, while observation error variances remain constant. The term that is subtracted on the right-hand side of Eq. (19), which is a measure of the quality of the analysis, will decrease. As a consequence, the variance of the DmA difference will increase to tend asymptotically, as it must obviously do in the limit of a perfectly accurate analysis, to the variance of the data error. That constitutes a definitive proof, if one is needed, that the fit of an analysis to the data used in that analysis cannot be a measure of the quality of the analysed fields.

The objective function given by Eq. (6) assumes at its minimum the value

$$J_{\min} \equiv J(\mathbf{x}^a) = \frac{1}{2}(\boldsymbol{\Gamma}\mathbf{x}^a - \mathbf{z})^T \mathbf{S}^{-1}(\boldsymbol{\Gamma}\mathbf{x}^a - \mathbf{z}). \tag{21}$$

It is (half) the squared **S**–Mahalanobis norm of the DmA difference $\boldsymbol{\delta}$. In the $\{\boldsymbol{\Gamma}(\boldsymbol{\mathcal{S}}) - \perp\boldsymbol{\Gamma}(\boldsymbol{\mathcal{S}})\}$ decomposition of the data space $\boldsymbol{\mathcal{D}}$, Eq. (21) reads (see Eq. 7)

$$J_{\min} = \frac{1}{2}\boldsymbol{\varepsilon}_2^T \mathbf{S}^{-1}\boldsymbol{\varepsilon}_2.$$

Since $\boldsymbol{\varepsilon}_2$ is in one-to-one linear correspondence with the innovation **d**, and $\mathbf{S}_2 = \mathcal{E}([\boldsymbol{\varepsilon}_2\boldsymbol{\varepsilon}_2^T]$, invariance of the Mahalanobis scalar product in a linear transformation implies that

$$J_{\min} = \frac{1}{2}\mathbf{d}^T \mathcal{E}[\mathbf{d}\mathbf{d}^T]^{-1}\mathbf{d}. \tag{22}$$

The value of the objective function at its minimum is (half) the squared Mahalanobis norm of the innovation, with respect to its own covariance matrix. This is a deterministic result, valid for any realization of the minimization process. It is readily seen that $J_{\min}$ is also the value of the dual objective function (see Eq. 25 of chapter *Variational Assimilation*, Talagrand) at its minimum.

Equation (22) being valid in any system of coordinates, it is convenient to consider as coordinates the principal components $d_i$ ($i = 1, \ldots, p$) of $\mathbf{d}$, in which the covariance matrix $\mathcal{E}[\mathbf{d}\mathbf{d}^T]$ is the unit matrix $\mathbf{I}_p[\mathcal{E}[d_i\, d_j] = \delta_{ij}$, where $\delta_{ij}$ is the Kronecker symbol]. Equation (22) then reads

$$J_{\min} = \frac{1}{2}\sum_{i=1}^{p} d_i^2, \tag{23}$$

which shows that, for a consistent system, $J_{\min}$ has expectation

$$\mathcal{E}[J_{\min}] = \frac{p}{2}. \tag{24}$$

The expectation of the objective function at its minimum is half the number of observations. This provides a very simple overall check of consistency of an assimilation system. If the observed expectation of $J_{\min}$ is smaller (resp. larger) than $p/2$, this means that the assimilation system is inconsistent, and that the covariance matrix $\mathcal{E}[\mathbf{d}\mathbf{d}^T]$, as specified by the system, is too large (resp. too small). Note that the presence of a residual bias in the innovation (which can of course be directly checked) would lead to an increase of $J_{\min}$.

$J_{\min}$ is a direct output of variational algorithms, both in their primal and dual forms. It can also be computed, although at some numerical cost, in other assimilation algorithms, such as Kalman filtering (see, e.g., Ménard and Chang 2000). The criterion given by Eq. (24) seems to have been first described and used, in the context of oceanography and meteorology, by Bennett (1992). Since then, the test given by Eq. (24) has been performed for a fairly large number of assimilation systems. One can mention, among others, the works of Ménard and Chang (2000), Talagrand and Bouttier (2000), Cañizares et al. (2001), Muccino et al. (2004), Sadiki and Fischer (2005), Chapnik et al. (2006) and Elbern et al. (2007). A remark similar to the one that has been made about Eq. (19) can also be made here. Systems that have gone through extended operational validation and tuning, even if they have never been subject to the particular check given by Eq. (24), usually show a value of $\mathcal{E}[J_{\min}]$ that differs from its theoretical value $p/2$ by a factor of, at most, a few units.

The test given by Eq. (24) is often called the $\chi^2$*–test*. The $\chi^2$ probability distribution of order $p$ is the distribution of the sum of the squares of $p$ independent Gaussian variables, each with expectation 0 and variance 1. It has expectation $p$ and variance $2p$. It is seen from Eq. (23) that, if the data error (and therefore the innovation) is Gaussian, the quantity $2J_{\min}$ follows a $\chi^2$ distribution of order $p$. Both

the expectation and variance of $J_{\min}$ are then equal to $p/2$. But it is also seen from the above that the expectation of $J_{\min}$ is equal to $p/2$ independently of whether the data error is Gaussian or not. However, for large $p$, and even if the innovation is not Gaussian, the central limit theorem (which states that the sum of a large number of independent random variables is Gaussian) ensures that $2J_{\min}$ must approximately follow a $\chi^2$ distribution of order $p$, The distribution of $J_{\min}$, which has expectation $p/2$ and standard deviation $\sqrt{(p/2)}$, is then very strongly peaked. Experience shows that a few realizations of an assimilation system are sufficient for reliable estimation of $\mathcal{E}$ $[J_{\min}]$.

The objective function given by Eq. (6) will most often be the sum of a number of independent terms, viz.,

$$J(\mathbf{x}) = \sum_{k=1}^{K} J_k(\mathbf{x}),$$

where

$$J_k(\mathbf{x}) \equiv \frac{1}{2}(\mathbf{\Gamma}_k\mathbf{x} - \mathbf{z}_k)^T\mathbf{S}_k^{-1}(\mathbf{\Gamma}_k\mathbf{x} - \mathbf{z}_k). \tag{25}$$

In this equation, $\mathbf{z}_k$ is an $m_k$–dimensional component of the data vector $\mathbf{z}$ ($\mathbf{\Sigma}_k m_k = m$), and the rest of the notation is obvious. The inverse estimation error covariance matrix is easily obtained from Eq. (5) as

$$[\mathbf{P}^a]^{-1} = \sum_k \mathbf{\Gamma}_k^T\mathbf{S}_k^{-1}\mathbf{\Gamma}_k. \tag{26}$$

Left-multiplying by $\mathbf{P}^a$, and then taking the trace of the result, yields

$$1 = \frac{1}{n}\sum_k \mathrm{tr}(\mathbf{P}^a\mathbf{\Gamma}_k^T\mathbf{S}_k^{-1}\mathbf{\Gamma}_k) = \frac{1}{n}\sum_k \mathrm{tr}(\mathbf{S}_k^{-1/2}\mathbf{\Gamma}_k\mathbf{P}^a\mathbf{\Gamma}_k^T\mathbf{S}_k^{-1/2})$$

where use has been made, for obtaining the last equality, of the fact that the trace of the product of two matrices is not modified when the order of the factors is reversed. This expression shows that the quantity

$$I(\mathbf{z}_k) \equiv \frac{1}{n}\mathrm{tr}(\mathbf{S}_k^{-1/2}\mathbf{\Gamma}_k\mathbf{P}^a\mathbf{\Gamma}_k^T\mathbf{S}_k^{-1/2}) \tag{27}$$

(which, being the trace of a symmetric definite positive matrix, is necessarily positive) is a measure of the relative contribution of the subset of data $\mathbf{z}_k$ to the overall accuracy of the analysis, or of the (relative) information content of subset $\mathbf{z}_k$. In particular, in case of a background-observation decomposition of form given by Eq. (10) (for the background, $\mathbf{\Gamma}_k = \mathbf{I}_n$, and $\mathbf{S}_k = \mathbf{P}^b$),

$$I(\mathbf{x}^b) = \frac{1}{n}\mathrm{tr}[\mathbf{P}^a(\mathbf{P}^b)^{-1}] = 1 - \frac{1}{n}\mathrm{tr}(\mathbf{KH})$$
$$I(\mathbf{y}) = \frac{1}{n}\mathrm{tr}(\mathbf{KH}) \qquad .$$

Rodgers (2000) calls the quantity $I(\mathbf{z}_k)$ *Degrees of Freedom for Signal*, or *for Noise*, depending on whether the subset $\mathbf{z}_k$ belongs to observations or background. Equation (27) is absolutely general, and is valid for any subset of data, including subsets that may consist of data coming from both the background and the observations. That is clearly seen from the fact that, given any subset $\mathbf{v}$ of the data, the data vector $\mathbf{z}$ can always be transformed, through linear and invertible operations, into $\mathbf{z} \rightarrow (\mathbf{v}^T, \mathbf{w}^T)^T$, where the errors affecting $\mathbf{w}$ are uncorrelated with the errors affecting $\mathbf{v}$. In that transformation, the objective function $J(\mathbf{x})$ becomes the sum of two terms corresponding to $\mathbf{v}$ and $\mathbf{w}$ respectively, from which the information contents $I(\mathbf{v})$ and $I(\mathbf{w})$ are clearly defined.

The relative information content $I(\mathbf{z}_k)$ is in essence the sum of the weights assigned in the assimilation to the components of $\mathbf{z}_k$, normalized in such a way as to allow consistent comparison between data that have been produced by different operators $\mathbf{\Gamma}_k$. Everything else being equal, $I(\mathbf{z}_k)$ increases with decreasing error $\mathbf{S}_k$. Figure 1, extracted from Chapnik et al. (2006), shows the information content

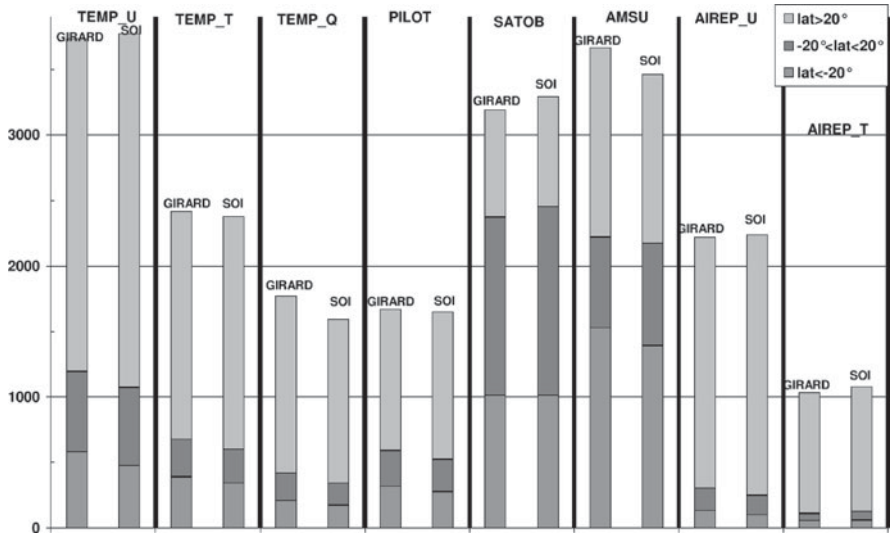



**Fig. 1** Relative information content (Eq. 27) for eight different subsets of observations, as estimated for the variational assimilation system of Météo-France. For each type of observations, the *two bars* correspond to two different algorithms for computing the relative information content (see text). Each *bar* is divided into three parts, corresponding respectively, from *top* to *bottom*, to the Northern Hemisphere (20°N–90°N), the tropical belt (20°S–20°N) and the Southern Hemisphere (20°S–90°S). For TEMP U and AIREP U (wind observations from radiosondes and aircraft respectively), observations of both horizontal components of the wind vector are included (© Royal Meteorological Society)

of eight subsets of observations, as determined for the variational assimilation algorithm of the ARPEGE Numerical Weather Prediction system of Météo-France. Each vertical bar is divided into three parts, corresponding respectively, from top to bottom, to observations performed northward of latitude 20°N, between latitudes 20°N and 20°S, and southward of latitude 20°S (for each type of observations, the two bars correspond, as will be explained below, to two numerical algorithms for the computation of the information content).

It is seen that the largest information content corresponds to observation subsets which contain the largest number of observations: radiosonde wind measurements TEMP U (which contain measurements of both horizontal components of the wind), and satellite observations (SATOB and AMSU). The impact of the geographical distribution of the observations is also clearly visible. The information content of the Northern Hemisphere dominates in the radiosonde (TEMP), pilot (PILOT) and aircraft (AIREP) observations, which are much more numerous in the Northern Hemisphere. For satellite observations, the impact of both hemispheres is the same, with larger relative impact of the tropical belt for SATOB (wind estimates for geostationary satellites) than for AMSU observations (infrared radiation measurements performed from satellites on polar orbits).

Figure 2 shows the same information contents as in Fig. 1, divided now by the number of individual observations in each of the eight subsets. It is the intrinsic information content of individual observations, independent of the number of

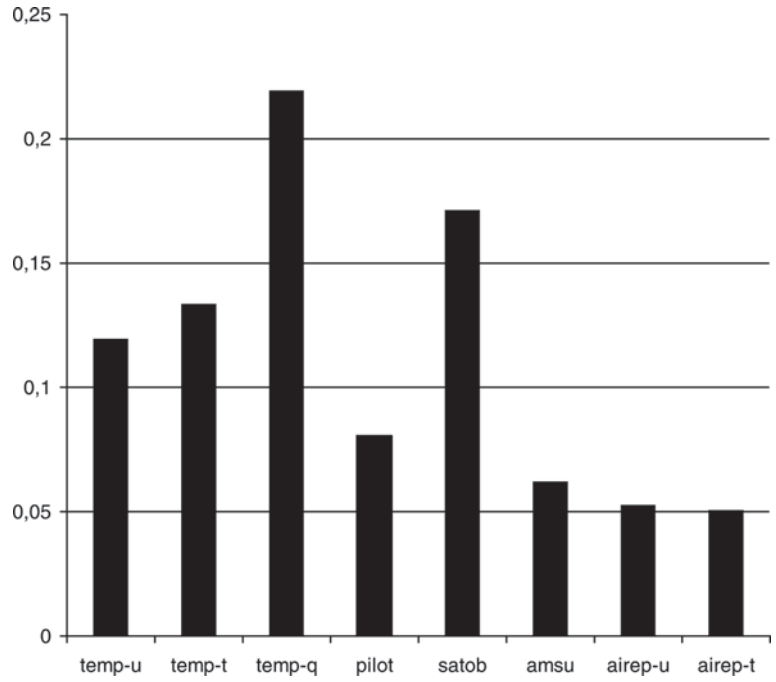



**Fig. 2** Same as Fig. 1, but averaged for individual observations in each class

observations in a given subset. It is seen that it is the radiosonde measurements of humidity that have here highest individual information content.

Different, but fundamentally similar diagnostics have been defined and studied by other authors (see, e.g., Fisher 2003). These types of diagnostics are very useful. Not only do they provide instructive a posteriori information, but they can be used for a priori estimation of information content. They have been used, for instance, by Rabier et al. (2002) for selecting the most informative channels in satellite radiance measurements. More generally, they can be used as part of *Observing System Simulation Experiments* (*OSSEs*) for a priori estimation of the gain that can be expected from new instruments (see chapter *Observing System Simulation Experiments*, Masutani et al.).

On the other hand, these diagnostics are based on the assumption that the expectations and variances of data errors have been correctly specified. That is, of course, not necessarily the case and these diagnostics may, in consequence, be misleading. In particular, they cannot be used in isolation for detecting a possible misspecification of the required expectations and variances.

It can be shown (Talagrand 1999; Desroziers and Ivanov 2001) that the expectation of the term $J_k(\mathbf{x})$ (Eq. 25) at the minimum of the objective function is equal to

$$\mathcal{E}[J(\mathbf{x}^a)] \equiv \frac{1}{2}[m_k - \mathrm{tr}(\mathbf{S}_k^{-\frac{1}{2}} \mathbf{\Gamma}_k \mathbf{P}^a \mathbf{\Gamma}_k^T \mathbf{S}_\mathbf{k}^{-\frac{1}{2}})], \tag{28}$$

where the same trace is present on the right-hand-side as in Eq. (27). Equation (28) includes Eq. (24) as a particular case. It shows that, everything else being equal, $\mathcal{E}\,[J_k(\mathbf{x}^a)]$ will be smaller for more accurate data (smaller norm for $\mathbf{S}_k$). It is obvious that the fit of the analysis must be closer to more accurate data. But Eq. (28) shows that this remains true even when the fit to the data is divided by the covariance matrix of the data error.

Equation (28) also provides the basis for further evaluation of the consistency of an assimilation scheme. It suffices to compare the trace of $\mathbf{S}_k^{-1/2} \mathbf{\Gamma}_k \mathbf{P}^a \mathbf{\Gamma}_k^T \mathbf{S}_k^{-1/2}$, as computed directly and as determined statistically, through Eq. (28), from results of assimilation experiments. Desroziers and Ivanov (2001) have shown that, if the observation error is supposed to be uncorrelated in space and uncorrelated with the background error, Eq. (28) can be used for estimating the observation and background error variances. This is, in essence, a systematic extension of the already mentioned work by Hollingsworth and Lönnberg (1989) and Daley (1993). Along the same lines, Chapnik et al. (2006) have used Eq. (28) to tune the variances of the observational errors in the various channels of the TOVS instrument, carried by the satellites of the NOAA series (*Appendix* lists acronyms.). This has led to a significant change for several of the variances (reduction by a factor of 9 in one case). It has also led to a modest, but distinct, improvement in the quality of the ensuing forecasts.

As a side remark, Eq. (28) also provides what is likely the simplest way of computing the trace $\mathrm{tr}(\mathbf{S}_k^{-1/2} \mathbf{\Gamma}_k \mathbf{P}^a \mathbf{\Gamma}_k^T \mathbf{S}_k^{-1/2})$. The matrix $\mathbf{S}_k^{-1/2} \mathbf{\Gamma}_k \mathbf{P}^a \mathbf{\Gamma}_k^T \mathbf{S}_k^{-1/2}$ has dimension $m_k \times m_k$, where $m_k$ can reach values of order $O(10^6)$. Computing the

trace of large matrices that are not explicitly available (which is the case in assimilation of meteorological or oceanographical observations) raises specific difficulties. A simple way to compute the trace $\mathrm{tr}(\mathbf{S}_k^{-1/2}\boldsymbol{\Gamma}_k\mathbf{P}^a\boldsymbol{\Gamma}_k^T\mathbf{S}_k^{-1/2})$ is to run the assimilation code from unbiased synthetic data affected with errors that have covariance matrix $\mathbf{S}_k$, and to determine the trace from the sample average of $J_k(\mathbf{x}^a)$. Experience shows that, for large values of $m_k$ (and similarly to what has been said above concerning the test given by Eq. 24), a sample of a few elements is sufficient for determination of $\mathcal{E}[J_k(\mathbf{x}^a)]$. It is that particular method that has been used for determining the values marked *SOI* (for *Simulated Optimal Innovation*) in Fig. 1 (the method identified as *Girard* is also a Monte Carlo type method, described in Girard 1987).

We have described a number of diagnostics, in particular diagnostics of internal consistency, that can be implemented on an assimilation system. Many other such diagnostics can be defined, all based on statistics of the innovation or of the DmA differences. Two particular diagnostics have been defined by Desroziers et al. (2005). Assuming the background and observation errors to be uncorrelated, and to have respective covariance matrices $\mathbf{P}^b$ and $\mathbf{R}$ (see Eq. 15), then, in a consistent system

$$\mathcal{E}[\mathbf{H}(\mathbf{x}^a - \mathbf{x}^b)\mathbf{d}^{\mathbf{T}}] = \mathbf{H}\mathbf{P}^b\mathbf{H}^T,$$
$$\mathcal{E}[(\mathbf{y} - \mathbf{H}\mathbf{x}^a)\mathbf{d}^T] = \mathbf{R}.$$

This allows direct comparison with the a priori specified values for $\mathbf{P}^b$ and $\mathbf{R}$ (although of course, an inconsistency in, say, the second of those equations, does not mean that the misspecification lies only in $\mathbf{R}$; actually, it does not even mean that $\mathbf{R}$ is mispecified at all).

It is worth making a few additional remarks concerning the information content given by Eq. (27). As a simple example, consider the case of a scalar $x$ that evolves in time $t$ according to the equation

$$x_{t+1} = \alpha x_t$$

with $\alpha > 0$. Assume two equally accurate observations of $x$ have been performed at times $t$ and $t + 1$. The corresponding information contents are easily seen to be in the proportion $(1/\alpha, \alpha)$. For stable systems ($\alpha > 1$), the later observation is more informative; it is less informative for unstable systems ($\alpha < 1$). The two quantities $x_t$and $x_{t+1}$ being in one-to-one correspondence, this is true independently of the time at which $x_t$ is to be estimated.

Given two data subsets $\mathbf{v}_1$ and $\mathbf{v}_2$, with respective information contents $I(\mathbf{v}_1)$ and $I(\mathbf{v}_2)$, the information content $I(\mathbf{v})$ of the union set $\mathbf{v} = ({\mathbf{v}_1}^T, {\mathbf{v}_2}^T)^T$ is equal to $I(\mathbf{v}_1) + I(\mathbf{v}_2)$ if the errors affecting $\mathbf{v}_1$ and $\mathbf{v}_2$are uncorrelated. If that is not the case, $\mathbf{v}_1$ and $\mathbf{v}_2$ can be said to be *positively*, or *negatively* correlated depending on whether $I(\mathbf{v}) < I(\mathbf{v}_1) + I(\mathbf{v}_2)$ or $I(\mathbf{v}) > I(\mathbf{v}_1) + I(\mathbf{v}_2)$. This defines a sign (and actually a magnitude) for the correlation between two subsets of data. The information content being invariant in a linear transformation in data space (and in particular in a change of sign in any

of the individual data), that correlation is not systematically related to the sign of the numerical correlations between the components of the errors affecting $\mathbf{v}_1$ and $\mathbf{v}_2$ (in general, those correlations make up a whole matrix, with no unambiguous sign).

The information content $I(\mathbf{z}_k)$ quantifies the relative contribution of subset $\mathbf{z}_k$ to the overall accuracy of the estimate of the state vector $\mathbf{x}$. That notion can be extended to the measure of the contribution of $\mathbf{z}_k$ to the accuracy of the estimate of any subset, say $\mathbf{u}_1$, of $\mathbf{x}$. To see that, denote $n_1$ the dimension of $\mathbf{u}_1$, and decompose the state vector $\mathbf{x}$ into $({\mathbf{u}_1}^T, {\mathbf{u}_2}^T)^T$ where $\mathbf{u}_2$ is the projection of $\mathbf{x}$ onto the subspace that is $\mathbf{P}^a$–Mahalanobis orthogonal to $\mathbf{u}_1$. In that decomposition, the estimation error covariance matrix $\mathbf{P}^a$ reads

$$\mathbf{P} = \mathrm{diag}(\mathbf{P}_1^a, \mathbf{P}_2^a)$$

with

$$\mathbf{P}_1^a = \mathcal{E}[(\mathbf{u}_1^a - \mathbf{u}_1^t)(\mathbf{u}_1^a - \mathbf{u}_1^t)^T],$$
$$\mathbf{P}_2^a = \mathcal{E}[(\mathbf{u}_2^a - \mathbf{u}_2^t)(\mathbf{u}_2^a - \mathbf{u}_2^t)^T],$$

(where the superscript $a$ denotes, as before, analysis). As for the data operator $\boldsymbol{\Gamma}_k$, it decomposes into

$$\boldsymbol{\Gamma}_k = (\boldsymbol{\Gamma}_{k,1},\ \boldsymbol{\Gamma}_{k,2}),$$

where $\boldsymbol{\Gamma}_{k,1}$ ($\boldsymbol{\Gamma}_{k,2}$) defines the contribution of $\mathbf{u}_1$ ($\mathbf{u}_2$) to the data subset $\mathbf{z}_k$. Equation (5) decomposes in turn into

$$[\mathbf{P}_1^a]^{-1} = \sum_k \boldsymbol{\Gamma}_{k,1}^T \mathbf{S}_k^{-1} \Gamma_{k,1}, \tag{29a}$$

$$[\mathbf{P}_2^a]^{-1} = \sum_k \boldsymbol{\Gamma}_{k,2}^T \mathbf{S}_k^{-1} \Gamma_{k,2} \tag{29b}$$

The same derivation that has led from Eqs. (26) and (27), started this time from Eq. (29a), leads to defining

$$I_1(\mathbf{z}_k) \equiv \frac{1}{n} \mathrm{tr}(\mathbf{S}_k^{-1/2} \boldsymbol{\Gamma}_{k,1} \mathbf{P}^a \boldsymbol{\Gamma}_{k,1}^T \mathbf{S}_k^{-1/2}), \tag{30}$$

as being the relative contribution of the data subset $\mathbf{z}_k$ to the accuracy of the estimation of $\mathbf{u}_1$. One can thus define the relative contribution of any subset of the data (for instance, the infrared radiances in a given channel over a given geographical area) to the accuracy of the estimate of any subset of the analysed fields (for instance, the estimate of humidity over that same area).

Numerical determination of $I_1(\mathbf{z}_k)$ seems, however, to raise serious problems, since it requires the identification, in one form of another, of the subspace in $\boldsymbol{\mathcal{S}}$ that is $\mathbf{P}^a$–Mahalanobis orthogonal to $\mathbf{u}_1$. It is not clear how that could be achieved in practice.

## 6 Diagnostics of Optimality of Assimilation Algorithms

The various diagnostics that have been presented in the previous sections allow objective comparison of the quality of different assimilation schemes, or evaluation of the internal consistency of a given scheme. They say nothing as to the optimality, or otherwise, of a given scheme. The *BLUE* is defined on conditions of statistical unbiasedness and minimum estimation error variance. As a consequence, the estimation error $\mathbf{x}^a$–$\mathbf{x}^t$, in addition to being unbiased, must be statistically uncorrelated with the DmA difference or, equivalently, with the innovation vector. This is expressed by Eq. (10a), where the second term on the right-hand side is the orthogonal projection, in the sense of covariance, of (minus) the background error $\mathbf{x}^t$– $\mathbf{x}^b$ onto the space spanned by the innovation $\mathbf{y} - \mathbf{H}\mathbf{x}^b$ (see also Eq. 20). The optimality condition is often expressed, in an exactly equivalent way, by saying that a sequential algorithm for assimilation is optimal if, and only if, the temporal sequence of innovation vectors is unbiased and uncorrelated (Kailath 1968).

This optimality condition can be objectively checked against independent observations. Let us consider an observation of the form

$$\mathbf{q} = \mathbf{D}\mathbf{x}^t + \boldsymbol{\gamma},$$

where $\mathbf{D}$ is a known linear operator, and the error $\boldsymbol{\gamma}$ is assumed to be unbiased and uncorrelated with the data error $\boldsymbol{\varepsilon}$, and therefore with the innovation $\mathbf{d}$. Optimality of the estimate $\mathbf{q}^a = \mathbf{D}\mathbf{x}^a$ of $\mathbf{w}$ is equivalent to the conditions that it be statistically unbiased

$$\mathcal{E}[\mathbf{q} - \mathbf{D}\mathbf{x}^a] = 0, \tag{31}$$

and uncorrelated with the innovation

$$\mathcal{E}[(\mathbf{q} - \mathbf{D}\mathbf{x}^a)\mathbf{d}^T] = 0. \tag{32}$$

If the unbiasedness condition given by Eq. (31) is usually checked in assimilation systems, the uncorrelatedness condition given by Eq. (32), in spite of its simplicity, has so far been rarely used. One of the few examples is a work by Daley (1992), who computed the correlation of the innovation sequence for the sequential assimilation system that was then in use at the Canadian Meteorological Centre (that system is described by Mitchell et al. 1990). He found significantly non-zero correlations, reaching values of more than 0.4 for the 500 hPa geopotential innovation, at a time-lag of 12 h. Similar tests, performed more recently on a system for assimilation of oceanographical observations, led to correlation values around 0.3 (Miller, personal communication).

The diagnostic given by Eqs. (31) and (32), if used alone, is actually a "one-way" diagnostic. If the observed correlation is found to be significantly different from 0, as in the two examples above, that is a proof that the assimilation system is suboptimal, and can be improved. But if the correlation is found to be statistically

undistinguishable from 0, that does not mean that the system cannot be improved. To see that, consider a system which uses as background a short-range forecast produced by a high-quality Numerical Weather Prediction model, and suppose the system uses as background error covariance matrix $\mathbf{P}^b$ the matrix of climatological covariances. That is not erroneous, since the long term statistical distribution of the background must be close to the climatological distribution. And, provided the covariance matrix of observation error is correctly specified, one can expect that the covariance (Eq. 32) will be 0. However, in view of the quality of present short range numerical weather forecasts, it is clear that such a system could be significantly improved. Actually, a system that is suboptimal by the criterion given by Eq. (32) can very well produce much more accurate estimates than an optimal "climatological" system.

This first shows that the diagnostic given by Eq. (32) does not have much meaning if it not associated with diagnostics of the magnitude of the difference $\mathbf{q} - \mathbf{D}\mathbf{x}^a$. That is not a problem inasmuch as such diagnostics are performed routinely. But this short discussion also shows that it is impossible to objectively determine, at least on the basis of diagnostics of form given by Eqs. (31) and (32), whether an assimilation system makes the best possible use of the available data.

On the other hand, that certainly does not mean that diagnostics of form Eq. (32) should not be used at all. As mentioned, they have rarely been used so far, but they can objectively detect suboptimality, and would certainly be a useful complement to other commonly used diagnostics.

## 7 Conclusions

We have studied in some detail, in the context of the *BLUE*, the three questions stated in the Introduction. The answer to the first question ($Q1$), relative to the possibility of objectively evaluating the quality of an assimilation algorithm, is fairly obvious. Such an evaluation can be made only against unbiased observations that have not only not been used in the assimilation, but are affected by errors that are uncorrelated with the errors affecting the data that have been used in the assimilation (in the general case of a non-linear estimation scheme, the condition would be that the errors affecting the verifying observations must be statistically independent of the errors affecting the data that have been used in the assimilation).

The second question ($Q2$) was relative to the possibility of objectively determining the probability distribution of the errors affecting the data (the expectation $\boldsymbol{\mu}$ and the covariance matrix $\mathbf{S}$ in the case of the *BLUE*). It has led to the conclusion that (except for trial and error tuning, which cannot be exhaustive in meteorological or oceanographical applications) this will always require external hypotheses, i.e., hypotheses that cannot be objectively validated on the basis of the data only (incidentally, the author does not know if this result, which has been shown here on the basis of a fundamentally linear argument, extends to non-linear estimation). Appropriate external information is always available in meteorological

and oceanographical applications, but is largely unsufficient to entirely define the required quantities $\boldsymbol{\mu}$ and $\mathbf{S}$. Now, there is no other choice in practice than making hypotheses about the statistics of the errors affecting the data. It is important to distinguish as clearly as possible between hypotheses that are very unlikely to be ever modified (such as, for instance, that errors in radiosonde observations performed a long distance apart are uncorrelated), from hypotheses that are reasonable but probably disputable (such as, for instance, that observation errors are statistically uncorrelated with background errors), and from hypotheses that are made for convenience, but are very presumably erroneous (such as, for instance, that model errors are absent, or even only uncorrelated in time). Ideally, one might wish to define a minimum set of reliable hypotheses such that all remaining necessary error statistics can be objectively determined from statistics of the innovation. That goal seems, however, to be somewhat elusive in the present state of assimilation of meteorological and oceanographical observations. On the other hand, methods such as generalized cross validation (Wahba et al. 1995), which are ultimately trial and error experimentation, but are based on a solid methodological approach, have certainly not received enough attention in meteorological and oceanographical applications.

Note that systematic comparison between a priori assumed and a posteriori statistics of the innovation (or equivalently of the DmA difference) can reveal inconsistencies for which they cannot be unambiguous interpretation, but which can, if used with good judgment, help improve the a priori specification of $\boldsymbol{\mu}$ and $\mathbf{S}$.

Concerning objective estimation of the optimality of an assimilation algorithm (*Q*3), the decorrelation criterion (Eq. 32) is valid only for least squares estimation (but can extend to non-linear least squares estimation). Although it can prove nothing as to the accuracy of the assimilation, it can nevertheless be useful, and has likely also not received enough attention.

**Acknowledgments** The author thanks numerous colleagues, in particular F. Bouttier, B. Chapnik and G. Desroziers, for stimulating discussions. B. Chapnik provided Fig. 2.

## References

Bennett, A.F., 1992. *Inverse Methods in Physical Oceanography,* Cambridge University Press, Cambridge, UK, 346pp.

Cañizares, R., A. Kaplan, M.A. Cane, D. Chen and S.E. Zebiak, 2001. Use of data assimilation via linear low-order models for the initialization of El Niño – Southern Oscillation predictions. *J. Geophys. Res.*, **106**, 30947–30959.

Chapnik, B., G. Desroziers, F. Rabier and O. Talagrand, 2006. Diagnosis and tuning of observational error statistics in a quasi-operational data assimilation setting. *Q. J. R. Meteorol. Soc*., **132**, 543–565, doi:10.1256/qj.04.102.

Daley, R., 1992. The lagged innovation covariance: A performance diagnostic for atmospheric data assimilation. *Mon. Weather Rev*., **120**, 178–196.

Daley, R., 1993. Estimating observation error statistics for atmospheric data assimilation. *Ann. Geophysicae*, **11**, 634–647.

Desroziers, G., P. Brousseau and B. Chapnik, 2005. Use of randomization to diagnose the impact of observations on analyses and forecasts. *Q. J. R. Meteorol. Soc*., **131**, 2821–2837, doi:10.1256/qj.04.151.

Desroziers, G. and S. Ivanov, 2001. Diagnosis and adaptive tuning of observation-error parameters in a variational assimilation. *Q. J. R. Meteorol. Soc*., **127**, 1433–1452.

Elbern, H., A. Strunk, H. Schmidt and O. Talagrand, 2007. Emission rate and chemical state estimation by 4-dimensional variational inversion. *Atmos. Chem. Phys*., **7**, 3749–3769.

Fisher, M., 2003. *Estimation of Entropy Reduction and Degrees of Freedom for Signal for Large Variational Analysis Systems*, Technical Memorandum No 397, Research Department, ECMWF, Reading, UK, 18pp., available at the address http://www.ecmwf.int/publications/library/do/references/show?id=83951.

Girard, D., 1987. *A fast Monte-Carlo Cross-Validation Procedure for Large Least Square Problems with Noisy Data*, Technical report RR 687-M, IMAG. Université de Grenoble, Grenoble, France, 22pp.

Hollingsworth, A. and P. Lönnberg, 1989. The verification of objective analyses: Diagnostic of analysis system performance. *Meteorol. Atmos. Phys*., **40**, 3–27.

Kailath, T., 1968. An innovations approach to least-squares estimation. Part I: Linear filtering in additive white noise. *IEEE Trans. Automat. Contr*., **AC-13**(6), 646–655.

Ménard, R. and L.-P. Chang, 2000. Assimilation of stratospheric chemical tracer observations using a Kalman filter. Part II: chi2-validated results and analysis of variance and correlation dynamics. *Mon. Weather Rev*., **128**, 2672–2686.

Mitchell, H., C. Charette, C. Chouinard and B. Brasnett, 1990. Revised interpolation statistics for the Canadian data assimilation procedure: Their derivation and application. *Mon. Weather Rev*., **118**, 1591–1614.

Muccino, J.C., N.F. Hubele and A.F. Bennett, 2004. Significance testing for variational assimilation. *Q. J. R. Meteorol. Soc.*, **130**, 1815–1838, doi:10.1256/qj.03.47.

Rabier, F., N. Fourrié, D. Chafaï and P. Prunet, 2002. Channel selection methods for infrared atmospheric sounding interferometer radiances. *Q. J. R. Meteorol. Soc.,* **128**, 1011–1027.

Rodgers, C.D., 2000. *Inverse Methods for Atmospheric Sounding: Theory and Practice*, World Scientific Publishing Co. Ltd, London, UK, 238pp.

Sadiki, W. and C. Fischer, 2005. A posteriori validation applied to the 3D-VAR Arpège and Aladin data assimilation systems. *Tellu*s, **57A**, 21–34.

Talagrand, O., 1999. A posteriori evaluation and verification of analysis and assimilation algorithms. In *Proceedings of Workshop on Diagnosis of Data Assimilation Systems* (November 1998), ECMWF. Reading, England, 17–28, available at the address http://www.ecmwf.int/publications/library/do/references/show?id=87283.

Talagrand, O. and F. Bouttier, 2000. Internal diagnostics of data assimilation systems. In *Proceedings of Seminar on Diagnosis of Models and Data Assimilation Systems* (September 1999), ECMWF. Reading, UK, pp. 407–409.

Wahba, G., D. Johnson, F. Gao and J. Gong, 1995. Adaptive tuning of numerical weather prediction models: Randomized GCV in three and four dimensional data assimilation. *Mon. Weather Rev*., **123**, 3358–3369.

# Initialization

**Peter Lynch and Xiang-Yu Huang**

## 1 Introduction

The spectrum of atmospheric motions is vast, encompassing phenomena having periods ranging from seconds to millennia. The motions of interest to the forecaster typically have timescales of a day or longer, but the mathematical models used for numerical prediction describe a broader span of dynamical features than those of direct concern. For many purposes these higher frequency components can be regarded as *noise* contaminating the motions of meteorological interest. The elimination of this noise is achieved by adjustment of the initial fields, a process called *initialization*.

The natural oscillations of the atmosphere fall into two groups (see e.g. Kasahara 1976). The solutions of meteorological interest have low frequencies and are close to geostrophic balance. They are called rotational or vortical modes, since their vorticity is greater than their divergence; if divergence is ignored, these modes reduce to the Rossby-Haurwitz waves. There are also very fast gravity-inertia wave solutions, with phase speeds of hundreds of metres per second and large divergence. For typical conditions of large scale atmospheric flow (when the Rossby and Froude numbers are small) the two types of motion are clearly separated and interactions between them are weak. The high frequency gravity-inertia waves may be locally significant in the vicinity of steep orography, where there is strong thermal forcing or where very rapid changes are occurring; but overall they are of minor importance.

A subtle and delicate state of balance exists in the atmosphere between the wind and pressure fields, ensuring that the fast gravity waves have much smaller amplitude than the slow rotational part of the flow. Observations show that the pressure and wind fields in regions not too near the Equator are close to a state of geostrophic

P. Lynch (✉)
University College Dublin, Dublin, Ireland
e-mail: peter.lynch@ucd.ie

W. Lahoz et al. (eds.), *Data Assimilation*, DOI 10.1007/978-3-540-74703-1_9,

balance and the flow is quasi-non-divergent. The bulk of the energy is contained in the slow rotational motions and the amplitude of the high frequency components is small. The existence of this geostrophic balance is a perennial source of interest; it is a consequence of the forcing mechanisms and dominant modes of hydrodynamic instability and of the manner in which energy is dispersed and dissipated in the atmosphere. For a recent review of balanced flow, see McIntyre (2003). The gravity-inertia waves are instrumental in the process by which the balance is maintained, but the nature of the sources of energy ensures that the low frequency components predominate in the large scale flow. The atmospheric balance is subtle, and difficult to specify precisely. It is *delicate* in that minor perturbations may disrupt it but *robust* in that local imbalance tends to be rapidly removed through radiation of gravity-inertia waves in a process known as geostrophic adjustment.

When the basic equations are used for numerical prediction the forecast may contain spurious large amplitude high frequency oscillations. These result from anomalously large gravity-inertia waves which occur because the balance between the mass and velocity fields is not reflected faithfully in the analysed fields. High frequency oscillations of large amplitude are engendered, and these may persist for a considerable time unless strong dissipative processes are incorporated in the forecast model. It was the presence of such imbalance in the initial fields which gave rise to the totally unrealistic pressure tendency of 145 hPa/6 h obtained by Lewis Fry Richardson in the first-ever objective numerical weather forecast (Richardson 1922, Lynch 2006).

Although they have little effect on the long-term evolution of the flow, gravity waves may profoundly influence the way it changes on shorter time-scales. Figure 1 schematically depicts the pressure variation over a period of 1 day. The smooth curve represents the variation due to meteorological effects; its gentle slope (dotted line) indicates the long-term change (Phillips 1973). The rapidly varying curve represents the actual pressure changes when gravity waves are superimposed on the meteorological flow: the slope of the oscillating curve (dashed line) is precipitous and, if used to determine long-range variations, yields totally misleading results. What Richardson calculated was the instantaneous rate of change in pressure for an atmospheric state having gravity wave components of large amplitude.

If the fields are not initialized, the spurious oscillations which occur in the forecast can lead to various problems. In particular, new observations are checked for accuracy against a short-range forecast. If this forecast is noisy, good observations may be rejected or erroneous ones accepted. Thus, *initialization is essential for satisfactory data assimilation* (see other chapters in Part I, *Theory*, for a discussion of data assimilation). Another problem occurs with precipitation forecasting. A noisy forecast has unrealistically large vertical velocity. This interacts with the humidity field to give hopelessly inaccurate rainfall patterns. To avoid this *spin-up*, we must control the gravity wave oscillations.

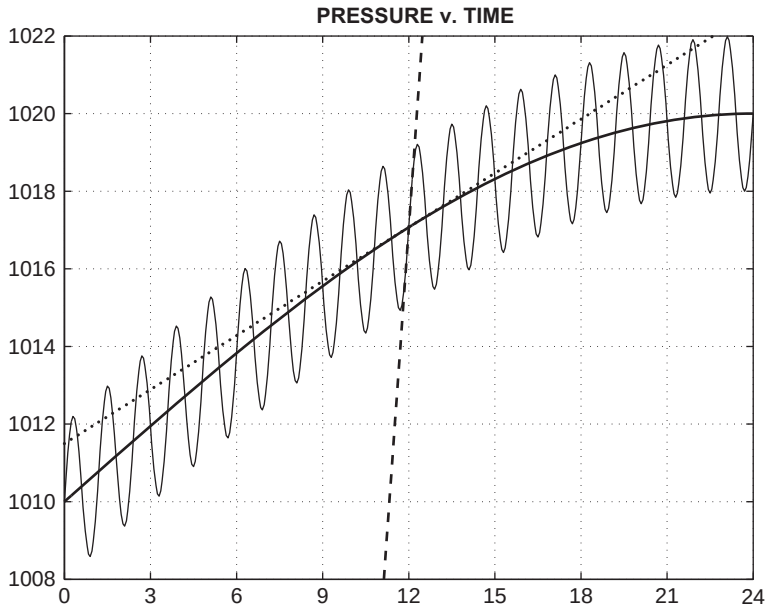



**Fig. 1** Schematic illustration of pressure variation over a 24 h period. The *thick line* is the mean, long-term variation, the *thin line* is the actual pressure, with high frequency noise. The *dotted line* shows the rate of change, at 12 h, of the mean pressure and the *dashed line* shows the corresponding rate of change of the actual pressure (after Phillips 1973)

## 2 Early Initialization Methods

### *2.1 The Filtered Equations*

The first computer forecast was made in 1950 by Charney, Fjørtoft and Von Neumann (Charney et al. 1950). In order to avoid Richardson's error, they modified the prediction equations in such a way as to eliminate the high frequency solutions. This process is known as filtering. The basic filtered system is the set of quasi-geostrophic equations. These equations were used in operational forecasting for a number of years. However, they involve approximations which are not always valid, and this can result in poor forecasts. A more accurate filtering of the primitive equations leads to the *balance equations*. This system is more complicated to solve than the quasi-geostrophic system, and has not been widely used.

### *2.2 Static Initialization*

Hinkelmann (1951) investigated the problem of noise in numerical integrations and concluded that if the initial winds were geostrophic, high frequency oscillations

would occur but would remain small in amplitude. He later succeeded in integrating the primitive equations, using a very short timestep, with geostrophic initial winds (Hinkelmann 1959). Forecasts made with the primitive equations were soon shown to be clearly superior to those using the quasi-geostrophic system. However, the use of geostrophic initial winds had a huge disadvantage: the valuable information contained in the observations of the wind field was completely ignored. Moreover, the remaining noise level is not tolerable in practice. Charney (1955) proposed that a better estimate of the initial wind field could be obtained by using the non-linear balance equation. This equation — part of the balance system — is a diagnostic relationship between the pressure and wind fields. It implies that the wind is non-divergent. It was later argued by Phillips (1960) that a further improvement would result if the divergence of the initial field were set equal to that implied by quasi-geostrophic theory. Each of these steps represented some progress, but the noise problem still remained essentially unsolved.

## *2.3 Dynamic Initialization*

Another approach, called dynamic initialization, uses the forecast model itself to define the initial fields (Miyakoda and Moyer 1968). The dissipative processes in the model can damp out high frequency noise as the forecast proceeds. We integrate the model first forward and then backward in time, keeping the dissipation active all the time. We repeat this forward–backward cycle many times until we finally obtain fields, valid at the initial time, from which the high frequency components have been damped out. The forecast starting from these fields is noise-free. However, the procedure is expensive in computer time, and damps the meteorologically significant motions as well as the gravity waves, so it is no longer popular. Digital filtering initialization, described below, is essentially a refinement of dynamic initialization. Because it used a highly selective filtering technique, it is computationally more efficient than the older method.

## *2.4 Variational Initialization*

An elegant initialization method based on the calculus of variations was introduced by Sasaki (1958). We consider the simplest case: given an analysis of the mass and wind fields, how can they be minimally modified so as to impose geostrophic balance? This problem can be formulated as the minimization of an integral representing the deviation of the resulting fields from balance. The variation of the integral leads to the Euler-Lagrange equations, which yield diagnostic relationships for the new mass and wind fields in terms of the incoming analysis. Although the method was not widely used, the variational method is now at the centre of modern data assimilation practice. In Sect. 6 below we discuss the use of a digital filter as a weak constraint in four-dimensional variational assimilation (4D-Var; see chapter *Variational Assimilation*, Talagrand).

## 3 Atmospheric Normal Mode Oscillations

The solutions of the model equations can be separated, by a process of spectral analysis, into two sets of components or linear normal modes, slow rotational components or Rossby modes, and high frequency gravity modes. We assume that the amplitude of the motion is so small that all non-linear terms can be neglected. The horizontal structure is then governed by a system equivalent to the linear shallow water equations which describe the small-amplitude motions of a shallow layer of incompressible fluid. These equations were first derived by Laplace in his discussion of tides in the atmosphere and ocean, and are called the Laplace tidal equations. The simplest means of deriving the linear shallow water equations from the primitive equations is to assume that the vertical velocity vanishes identically.

### *3.1 The Laplace Tidal Equations*

Let us assume that the motions under consideration can be described as small perturbations about a state of rest, in which the temperature is a constant, $T_0$, and the pressure $\bar{p}(z)$ and density $\bar{\rho}(z)$ vary only with height. The basic state variables satisfy the gas law and are in hydrostatic balance: $\bar{p} = \mathcal{R}\bar{\rho}T_0$ and $d\bar{p}/dz = -g\bar{\rho}$. The variations of mean pressure and density follow immediately:

$$\bar{p}(z) = p_0 \exp(-z/H), \qquad \bar{\rho}(z) = \rho_0 \exp(-z/H),$$

where $H = p_0/g\rho_0 = \mathcal{R}T_0/g$ is the scale-height of the atmosphere. We consider only motions for which the vertical component of velocity vanishes identically, $w \equiv 0$. Let $u$, $v$, $p$ and $\rho$ denote variations about the basic state, each of these being a small quantity. The horizontal momentum, continuity and thermodynamic equations, with standard notation, are (see chapters *The Role of the Model in the Data Assimilation System*, Rood; *General Concepts in Meteorology and Dynamics*, Charlton-Perez et al.)

$$\frac{\partial \bar{\rho}u}{\partial t} - f\bar{\rho}v + \frac{\partial p}{\partial x} = 0 \tag{1}$$

$$\frac{\partial \bar{\rho}v}{\partial t} + f\bar{\rho}u + \frac{\partial p}{\partial y} = 0 \tag{2}$$

$$\frac{\partial \rho}{\partial t} + \nabla\cdot\bar{\rho}\mathbf{V} = 0 \tag{3}$$

$$\frac{1}{\gamma\bar{p}}\frac{\partial p}{\partial t} - \frac{1}{\bar{\rho}}\frac{\partial \rho}{\partial t} = 0 \tag{4}$$

Density can be eliminated from the continuity equation, Eq. (3), by means of the thermodynamic equation, Eq. (4). Now let us assume that the horizontal and vertical dependencies of the perturbation quantities are separable:

$$\begin{Bmatrix} \bar{\rho}u \\ \bar{\rho}v \\ p \end{Bmatrix} = \begin{Bmatrix} U(x,y,t) \\ V(x,y,t) \\ P(x,y,t) \end{Bmatrix} Z(z). \tag{5}$$

The momentum and continuity equations can then be written

$$\frac{\partial U}{\partial t} - fV + \frac{\partial P}{\partial x} = 0 \tag{6}$$

$$\frac{\partial V}{\partial t} + fU + \frac{\partial P}{\partial y} = 0 \tag{7}$$

$$\frac{\partial P}{\partial t} + (gh)\nabla \cdot V = 0 \tag{8}$$

where $\mathbf{V} = (U, V)$ is the momentum vector and $h = \gamma H = \gamma \mathcal{R} T_0/g$. This is a set of three equations for the three dependent variables $U$, $V$, and $P$. They are mathematically isomorphic to the Laplace tidal equations with a mean depth $h$. The quantity $h$ is called the equivalent depth. There is no dependence in this system on the vertical coordinate $z$.

The vertical structure follows from the hydrostatic equation, together with the relationship $p = (\gamma g H)\rho$ implied by the thermodynamic equation. It is determined by

$$\frac{dZ}{dz} + \frac{Z}{\gamma H} = 0, \tag{9}$$

the solution of which is $Z = Z_0 \exp(-z/\gamma H)$, where $Z_0$ is the amplitude at $z = 0$. If we set $Z_0 = 1$, then $U$, $V$ and $P$ give the momentum and pressure fields at the Earth's surface. These variables all decay exponentially with height. It follows from Eq. (5) that $u$ and $v$ actually increase with height as $\exp(\kappa z/H)$, but the kinetic energy decays.

## *3.2 Vorticity and Divergence*

We examine the solutions of the Laplace tidal equations in some enlightening limiting cases. Holton (1992) gives a more extensive analysis, including treatments of the equatorial and mid-latitude $\beta$-plane approximations. By means of the Helmholtz Theorem, a general horizontal wind field $\mathbf{V}$ may be partitioned into rotational and divergent components

$$\mathbf{V} = \mathbf{V}_\psi + \mathbf{V}_\chi = k \times \nabla\psi + \nabla\chi \, .$$

The stream function $\psi$ and velocity potential $\chi$ are related to the vorticity and divergence by the Poisson equations $\nabla^2\psi = \zeta$ and $\nabla^2\chi = \delta$, respectively. It is straightforward to derive equations for the vorticity and divergence tendencies.

Together with the continuity equation, they are

$$\frac{\partial \zeta}{\partial t} + f\delta + \beta v = 0 \tag{10}$$

$$\frac{\partial \delta}{\partial t} - f\zeta + \beta u + \nabla^2 P = 0 \tag{11}$$

$$\frac{\partial P}{\partial t} + gh\delta = 0\,. \tag{12}$$

These equations are completely equivalent to Eqs. (6), (7), and (8); no additional approximations have yet been made. However, the vorticity and divergence forms enable us to examine various simple approximate solutions.

### *3.3 Rossby-Haurwitz Modes*

If we suppose that the solution is quasi-non-divergent, i.e., we assume $|\delta| \ll |\zeta|$, the wind is given approximately in terms of the stream function $(u, v) \approx (-\psi_y, \psi_x)$, and the vorticity equation becomes

$$\nabla^2 \psi_t + \beta \psi_x = O(\delta)\,, \tag{13}$$

and we can ignore the right-hand side. Assuming the stream function has the wave-like structure of a spherical harmonic, $Y_n^m(\lambda, \phi) = P_n^m(\sin\phi)\exp(im\lambda)$, we substitute the expression $\psi = \psi_0 Y_n^m(\lambda, \phi)\exp(-ivt)$ in the vorticity equation and immediately deduce an expression for the frequency:

$$v = v_R \equiv -\frac{2\Omega m}{n(n+1)}\,. \tag{14}$$

This is the celebrated dispersion relation for Rossby-Haurwitz waves (Haurwitz 1940). If we ignore sphericity (the $\beta$-plane approximation) and assume harmonic dependence $\psi(x, y, t) = \psi_0 \exp[i(kx + \ell y - vt)]$, then Eq. (13) has the dispersion relation

$$c = \frac{v}{k} = -\frac{\beta}{k^2 + \ell^2}\,,$$

which is the expression for phase-speed found by Rossby (1939). The Rossby or Rossby-Haurwitz waves are, to the first approximation, non-divergent waves which travel westward, the phase speed being greatest for the waves of largest scale. They are of relatively low frequency — Eq. (14) implies that $|v| \leq \Omega$ — and the frequency decreases as the spatial scale decreases.

To the same degree of approximation, we may write the divergence equation, Eq. (11), as

$$\nabla^2 P - f\zeta - \beta\psi_y = O(\delta)\,. \tag{15}$$

Ignoring the right-hand side of Eq. (11), we get the *linear balance equation*

$$\nabla^2 P = \boldsymbol{\nabla}\cdot f\boldsymbol{\nabla}\boldsymbol{\psi}\,, \tag{16}$$

a diagnostic relationship between the geopotential and the stream function. This also follows immediately from the assumption that the wind is both non-divergent ($\mathbf{V} = k \times \nabla\psi$) and geostrophic ($f\mathbf{V} = k \times \nabla P$). If variations of $f$ are ignored, we can assume $P = f\psi$. The wind and pressure are in approximate geostrophic balance for Rossby-Haurwitz waves.

## 3.4 Gravity Wave Modes

If we assume now that the solution is quasi-irrotational, i.e. that $|\zeta| \ll |\delta|$, then the wind is given approximately by $(u, v) \approx (\chi_x, \chi_y)$ and the divergence equation becomes

$$\nabla^2\chi_t + \beta\chi_x + \nabla^2 P = O(\zeta)$$

with the right-hand side negligible. Using the continuity equation to eliminate $P$, we get

$$\nabla^2\chi_{tt} + \beta\chi_{xt} - gh\nabla^4\chi = 0\,.$$

Seeking a solution $\chi = \chi_0 Y_n^m(\lambda, \phi)\exp(-i\nu t)$, we find that

$$\nu^2 + \left(-\frac{2\Omega m}{n(n+1)}\right)\nu - \frac{n(n+1)gh}{a^2} = 0\,. \tag{17}$$

The coefficient of the second term is just the Rossby-Haurwitz frequency $\nu_R$ found in Eq. (14) above, so that

$$\nu = \pm\sqrt{\nu_G^2 + (\frac{1}{2}\nu_R)^2} - \frac{1}{2}\nu_R\,, \qquad \text{where} \qquad \nu_G \equiv \sqrt{\frac{n(n+1)gh}{a^2}}\,.$$

Noting that $|\nu_G| \gg |\nu_R|$, it follows that

$$\nu_\pm \approx \pm\nu_G\,,$$

the frequency of pure gravity waves. There are then two solutions, representing waves travelling eastward and westward with equal speeds. The frequency increases approximately linearly with the total wavenumber $n$.

## 4 Normal Mode Initialization

The model equations, Eqs. (10), (11), and (12) can be written schematically in the form

$$\dot{\mathbf{X}} + i\mathbf{L}\mathbf{X} + \mathcal{N}(\mathbf{X}) = \mathbf{0} \tag{18}$$

with $\mathbf{X}$ the state vector, $\mathbf{L}$ a matrix and $\mathcal{N}$ a non-linear vector function. If $\mathbf{L}$ is diagonalized, the system separates into two subsystems, for the low and high frequency components (LF and HF, respectively):

$$\dot{\mathbf{Y}} + i\Lambda_Y\mathbf{Y} + \mathcal{N}(\mathbf{Y}, \mathbf{Z}) = \mathbf{0} \tag{19}$$

$$\dot{\mathbf{Z}} + i\Lambda_Z\mathbf{Z} + \mathcal{N}(\mathbf{Y}, \mathbf{Z}) = \mathbf{0} \tag{20}$$

where $\mathbf{Y}$ and $\mathbf{Z}$ are the coefficients of the LF and HF components of the flow, referred to colloquially as the *slow* and *fast* components respectively, and $\Lambda_Y$ and $\Lambda_Z$ are diagonal matrices of eigenfrequencies for the two types of modes.

Let us suppose that the initial fields are separated into slow and fast parts, and that the latter are removed so as to leave only the Rossby waves. It might be hoped that this process of "linear normal mode initialization", imposing the condition

$$\mathbf{Z} = \mathbf{0} \quad \text{at} \quad t = 0$$

would ensure a noise-free forecast. However, the results of the technique are disappointing: the noise is reduced initially, but soon reappears; the forecasting equations are non-linear, and the slow components interact non-linearly in such a way as to generate gravity waves. The problem of noise remains: the gravity waves are small to begin with, but they grow rapidly (see Daley 1991; Chap. 9).

Machenhauer (1977) examined gravity wave dynamics in simple systems and found that the amplitude of the high-frequency components is quasi-stationary. To control the growth of HF components, he proposed setting their initial rate of change to zero, in the hope that they would remain small throughout the forecast. Baer (1977) proposed a somewhat more general method, using a two-timing perturbation technique. The forecast, starting from initial fields modified so that $\dot{\mathbf{Z}} = \mathbf{0}$ at $t = 0$ is very smooth and the spurious gravity wave oscillations are almost completely removed. The method takes account of the non-linear nature of the equations, and is referred to as non-linear normal mode initialization:

$$\dot{\mathbf{Z}} = \mathbf{0} \quad \text{at} \quad t = 0\,.$$

The method is comprehensively reviewed in Daley (1991).

In Fig. 2, we show the evolution of surface pressure for three 24-h forecasts (Williamson and Temperton 1981). The solid lines (in both panels) are the pressure variation for forecasts from uninitialized data. Forecasts from linearly initialized

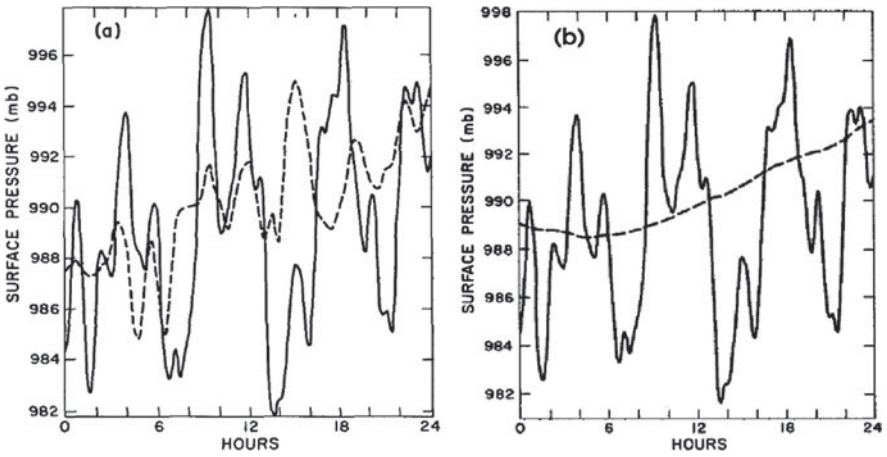



**Fig. 2** Pressure variation over a 24-h period for forecasts from uninitialized data (*solid lines, both panels*, LNMI data (*dashed line*, *left panel*) and NNMI data (*dashed line*, *right panel*). LNMI = linearly initialized data; NNMI = non-linearly initialized data. From Williamson and Temperton (1981)

data (LNMI) are shown by the dashed line in the left panel. Forecasts from data that is non-linearly initialized (NNMI) are shown by the dashed line in the right panel. It is clear that LNMI is ineffective in removing spurious oscillations. NNMI is excellent in this regard.

## 5 Digital Filter Initialization

Normal mode initialization, or NMI, has been used in many NWP (Numerical Weather Prediction) centres, and has performed satisfactorily. Its most natural context is for global models, for which the horizontal structure of the normal modes corresponds to the Hough functions, the eigenmodes of the Laplace tidal equations. For limited area models, normal modes can also be derived, but the lateral boundaries force the introduction of simplifying assumptions. An alternative method of initialization, called digital filter initialization (DFI), was introduced by Lynch and Huang (1992). It was generalized to allow for diabatic effects by Huang and Lynch (1993). The latter paper also discussed the use of an optimal filter. A much simpler filter, the Dolph-Chebyshev filter, which is a special case of the optimal filter, was applied to the initialization problem by Lynch (1997). A more efficient formulation of DFI was presented by Lynch et al. (1997).

Digital filter initialization (DFI) uses filters similar to those arising in signal processing. The selection principle for these is generally based on the frequency of the signal components. There are a number of ideal types — lowpass, highpass, bandpass and bandstop — corresponding to the range of frequencies which pass through the filter and those which are rejected. In many cases the input consists of a low frequency (LF) signal contaminated by high frequency (HF) noise, and the information

in the signal can be isolated by using a lowpass filter which rejects the noise. Such a situation is typical for the application to meteorology discussed below.

The method of digital filter initialization has significant advantages over alternative methods, and is now in use operationally at several major weather prediction centres (see chapter *Numerical Weather Prediction*, Swinbank). In DFI there is no need to compute or store normal modes; this advantage becomes more pronounced as the number of degrees of freedom of the model increases. There is no need to separate the vertical modes; NMI requires the introduction of an auxiliary geopotential variable, and partitioning of its changes between the temperature and surface pressure involves an ad hoc assumption. DFI is free from this problem. There is complete compatibility with model discretization, eliminating discretization errors due to grid disparities. DFI is applicable to exotic grids on arbitrary domains, facilitating its use with stretched or irregular model grids. There is no iterative numerical procedure which may diverge; therefore, all vertical modes can be initialized effectively. The simplicity of the method makes it easy to implement and maintain. The method is applicable to all prognostic model variables; thus, DFI produces initial fields for these variables which are compatible with the basic dynamical fields. Last but not least, DFI filters the additional prognostic variables in non-hydrostatic models in a manner identical to the basic variables. The DFI method is thus immediately suitable for non-hydrostatic models (Bubnová et al. 1995; Chen and Huang 2006).

## *5.1 Design of Non-recursive Filters*

Consider a function of time, $f(t)$, with low and high frequency components. To filter out the high frequencies one may proceed as follows:

[1] Calculate the Fourier transform $F(\omega)$ of $f(t)$;
[2] Set the coefficients of the high frequencies to zero;
[3] Calculate the inverse transform.

(See Fig. 3). Step [2] may be performed by multiplying $F(\omega)$ by an appropriate weighting function $H(\omega)$.

Suppose that $f$ is known only at discrete moments $t_n = n\Delta t$, so that the sequence $\{\cdots, f_{-2}, f_{-1}, f_0, f_1, f_2, \cdots\}$ is given. For example, $f_n$ could be the value of some model variable at a particular grid point at time $t_n$. The shortest period component that can be represented with a time step $\Delta t$ is $\tau_N = 2\Delta t$, corresponding to a maximum frequency, the so-called Nyquist frequency, $\omega_N = \pi/\Delta t$. The sequence $\{f_n\}$ may be regarded as the Fourier coefficients of a function $F(\theta)$:

$$F(\theta) = \sum_{n=-\infty}^{\infty} f_n e^{-in\theta}, \tag{21}$$

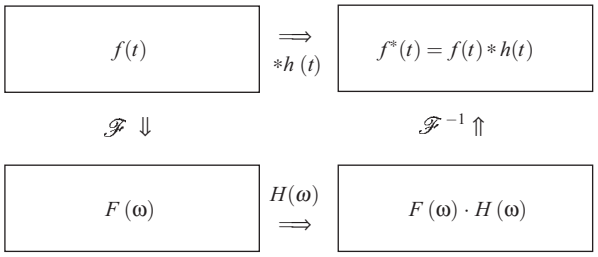



**Fig. 3** Schematic representation of the equivalence between convolution and filtering in Fourier space.

where $\theta = \omega\Delta t$ is the *digital frequency* and $F(\theta)$ is periodic, $F(\theta) = F(\theta + 2\pi)$. High frequency components of the sequence may be eliminated by multiplying $F(\theta)$ by a function $H(\theta)$ defined by

$$H(\theta) = \begin{cases} 1, & |\theta| \leq |\theta_c|; \\ 0, & |\theta| > |\theta_c|, \end{cases} \tag{22}$$

where the cutoff frequency $\theta_c = \omega_c \Delta t$ is assumed to fall in the Nyquist range $(-\pi, \pi)$ and $H(\theta)$ has period $2\pi$. This function may be expanded:

$$H(\theta) = \sum_{n=-\infty}^{\infty} h_n e^{-in\theta} \qquad ; \qquad h_n = \frac{1}{2\pi} \int_{-\pi}^{\pi} H(\theta) e^{in\theta} d\theta. \tag{23}$$

The values of the coefficients $h_n$ follow immediately from Eqs. 22 and 23:

$$h_n = \frac{\sin n\theta_c}{n\pi}. \tag{24}$$

Let $\{f_n^\star\}$ denote the low frequency part of $\{f_n\}$, from which all components with frequency greater than $\theta_c$ have been removed. Clearly,

$$H(\theta) \cdot F(\theta) = \sum_{n=-\infty}^{\infty} f_n^\star e^{-in\theta}.$$

The convolution theorem for Fourier series now implies that $H(\theta) \cdot F(\theta)$ is the transform of the convolution of $\{h_n\}$ with $\{f_n\}$:

$$f_n^\star = (h * f)_n = \sum_{k=-\infty}^{\infty} h_k f_{n-k}. \tag{25}$$

This enables the filtering to be performed directly on the given sequence $\{f_n\}$. In practice the summation must be truncated at some finite value of $k$. Thus, an

approximation to the low frequency part of $\{f_n\}$ is given by

$$f_n^\star = \sum_{k=-N}^{N} h_k f_{n-k}. \tag{26}$$

A more sophisticated method uses the Chebyshev alternation theorem to obtain a filter whose maximum error in the pass- and stop-bands is minimized. This method yields a filter meeting required specifications with fewer coefficients that the other methods. The design of non-recursive and recursive filters is outlined in Hamming (1989), where several methods are described, and fuller treatments may be found in Oppenheim and Schafer (1989).

## *5.2 Application of a Non-recursive Digital Filter to Initialization*

An initialization scheme using a non-recursive digital filter has been developed by Lynch and Huang (1992) for the HIRLAM (High Resolution Limited Area Model) model. The uninitialized fields of surface pressure, temperature, humidity and winds were first integrated forward for 3 h, and running sums of the form

$$f_F^\star(0) = \frac{1}{2} h_0 f_0 + \sum_{n=1}^{N} h_{-n} f_n, \tag{27}$$

where $f_n = f(n\Delta t)$, were calculated for each field at each gridpoint and on each model level. These were stored at the end of the 3 h forecast. The original fields were then used to make a 3 h "hindcast", during which running sums of the form

$$f_B^\star(0) = \frac{1}{2} h_0 f_0 + \sum_{n=-1}^{-N} h_{-n} f_n \tag{28}$$

were accumulated for each field, and stored as before. The two sums were then combined to form the required summations:

$$f^\star(0) = f_F^\star(0) + f_B^\star(0). \tag{29}$$

These fields correspond to the application of the digital filter Eq. (26) to the original data, and will be referred to as the filtered data.

Complete technical details of the original implementation of DFI in the HIRLAM model may be found in Lynch et al. (1999). A reformulation of the implementation, with further testing and evaluation, is presented in Huang and Yang (2002).

### *5.3 Initialization Example*

A detailed case study based on the implementation in HIRLAM was carried out to check the effect of the initialization on the initial fields and on the forecast, and to examine the efficacy of DFI in eliminating high frequency noise. The digital filter initialization was compared to the reference implicit normal mode initialization (NMI) scheme, and to forecasts with no initialization (NIL). Forecasts starting from the analysis valid at 1200 UTC on 10 February, 1999 were compared.

We first checked the effect of DFI on the analysis and forecast fields. The maximum change in surface pressure due to initialization was 2.2 hPa, with a RMS (root-mean-square) change of about 0.5 hPa. The changes to the other analysed variables were in general comparable in size to analysis errors, and considerably smaller in magnitude than typical changes brought about by the analysis itself: the RMS change in surface pressure from first guess to analysis was about 1 hPa. The RMS and maximum differences between the uninitialized 24-h forecast (NIL) and the filtered forecast (DFI) for all prognostic variables were examined. When we compare these values to the differences at the initial time they were seen to be generally smaller. The changes made by DFI are to the high frequency components; since these are selectively damped during the course of the forecast, the two forecasts were very similar. After 24-h the maximum difference in surface pressure was less than 1 hPa and the RMS difference is only 0.1 hPa.

The basic measure of noise is the mean absolute value of the surface pressure tendency

$$N_1 = \left(\frac{1}{\mathrm{N}}\right) \sum_{n=1}^{N} \left| \frac{\partial p_s}{\partial t} \right| .$$

For well-balanced fields this quantity has a value of about 1 hPa/3 h. For uninitialized fields it can be an order of magnitude larger. In Fig. 4 we plot the value of $N_1$ for three forecasts. The solid line represents the forecast from uninitialized data: we see that the value of $N_1$ at the beginning of the forecast is about 12 hPa/3 h. This large value reflects the lack of an effective multivariate balance in the analysis. It takes about 6 h to fall to a reasonable value. The dashed line is for a forecast starting from data initialized using the implicit normal mode method (NMI). The starting value is about 3 hPa/3 h, falling to about 1.5 hPa/3 h after 12 h. The final graph (the dotted line) is for the digitally filtered data (DFI). The initial value of $N_1$ is now about 1.5, and remains more or less constant throughout the forecast. It is clear from this measure that DFI is more effective in removing high frequency noise than NMI.

The measure $N_1$ indicates the noise in the vertically integrated divergence field. However, even when this is small, there may be significant activity in the internal gravity wave modes. To see this, we look at the vertical velocity field at 500 hPa for the NIL and DFI analyses. The left panel in Fig. 5 shows the uninitialized vertical velocity field, zoomed in over western Europe and the eastern North Atlantic.

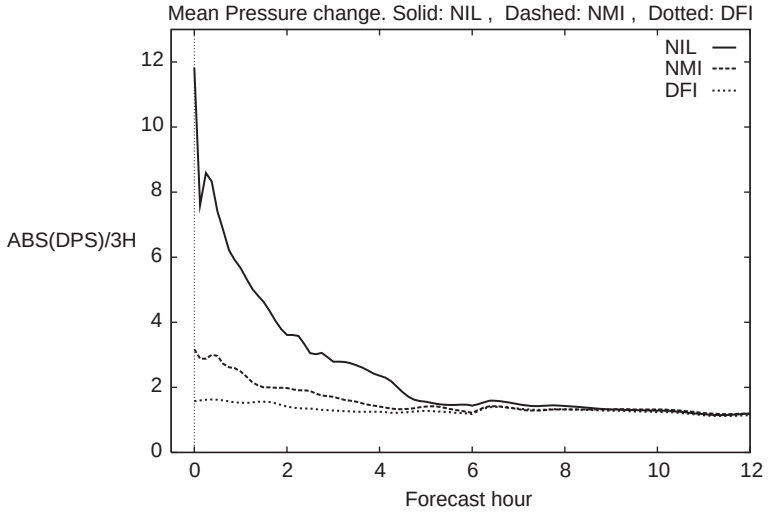



**Fig. 4** Mean absolute surface pressure tendency for three forecasts. *Solid*: uninitialized analysis (NIL). *Dashed*: Normal mode initialization (NMI). *Dotted*: Digital filter initialization (DFI). Units are hPa/3 h.

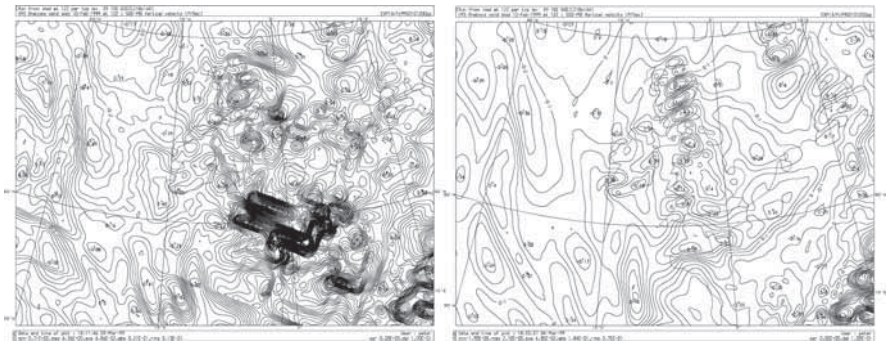


**Fig. 5** Vertical velocity at 500 hPa over western Europe and the eastern North Atlantic. (*Left*) Uninitialized analysis (NIL); (*Right*) after digital filtering (DFI)

There is clearly substantial gravity wave noise in this field. In fact, the field is physically quite unrealistic. The right panel shows the DFI vertical velocity. It is much smoother; the spurious features have been eliminated and the large values with small horizontal scales which remain are clearly associated with the Scottish Highlands, the Norwegian Mountains and the Alps. Comparison with the NMI method (see Lynch et al. 1999, for details) indicates that DFI is more effective than NMI in dealing with internal gravity wave noise. It is noteworthy that stationary mountain waves are unaffected by digital filtering, since they have zero frequency. This is a desirable characteristic of the DFI scheme.

### *5.4 Benefits for the Data Assimilation Cycle*

In Lynch et al. (1999), a parallel test of data for one of the FASTEX (Fronts and Atlantic Storm Track EXperiment) intensive observing periods showed that the DFI method resulted in slightly improved scores compared to NMI. As it is not usual for an initialization scheme to yield significant improvements in forecast accuracy, some discussion is merited. We cannot demonstrate beyond question the reason for this improvement. However, the comparative results showed up some definite defects in the implicit normal mode initialization as implemented in the reference HIRLAM model. It was clear that the NMI scheme did not eliminate imbalance at lower model levels. Moreover, although the noise level indicated by the parameter $N_1$ fell to a reasonable level in 6 h, there was still internal gravity wave noise, not measured by this parameter. Any noise in the 6 h forecast will be carried through to the next analysis cycle, and will affect the quality control and assimilation of new observational data. It is believed that the DFI scheme, with its superior ability to establish atmospheric balance, results in improved assimilation of data and consequently in a reduction of forecast errors.

## 6 Constraints in 4D-Var

We conclude with a discussion on the application of a digital filter as a weak constraint in four-dimensional variational assimilation (4D-Var; see chapter *Variational Data Assimilation*, Talagrand). The idea is that if the state of the system is noise-free at a particular time, i.e., is close to the slow manifold, it will remain noise-free, since the slow manifold is an invariant subset of phase-space (Leith 1980). We consider a sequence of values $\{x_0, x_1, x_2, \cdots x_N\}$ and form the filtered value

$$\bar{x} = \sum_{n=0}^{N} h_n x_n. \tag{30}$$

The evolution is constrained, so that the value at the mid-point in time is close to this filtered value, by addition of a term

$$J_c = \frac{1}{2}\gamma ||x_{N/2} - \bar{x}||^2$$

to the cost function to be minimized ($\gamma$ is an adjustable parameter). It is straightforward to derive the adjoint of the filter operator (Gustafsson 1992). Gauthier and Thépaut (2001) applied such a constraint to the 4D-Var system of Météo-France. They found that a digital filter weak constraint imposed on the low-resolution increments efficiently controlled the emergence of fast oscillations while maintaining a close fit to the observations. As the values required for input to the filter are already available, there is essentially no computational overhead in applying this procedure.

The dynamical imbalance was significantly less in 4D-Var than in 3D-Var (three-dimensional variational assimilation). Wee and Kuo (2004) included a $J_c$ term in the MM5 4D-Var. They found that the weak contraint not only reduces the dynamic imbalance in the 4D-Var solution, but also improves the quality of the analysis and forecast significantly.

To illustrate the impact of the $J_c$ constraint, experiments are carried out using the WRF (Weather Research and Forecasting) 4D-Var system Huang et al. (2009), (a) without and (b) with the penalty term in the minimization. The cost functions ($J_o$, $J_b$ and $J_c$) are shown in Fig. 6. In both panels, $J_c$ is computed with $\gamma = 0.1$. It is clear that the unconstrained 4D-Var analysis contains a significant amount of noise, with $J_c$ large, and the weak constraint $J_c$ is able to control the noise level. In most of our experiments, $J_c$ also helps the convergence of the minimization.

To further demonstrate the noise control effect of $J_c$, we computed $N_1$ during the forecasts from WRF 4D-Var analyses using different $\gamma$. The results from five experiments are shown in Fig. 7. NoJcDF: forecast start from a WRF 4D-Var analysis without $J_c$ or $\gamma = 0$. JcDF(0.1): forecast start from a WRF 4D-Var analysis with $J_c$ and $\gamma = 0.1$. JcDF(1): forecast start from a WRF 4D-Var analysis with $J_c$ and $\gamma = 1$. JcDF(10): forecast start from a WRF 4D-Var analysis with $J_c$ and

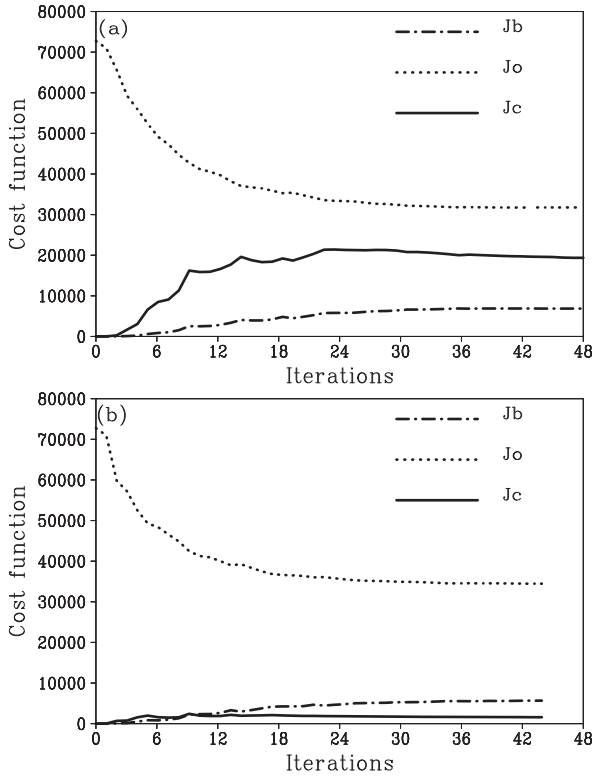



**Fig. 6** Cost functions for experiment (**a**) without $J_c$ and (**b**) with $J_c$ in the minimization

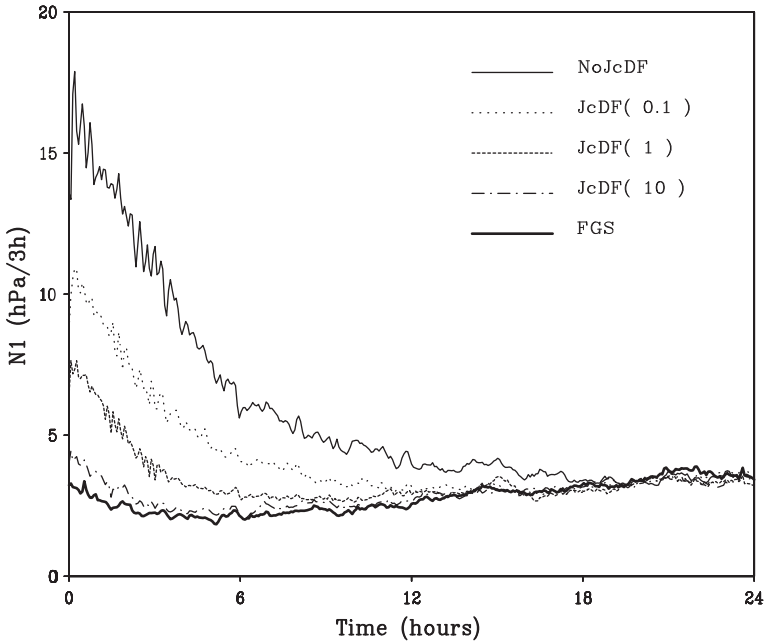



**Fig. 7** Mean absolute surface pressure tendency for 5 forecasts

$\gamma = 10$. FGS: forecast start from first guess, which is a 3-h forecast from previous analysis cycle and can be considered as a noise-free forecast. The larger the weight we assign to $J_c$ in the 4D-Var minimization, the lower the noise level becomes in the subsequent forecast. However, a larger weight in $J_c$ may compromise the fit to observations (a larger $J_o$ at the end of minimization). The tuning of $\gamma$ is necessary.

## 7 Conclusion

We have described several methods of eliminating noise from the forecast by removal of spuriously large-amplitude gravity-wave components from the initial data. This is essential for practical reasons and, in particular, for avoidance of problems in the assimilation cycle. The benefits of initialization are clear. However, it is noteworthy that modern variational assimilation methods are capable of producing fields in good balance, so that a separate initialization stage is less important now. Constraints to ensure good balance can be incorporated directly into variational assimilation schemes. The digital filter method is particularly attractive in this respect, and is a natural choice for variational analysis.

## References


Baer, F., 1977. Adjustment of initial conditions required to suppress gravity oscillations in nonlinear flows. *Beitr. Phys. Atmos.*, **50**, 350–366.

Bubnová, R., G. Hello, P. Bénard and J.-F. Geleyn, 1995. Integration of the fully elastic equations cast in the hydrostatic pressure terrain-following coordinate in the framework of the ARPEGE/Aladin NWP system. *Mon. Weather Rev.*, **123**, 515–535.

Charney, J.G., 1955. The use of the primitive equations of motion in numerical prediction. *Tellus*, **7**, 22–26.

Charney, J.G., R. Fjørtoft and J. von Neumann, 1950. Numerical integration of the barotropic vorticity equation. *Tellus*, **2**, 237–254.

Chen, M. and X.-Y. Huang, 2006. Digital filter initialization for MM5. *Mon. Weather Rev.*, **134**, 1222–1236.

Daley, R., 1991. *Atmospheric Data Assimilation*. Cambridge University Press, Cambridge, 457pp.

Gauthier, P. and J.-N. Thépaut, 2001. Impact of the digital filter as a weak constraint in the preoperational 4D-Var assimilation system of Météo-France. *Mon. Weather Rev.*, **129**, 2089–2102.

Gustafsson, N., 1992. Use of a digital filter as weak constraint in variational data assimilation. *Proceedings of the ECMWF Workshop on Variational Assimilation, with Special Emphasis on Three-Dimensional Aspects*, pp 327–338. Available from the European Centre for Medium Range Weather Forecasting, Shinfield Park, Reading, Berks. RG2 9AX, UK.

Hamming, R.W., 1989. *Digital Filters*. Prentice-Hall International, Inc., Englewood Cliffs, NJ, 284pp.

Haurwitz, B., 1940. The motion of atmospheric disturbances on the spherical earth. *J. Marine Res.*, **3**, 254–267.

Hinkelmann, K., 1951. Der Mechanismus des meteorologischen Lärmes. *Tellus*, **3**, 285–296.

Hinkelmann, K., 1959. *Ein numerisches Experiment mit den primitiven Gleichungen*. Vol. Rossby Memorial Volume. Rockerfeller lnstitute Press, New York, NY, pp 486–500.

Holton, J.R., 1992. *An Introduction to Dynamic Meteorology*, 3rd edition. International Geophysics Series, vol. 48. Academic Press, San Diego. Chap. 7.

Huang, X.-Y. and P. Lynch, 1993. Diabatic digital filter initialization: Application to the HIRLAM model. *Mon. Weather Rev.*, **121**, 589–603.

Huang, X.-Y., Q. Xiao, D.M. Barker, X. Zhang, J. Michalakes, W. Huang, T. Henderson, J. Bray, Y. Chen, Z. Ma, J. Dudhia, Y. Guo, X. Zhang, D.-J. Won, H.-C. Lin and Y.-H. Kuo, 2009. Four-dimensional variational data assimilation for WRF: Formulation and preliminary results. *Mon. Weather Rev.*, **137**, 299–314.

Huang, X.-Y. and X. Yang, 2002. *A New Implementation of Digital Filtering Initialization Schemes for HIRLAM.* Technical Report 53, 36pp. Available from HIRLAM-5, c/o Per Undén, SMHI, S-60176 Norrköping, Sweden.

Kasahara, A., 1976. Normal modes of ultralong waves in the atmosphere. *Mon. Weather Rev.*, **104**, 669–690.

Leith, C.E., 1980. Nonlinear normal mode initialization and quasi-geostrophic theory. *J. Atmos. Sci.*, **37**, 958–968.

Lynch, P., 1997. The Dolph-Chebyshev window: A simple optimal filter. *Mon. Weather Rev.*, **125**, 655–660.

Lynch, P., 2006. *The Emergence of Numerical Weather Prediction: Richardson's Dream*. Cambridge University Press, Cambridge, 279pp.

Lynch, P., D. Giard and V. Ivanovici, 1997. Improving the efficiency of a digital filtering scheme. *Mon. Weather Rev.*, **125**, 1976–1982.

Lynch, P. and X.-Y. Huang, 1992. Initialization of the HIRLAM model using a digital filter. *Mon. Weather Rev.*, **120**, 1019–1034.

Lynch, P., R. McGrath and A. McDonald, 1999. *Digital Filter Initialization for HIRLAM*. HIRLAM Technical Report 42, 22pp. Available from HIRLAM-5, c/o Per Undén, SMHI, S-60176 Norrköping, Sweden.

Machenhauer, B., 1977. On the dynamics of gravity oscillations in a shallow water model with applications to normal mode initialization. *Beitr. Phys. Atmos.*, **50**, 253–271.

McIntyre, M.E., 2003. *Balanced Flow*. Vol. Encyclopedia of Atmospheric Sciences, J.R. Holton, J. Pyle, and J.A. Curry (eds.), 6 vols, ISBN 0-12-227090-8. Academic Press, London.

Miyakoda, K. and R.W. Moyer, 1968. A method for initialization for dynamic weather forecasting. *Tellus*, **20**, 115–128.

Oppenheim, A.V. and R.W. Schafer, 1989. *Discrete-Time Signal Processing*. Prentice-Hall International, Inc., Englewood Cliffs, NJ, 879pp.

Phillips, N.A., 1960. On the problem of initial data for the primitive equations. *Tellus*, **12**, 121–126.

Phillips, N.A., 1973. *Principles of Large Scale Numerical Weather Prediction*. Vol. Dynamic Meteorology, P. Morel (ed.). D. Reidel, Dordrecht, pp 1–96.

Richardson, L.F., 1922. *Weather Prediction by Numerical Process*. Cambridge University Press, Cambridge, 236pp. Reprinted by Dover Publications, New York, 1965.

Rossby, C.G., 1939. Relations between variations in the intensity of the zonal circulation of the atmosphere and the displacements of the semipermanent centers of action. *J. Marine Res.*, **2**, 38–55.

Sasaki, Y., 1958. An objective method based on the variational method. *J. Met. Soc. Jpn*, **36**, 77–88.

Wee, T.-K. and Y.-H. Kuo, 2004. Impact of a digital filter as a weak constraint in MM5 4DVAR: An observing system simulation experiment. *Mon. Weather Rev.*, **132**, 543–559.

Williamson, D. and C. Temperton, 1981. Normal mode initialization for a multilevel gridpoint model. Part II: Nonlinear aspects. *Mon. Weather Rev.*, **109**, 745–757.

# Part II
# Observations

# The Global Observing System

**Jean-Noël Thépaut and Erik Andersson**

## 1 Introduction

In this chapter we describe the main components of what is commonly known as the World Weather Watch Global Observing System (GOS), and review the different techniques to observe the atmosphere, the ocean and land surfaces. It should be stressed that the various observing systems generally tend to be complementary to one another, and that redundancy where it exists is valuable as it enables cross checking and inter-comparison of data. The emphasis is on the main observation types and those regularly used in Numerical Weather Prediction (NWP) systems. It thus complements the chapter *Assimilation of Operational Data* (Andersson and Thépaut) on one hand, which concentrates on the assimilation of operational data, and the chapter *Research Satellites* (Lahoz) on the other hand, which provides an overview of available and forthcoming research satellites. The different types of observations are here divided into two broad categories: in situ observations and remote sensing observations. We shall see that the different observing systems have different characteristics that need to be accounted for in assimilation of the data.

A number of acronyms are used in this chapter. The full list of acronyms is provided in the *Appendix*.

## 2 In Situ Observations

Those observation types that were in general use before the satellite era are sometimes referred to as "conventional observations". For the most part they are in situ measurements of meteorological parameters such as temperature, wind, pressure and humidity. In situ observations are generally considered to be point-wise and instantaneous, which are generally accurate assumptions in the context of operational NWP where the assimilating models typically have resolutions of

J.-N. Thépaut (✉)
European Centre for Medium-Range Weather Forecasts, ECMWF, Shinfield, UK
e-mail: jean-noel.thepaut@ecmwf.int

W. Lahoz et al. (eds.), *Data Assimilation*, DOI 10.1007/978-3-540-74703-1_10,

50 km or less in the horizontal, from a few 100 m (in the stratosphere) to a few tens of metres or less (in the boundary layer) in the vertical, and a few hours temporally. Each instrument and measurement technique is nevertheless associated with its own space- and time-scales due to its sampling characteristics. Instruments that travel through the atmosphere attached to a balloon or an aircraft may average over a relatively long distance in a short time, whereas stationary instruments may sample the same small volume of air over a longer period of time. Some in situ observations provide integrated information of the entire atmospheric column.

## *2.1 Surface and Marine Observations*

Near-surface measurements of temperature, wind and humidity have been made for centuries, forming an invaluable record of the climate (Compo et al. 2006), and providing a cornerstone for NWP. Stevenson screens are used to house thermometers to measure temperature at a standard height of 2 m and a wet bulb thermometer to determine dewpoint depression and hence humidity. A barometer, usually placed indoors, measures the air pressure. Wind velocity is measured at a standard height of 10 m with an anemometer. To aid comparison of pressure measurements from different stations they are normally adjusted to mean sea level, which is problematic in mountainous areas; assumptions have to be made about the temperature profile of the fictitious air column that would extend from the station location down to the sea level. In a further effort to aid comparison of measurements it is required that the surface observations are made simultaneously at four specific times during the day: the main so-called synoptic times are 0000, 0600, 1200 and 1800 UTC. Some stations are making 3-hourly observations, predominantly during day-time. As more stations are being automated hourly data are becoming more widely available. At the same time as reading the instruments, the observer (at manual stations) makes visual observations of clouds, visibility, and current weather. While all these data are very valuable to forecasters, it is still hard to use all parts of a surface observation report in a data assimilation system designed for NWP (see chapter *Numerical Weather Prediction*, Swinbank) as due to resolution and physics limitations, numerical models do not represent these observables very well. However, the situation is evolving, especially with the progress made in the development of very high resolution regional NWP systems.

Surface observations (SYNOP) are available over much of the densely populated regions of the world, particularly in the Northern Hemisphere, although there are extensive data voids over parts of Africa (Fig. 1, red markers). Surface observations from airports (METAR, available during the airfields' hours of operation) are shown with blue markers. Similar types of observations are also made from many ships, which helps fill the gaps over those parts of the ocean that are well covered by commercial shipping routes (cyan markers in Fig. 1). In recent years many drifting (and a few moored) buoys have been deployed to help fill the data voids (Fig. 2), not least

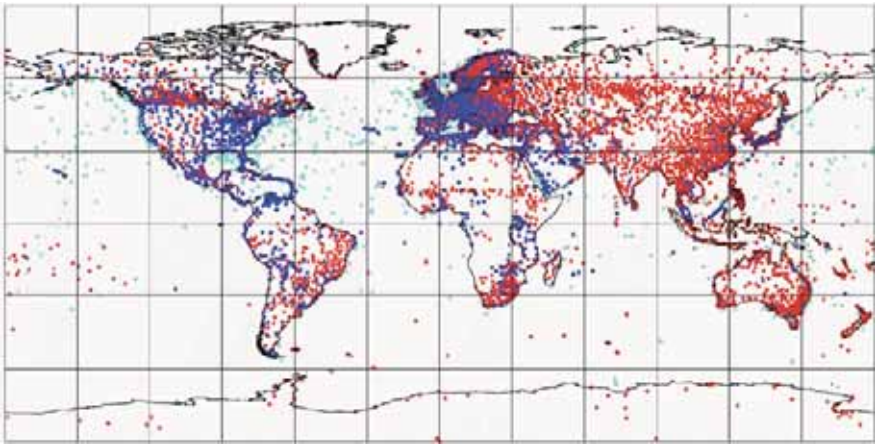


**Fig. 1** Typical data coverage of surface observations, 20070301 0900-1500 UTC, showing 16,550 SYNOP (*red*), 1,937 SHIP (*cyan*) and 12,383 METAR (*blue*)

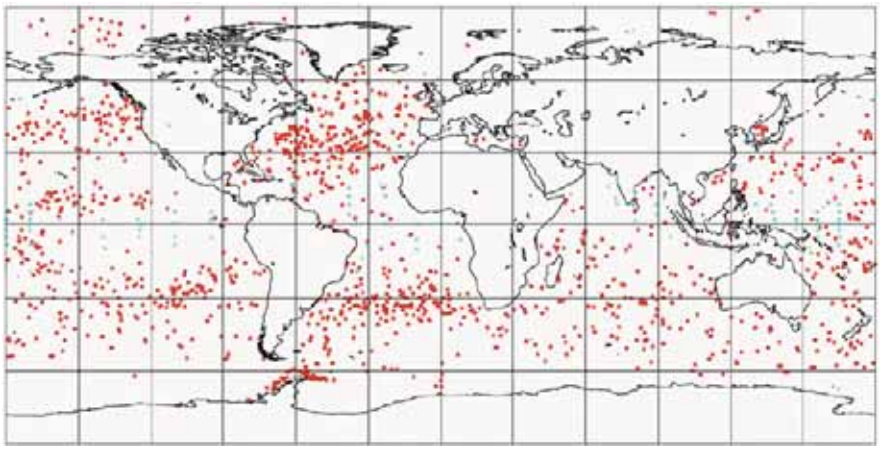


**Fig. 2** Typical data coverage of buoy observations, 20070301 0900-1500 UTC, showing 5,686 drifting buoys (*red*) and 140 moored buoys (*cyan*)

in the southern oceans. The buoys provide frequent surface pressure observations (hourly or in some cases every 10 min) which is particularly valuable to determine the intensification rate and movement of storms (Järvinen et al. 1999).

## *2.2 Radiosondes*

First attempts at making observations of the upper atmosphere (in this context, the free troposphere, i.e., above the boundary layer) were made during the second half of the nineteenth century. Labitzke and van Loon (1999) give some fascinating

accounts of early exploration of upper levels of the atmosphere, with particular emphasis on discoveries related to the stratosphere. Radiosondes came to be a crucial part of the Global Observing System following the International Geophysical Year, or IGY (1957).

Radiosondes are generally launched twice a day (at 0000 and 1200 UTC) from radiosonde stations across the world. As in the case of surface data, the stations tend to be concentrated in the main populated areas of the Northern Hemisphere (Fig. 3). Each radiosonde consists of a balloon carrying an instrument package, from which measurements are relayed to the ground station. The instrument package makes in situ measurements of pressure, temperature and humidity. Radar, or more recently GPS (Global Positioning System) navigation, is used to track the balloon and so ascertain the wind at the height of the balloon. In a radiosonde sounding, weather elements are reported at standard pressure levels. The reports also include "significant levels" to allow details of the measurement profile to be reconstructed between the standard pressure levels.

Radiosondes are a crucial part of the observation network. They are still heavily used by forecasters, particularly in developing countries. They make a major contribution to the NWP forecast performance (Bouttier and Kelly 2001), primarily because it is more difficult to use satellite data over land than ocean. Radiosondes are also essential for the calibration and bias correction of satellite data.

Since there are many fewer radiosonde observations than surface data, the data voids are even more severe. Some radiosonde ascents are also made from special weather ships, but, because of their high cost, they are being replaced to some degree by ASAP (Automated Shipboard Aerological Programme) systems that can automatically launch radiosonde balloons from commercial ships (cyan markers, Fig. 3). Radiosondes are also complemented by pilot balloons (red markers, Fig. 4),

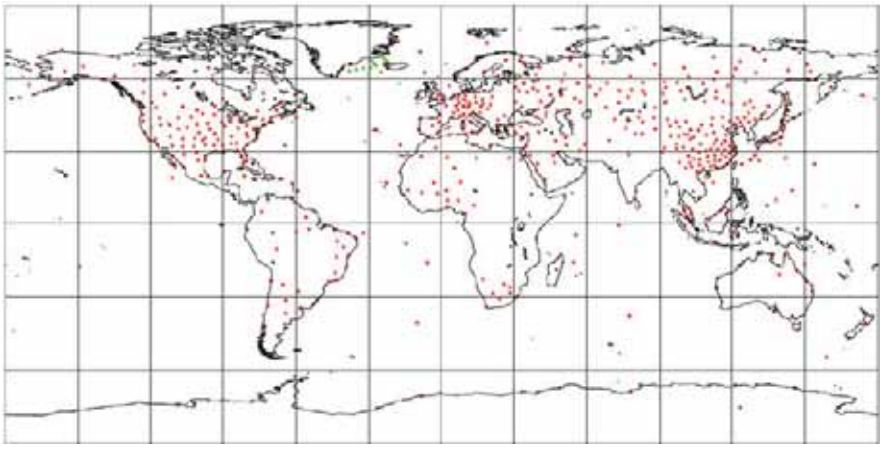


**Fig. 3** Typical data coverage of radiosonde observations, 20070301 0900-1500 UTC, showing 580 land stations (*red*), 8 ships (*cyan*) and 16 dropsondes (*green*)

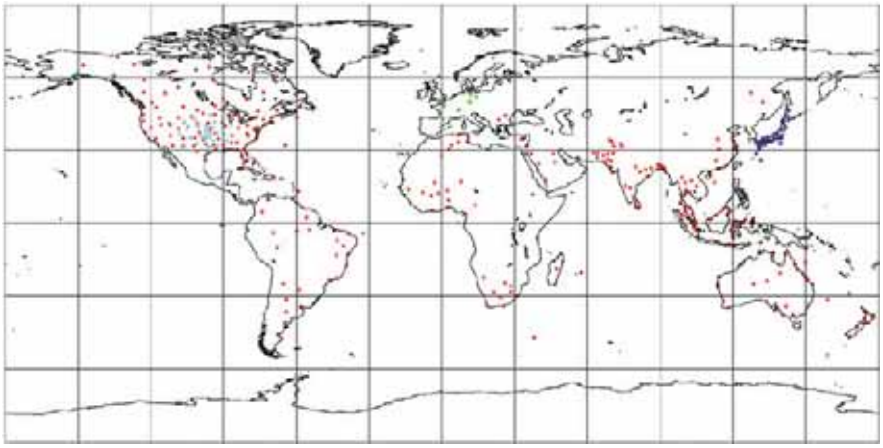


**Fig. 4** Typical data coverage of pilot balloons and profiler observations, 20070301 0900-1500 UTC, showing 305 PILOTs (*red*), 170 American (*cyan*), 128 European (*green*) and 186 Japanese (*blue*) profilers

which are launched simply to measure wind profiles, without carrying an instrument package.

Dropsonde observations are similar to radiosondes, except that the instrument packages are dropped from aircraft rather than flown from a balloon. These are often used in experimental campaigns, rather than in routine operations. They are also employed when targeting observations (see Sect. 4). The example in Fig. 3 shows some dropsonde data (green markers) deployed from an aircraft in the area between Iceland and Greenland.

## *2.3 Aircraft Observations*

Many commercial aircraft make in situ measurements of temperature, wind and pressure during their flights. The wind measurements need input from the navigation system since the wind is the difference between the aircraft's ground velocity and its air speed. Traditionally, aircraft observations are reported by the pilot at particular locations along the aircraft routes, e.g., at specific longitude crossings in the Atlantic as seen from the red markers in Fig. 5, labelled AIREP. More recently there have been very significant developments in establishing automatic transmission of observations from the aircraft (green and cyan markers, Fig. 5), under co-ordination of the AMDAR (Aircraft Meteorological Data Relay) programme of the WMO (World Meteorological Organization). The aircraft data are irregularly distributed over the globe with dense concentrations over the United States, Europe and along the main intercontinental air traffic routes. AMDAR reports are often produced at the specified frequency of one report per 7 min at cruise level, with additional reports at wind maxima. During ascent reporting is at 10 hPa intervals vertically for the first

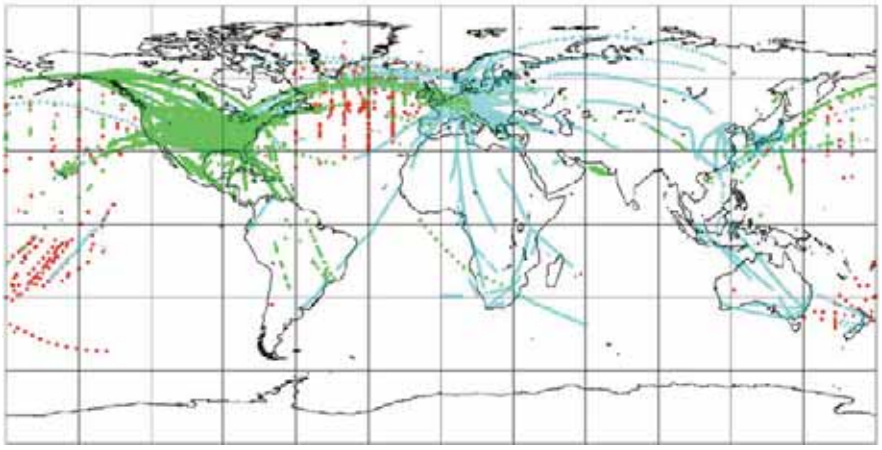


**Fig. 5** Typical data coverage of aircraft observations, 20070301 0900-1500 UTC, showing 3,175 AIREP (*red*), 22,979 AMDAR (*cyan*) and 31,587 ACAR (*green*)

100 hPa in the lower part of the profile and every 50 hPa above that layer to top of climb (around the tropopause and above) with the reverse applying during the descent phase. The AMDAR system thus provides data at altitude roughly every 70–100 km along the flight path as well as detailed profiles in the near vicinity of airports.

The European component of AMDAR is managed by EUCOS (The EUMETNET (The Network of European Meteorological Services) Composite Observing System). In an effort to optimize the benefit and the value-for-money of the European AMDAR programme, EUCOS has developed an elaborate and effective data collection strategy.

## *2.4 Targeted Observing*

The quality of numerical forecasts depends on the accuracy of initial conditions and consequently it depends on the full composite of the GOS. Observation targeting has been proposed as a cost-effective approach to complement the GOS with additional observations where they are most needed, and where they would have the greatest impact. Methods have been proposed and developed that make it possible to identify in advance regions of the atmosphere where forecasts are particularly sensitive to errors in initial conditions. If those areas are not well covered by routine observations it would be advantageous to make additional observations there.

Over the past decade many field campaigns took place and extensive work has been done by NWP centres to assess the impact on the forecast of the extra observations taken in specific, case-dependent target areas which were identified using objective and subjective methods (Langland 2005). The campaigns include, for

example, in 1997 FASTEX (Fronts and Atlantic Storm-Track Experiment), in 1998 NORPEX (NORth-Pacific Experiment) and CALJET (California Land-falling JETs experiment), in 1999 and 2000 the Winter Storm Reconnaissance programs (WSR99 and WSR00) and in 2003 NA-TReC. The overall conclusion was that, on average, targeted observations do increase the forecast skill by up to 10–15% in some cases over, e.g., Europe, but averaged over larger samples, the impact is quite small. For example, the use of targeted wind profile observations from an airborne Doppler Wind Lidar (DWL) instrument flying during the NA-TReC showed forecast skill improvements of about 3% in the verification area over Europe from forecast day 2 to 4 (Weissmann and Cardinali 2006).

Apart from identifying the target area, targeted observing also involves very significant logistical problems associated with the deployment of additional observations in remote areas, often located within the main data voids over the ocean. So far, deployment of dropsondes from specially equipped aircraft has been the most common approach. The aircraft has to be available on standby at strategic locations such as Alaska, Hawaii, Iceland or Ireland, for flights over the North Pacific and the North Atlantic, respectively. An example of targeted dropsonde data in the Iceland area are shown in Fig. 3 (green markers). Alternative observing systems that lend themselves to targeting include driftsondes, airborne DWL and DIAL (lidars for wind and humidity, respectively) and unmanned aircraft. During the NA-TReC, targeted collection of AMDAR data was developed and trialled, as well as commanding additional radiosonde launches from ships and regular stations in close vicinity of the target area. In future, it may also be feasible to enhance the sampling and collection of satellite data in sensitive areas on demand.

## 3 Remote Sensing Observations

The Space-based Global Observing System complements the in situ Observing System to provide the World Weather Watch's GOS. The main providers of Earth Observation satellite systems and space-based observations of the atmosphere for NWP centres are the American (NASA and NOAA), European (ESA and EUMETSAT) and Japanese (JAXA and JMA) space agencies. Other Earth Observation satellite systems are operated by the Russian Federation, People's Republic of China, the Indian Space Agency and other national space agencies.

Research and Development (R&D) Space Agencies (NASA, ESA, JAXA) usually promote demonstration missions, with innovative technologies, thus paving the way for future long-term operational missions (see the chapter *Research Satellites*, Lahoz, for a comprehensive review of the missions provided by R&D Space Agencies). The primary goal of R&D Space Agencies is in principle not the delivery of near real time products (typically 3 h or less) to the community. However, R&D satellite missions prove to be crucial to better characterize diverse features of the model (e.g. UARS, POLDER), develop new methodologies in view of assimilation

of future operational instruments (e.g. AIRS on AQUA, TRMM), or sometimes just improve the quality of the operational NWP assimilation system when data are of good quality and sufficient timeliness (e.g. ERS-2, QuikScat, AIRS, Envisat).

Operational Space Agencies (NOAA, EUMETSAT, JMA) operate instruments inherited from demonstration missions. Operational systems ensure a stabilized long-life mission technology (e.g. the HIRS instrument onboard NOAA satellites will have lasted more than 30 years), which eases the investment decisions at the NWP community end. Operational missions, moreover, ensure robustness in the processing chain and time delivery to the end-users in agreement with their requirements. Today, they constitute the backbone of the Global Observing System and provide the major part of the data currently assimilated in NWP.

Operational agencies ensure the long-term continuity of the operational systems in polar as well as geostationary orbits. Both ways of observing the Earth/atmosphere are very complementary. Geostationary platforms (GEOs), located at 36,000 km on an equatorial plane, orbit the Earth with the same angular velocity as the Earth and therefore provide an almost continuous view (repetition time of down to a few minutes) of the same part of the Earth. The high temporal resolution of the GEOs makes them essentially suitable for nowcasting applications, but also for NWP four-dimensional data assimilation systems through the provision of Atmospheric Motion Winds derived from cloud tracking or sequences of radiance data. The orbit geometry of GEOs makes them unable to observe Polar Regions. Figure 6 displays the current constellation of geostationary satellites currently assimilated at ECMWF (European Centre for Medium-Range Weather Forecasts). Low Earth Orbiting (LEO) satellites, at least the operational ones, orbit the Earth at around 800 km with a repetition time over the pole of about 100 min. Being closer to the atmosphere than the GEOs, these satellites are more suitable to

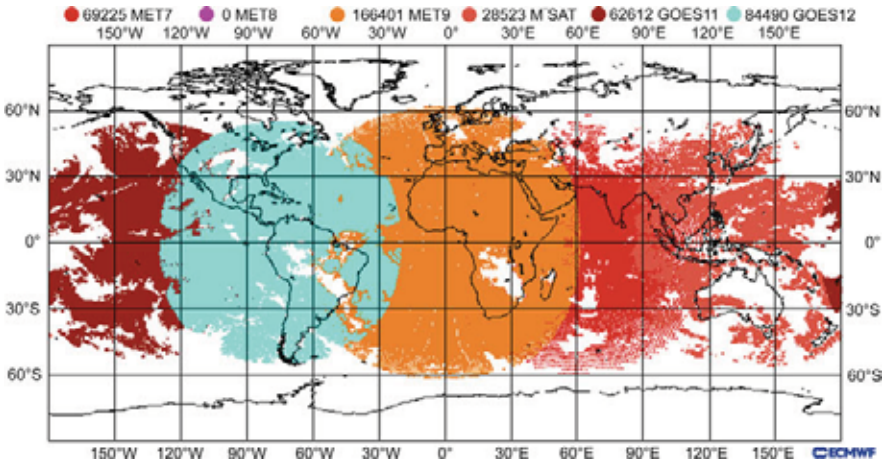



**Fig. 6** Typical data coverage provided by the Geostationary constellation: GOES-11 (*brown*), GOES-12 (*cyan*), Meteosat-7 (*red*), Meteosat-9 (*orange*) and MTSAT (*red–orange*)

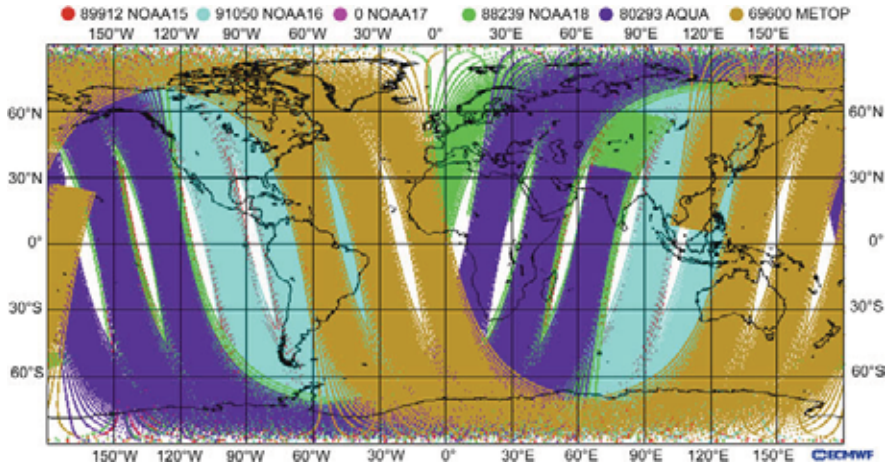



**Fig. 7** Typical data coverage provided by the LEO (Low Earth Orbit) constellation of AMSU-A instruments from NOAA, AQUA and METOP satellites: NOAA-15 in *red*, NOAA-16 in *cyan*, NOAA-18 in *green*, AQUA in *violet* and METOP in *brown*

sound the atmosphere. Figure 7 displays the constellation of the operational AMSU-A instruments on board NOAA, AQUA and METOP satellites. In both cases (GEOs and LEOs), a constellation of satellites is required to provide an adequate global coverage. This is currently achieved by the current respective operational satellite constellations.

## *3.1 Passive Technologies*

As mentioned in Thépaut (2003) and further discussed in the chapter *Assimilation of Operational Data* (Andersson and Thépaut) it is important to realize what is specific to satellite observations and in particular what they actually measure. Contrary to conventional in situ observations such as aircraft of radiosonde measurements, the quantities measured by satellite instruments do not relate directly to geophysical quantities. Satellite instruments do not measure temperature, do not measure humidity and do not measure wind. Satellite instruments measure essentially the radiation that reaches the top of the atmosphere at given frequencies. The key to using satellite observations lies in the data assimilation techniques to infer meteorological information from these radiance measurements; see the companion chapters in Part I, *Theory*.

Eyre (2000) provides an excellent overview of the different instrument technologies commonly used to observe the atmosphere from space, and a brief summary is given here. By selecting radiation at different frequencies (or channels), a satellite instrument can provide information on a range of geophysical variables (e.g. upper air temperature, moisture, ozone, surface parameters, clouds).

A distinction has to be made between passive and active instruments. Passive instruments sense radiation emitted by the surface and/or atmosphere (or the solar

radiation reflected by it). Active instruments emit radiation and measure how much of it is reflected or back-scattered by the surface and/or atmosphere. In general, the wavelengths (or channels) available from the most commonly used satellite instruments may be considered as one of 3 different types.

### 3.1.1 Atmospheric Sounding Channels from Passive Instruments

Sounding radiometers (infrared or microwave) sense primarily in strong gaseous absorption bands and measure the radiation emitted by molecules of these gases in the atmosphere. At a wavelength where the atmosphere is opaque, the instrument will measure radiation emanating high up in the atmosphere, while at another wavelength where the atmosphere is more transparent, the instrument will sense radiation emanating from the atmosphere at lower altitude. It is therefore a careful choice of the wavelengths detected that will allow the retrieval of atmospheric profile information. Emission from gases of known concentration (e.g. carbon dioxide or oxygen) will provide information on temperature profiles, while measurements from gases of variable concentration (e.g. ozone, water vapour) provide information on the mixing ratio of these gases. In the infrared, sounding radiometers are limited by the presence of clouds (especially opaque clouds), therefore no atmospheric information can be retrieved below the cloud. Information about the cloud cover and cloud top can, however, be obtained. Microwave sounding radiometers are less sensitive to clouds (except in precipitating areas) and are therefore complementary to infrared sounders in providing atmospheric profile information in all weather conditions.

Instruments of this type include HIRS (High Resolution Infrared Radiation Sounder) and AMSU (Advanced Microwave Sounding Unit) on NOAA and METOP satellites, and also SSMIS (Special Sensor Microwave Sounder/Imager) on DMSP (Defense Meteorological Satellite Program) satellites. Figure 8 represents the AMSU-A temperature "weighting functions" (indicating the altitude each AMSU-A channel is sensitive to), showing that a reasonable vertical sampling of the atmosphere can be achieved from an appropriate selection of channels with varying absorption strengths.

The vertical resolution of these classical sounders is rather low (around 3 km). However, the new generation of infrared sounders, such as the AIRS (Atmospheric InfraRed Sounder) on NASA's AQUA satellite and the IASI interferometer on EUMETSAT's METOP satellite, have somewhat changed the picture. Those instruments measure radiation in several thousands of different channels, and therefore provide atmospheric temperature and composition information at a much higher accuracy and vertical resolution than what can be achieved with the old generation of instruments such as HIRS. Indeed, while individual channels from advanced sounders only provide a broad layer measurement, it is their multiple combinations which provide significantly higher vertical resolution. Figure 9 displays the averaging kernels for the AIRS instrument (left panel) and HIRS instrument (right panel). Averaging kernels correspond to the rows of the Model Resolution Matrix (which is the convolution of the Kalman gain (or simply gain) matrix with the observation operator for a given type of observation – see, e.g., chapter *Mathematical Concepts*

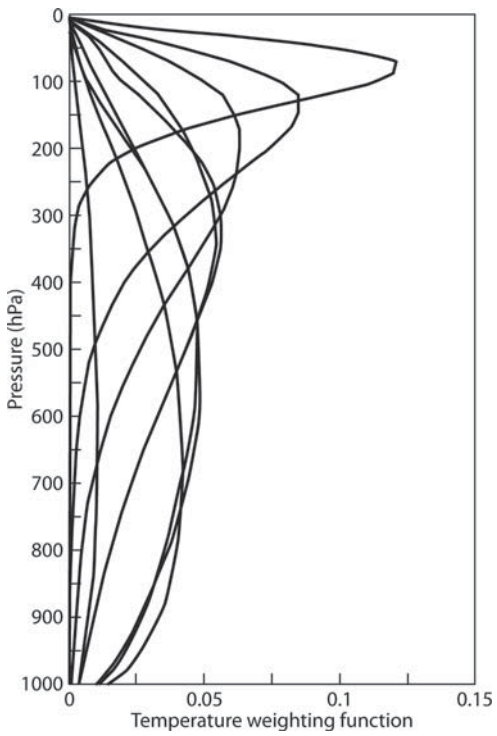



**Fig. 8** Temperature weighting functions (*x*-axis) of the AMSU-A instrument on board NOAA and METOP satellites. The *y*-axis represents the pressure levels at which a given channel is sensitive to

*of Data Assimilation*, Nichols); ideal averaging kernels for a perfectly observed system would show 60 delta function curves corresponding to the 60 levels of the L60 version of the ECMWF model. This is obviously not the case even with the AIRS instrument. However, it is clear from Fig. 9 that AIRS data offer a much higher vertical resolution than HIRS.

### 3.1.2 Surface Sensing Channels from Passive Instruments

These channels, called "imaging" channels, are located in atmospheric "window" regions of the infrared and microwave spectrum at frequencies where there is very little interaction with the atmosphere and the main contribution to the measured radiance in this case is the Earth or cloud top's surface emission. These channels are primarily used to obtain information on surface temperature; sea surface temperature derived from infrared imagery is for example widely used in NWP and global circulation modelling applications. The window channels are also used for quantities that influence the surface emissivity such as wind speed (through the roughness over sea), vegetation and snow cover. They can also be used to obtain information on cloud top (in the infrared), and rain (in the microwave). In addition, sequences of infrared images from geostationary satellites can be used to track the cloud movements and, indirectly, derive wind information. Last, and because various window channels are differentially sensitive to water vapour absorption,

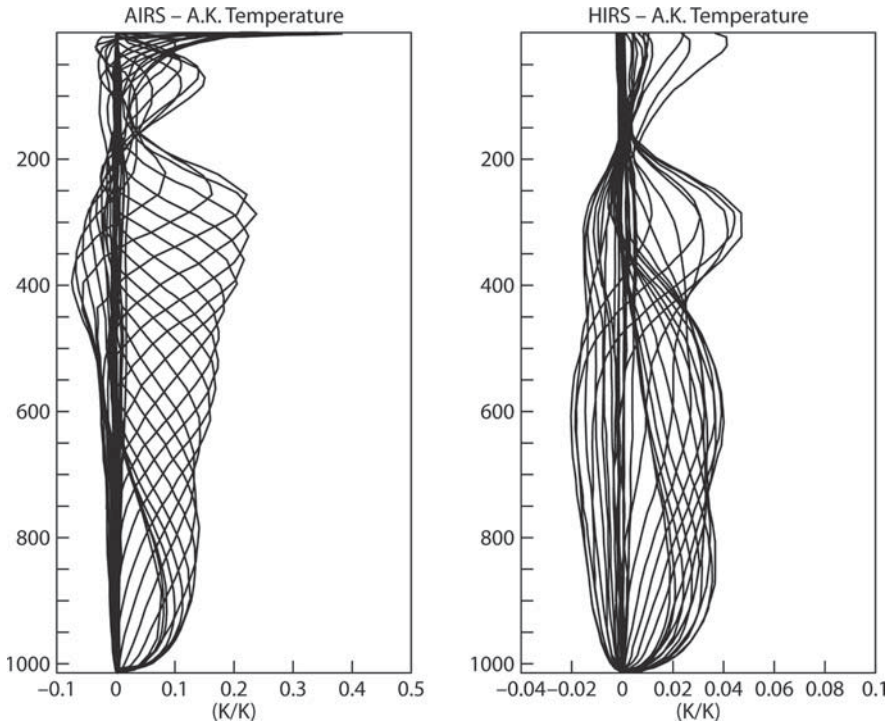



**Fig. 9** Averaging kernels (in K/K) for AIRS (*left panel*) and HIRS (*right panel*) instruments. The *y*-axis represents the altitude (in pressure level)

they can provide information on total column water vapour. Infrared instruments of this type include all imagers on GEO satellites, AVHRR on NOAA satellites, MODIS on NASA's TERRA and AQUA satellites. Microwave imagers include SSM/I (Special Sensor Microwave/Imager) and SSMIS on DMSP satellites, TMI (TRMM Microwave Imager) on the TRMM satellite and AMSR-E on NASA's AQUA satellite. New instruments such as SMOS from ESA use L-band frequencies (around 1.4 GHz) that will provide information about soil moisture over land and salinity over sea.

## *3.2 Active Technologies*

### 3.2.1 Surface Instruments

These instruments emit radiation towards the surface in the atmospheric window parts of the electromagnetic spectrum and measure what is scattered back from it. One widely used instrument of this type is the scatterometer. Scatterometers

emit microwave centimetric waves towards the sea surface. Some of the signal is reflected back to the satellite by Bragg reflection, the strength of the reflection being a function of both the amplitude of the ocean waves and the direction between the microwave beam and the wave orientation. These instruments provide, therefore, indirect information on the ocean wind speed and direction. Instruments of this type include the scatterometers on ESA's ERS satellites, Seawinds on NASA's QuikScat satellite and more recently ASCAT onboard EUMETSAT's METOP satellite. Scatterometers are also being progressively exploited over land to provide soil moisture information.

Among similar-class active instruments are altimeters which through the measurement of the time delay and the shape of the reflected signal, provide information on the sea surface height, the wave height and the wind, and SARs (Synthetic Aperture Radars) which provide information on wave height and spectra. Altimeters and SARs are carried for example on ESA's Envisat satellite.

### 3.2.2 Atmospheric Sensing Instruments

Active instruments operating in the visible (lidars) or the microwave (radars) can also analyse the signal backscattered from atmospheric targets such as molecules, aerosols, water droplets or ice particles. Their penetration capability allows the derivation of information on cloud base, cloud top, wind profiles (lidars) or cloud and rain profiles (radars). Such instruments are currently not carried on operational meteorological satellites, but several demonstration missions exist or are planned. A precipitation radar is currently flown on the TRMM satellite. A lidar and a cloud radar are also flying in tandem (Cloudsat and CALIPSO) together with the NASA's AQUA satellite (which together with a few other satellites constitute the so-called "A-train" – chapter *Research Satellites*, Lahoz), to provide complementary information on cloud and aerosol profiles. More importantly, a Doppler Wind Lidar will be flown by ESA (the principle being that by measuring the Doppler Shift of the return signal, the instrument provides information about the speed of the reflecting object along the line of sight and therefore on the wind) in 2011, providing for the first time global wind profile information.

## *3.3 Limb Technologies*

As described in Swinbank (2003), the instrument technologies described so far apply mainly to nadir-viewing instruments, that is instruments that measure the emission from, or the signal reflected by the Earth/atmosphere from a field of view below the satellite (although they may be scanned away from the nadir). Limb-viewing instruments, on the contrary, look at the atmospheric limb (i.e., the atmosphere above the horizon, as viewed from the satellite).

### 3.3.1 Limb Passive Sounders

These instruments have been used so far only on research satellites. They are particularly useful for stratospheric soundings. The instruments measure radiation emitted from close to the tangent point (see Fig. 2 chapter *Research Satellites*, Lahoz). The instruments need, therefore, to measure at wavelengths where the atmosphere is optically thin to avoid pollution from signals coming from part of the atmosphere closer to the satellite than the tangent point. This is the reason why the limb sounders are limited to the stratosphere and the upper troposphere. The main difference between limb sounders and nadir sounders is the resolution, limb sounders providing high vertical resolution (down to a few hundred metres) and rather poor horizontal resolution (around 400 km, which corresponds to the path length viewed by the instrument at the tangent point), which can of course be improved by increasing the spectral resolution of the instrument. Instruments of such class include MIPAS (Michelson Interferometer for Passive Atmospheric Sounding) onboard ESA's Envisat satellite and MLS (Microwave Limb Sounder) onboard NASA's Aura satellite.

### 3.3.2 GPS Technologies

Radio-occultation techniques using GPS (Global Positioning System) are another novel way of extracting atmospheric information. These techniques exploit an opportunity that the GPS constellation (originally designed for other applications) already exists. GPS receivers (such as the GRAS instrument on board METOP or the recently launched UCAR's COSMIC constellation) measure the Doppler shift of a GPS signal refracted along the atmospheric limb path. This refraction is proportional to (among other parameters) the density of the atmosphere, and therefore indirectly to temperature and humidity profiles. Provided a sufficient

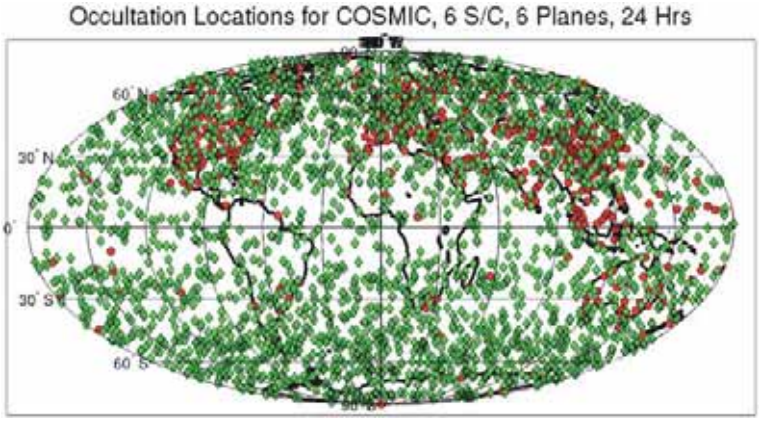



**Fig. 10** Twenty four-hour data coverage provided by the UCAR's COSMIC constellation

number of receivers are installed on LEOs, this technique offers high vertical resolution (balanced by a somewhat coarse horizontal resolution of ~200–300 km), self-calibrated and "all weather" observations of atmospheric temperature (and possibly humidity). These data are now operationally assimilated by a number of NWP centres (e.g. Healy and Thépaut 2006). Figure 10 represents the typical data coverage (occultation locations) provided by the COSMIC constellation for a period of 24 h.

## 4 Evolution of the Global Observing System

In the previous two sections we have outlined the main types of observations used for operational NWP. In this section, we will briefly mention some other techniques that might become important in the future.

### *4.1 Development of the In Situ Component of the GOS*

Radiosondes are currently the only observing system for which accurate time and positioning information is not provided with the data. The location and time of the launch locations are reported, but not the displacement due to the wind during the ascent, which lasts an hour or more. It is currently planned that enhanced reporting practices will become operational in the near future that will provide more detailed information of the balloon trajectory, and will allow the dissemination of an order of magnitude more data points in each profile (typically in excess of 1,000 levels rather than 60–100). Such improved and enhanced reporting is required to ensure full benefit of these data in present-day higher resolution NWP systems.

The time delay of signals sent from GPS satellites to ground stations can be measured extremely accurately. Since the refractive index of air varies with density and water vapour, this delay gives useful information about the integrated water vapour and temperature along the slant path between the ground station and the GPS satellites. In Europe, a system for near real time processing and distribution of several hundred ground based GPS data has been established. Similarly dense GPS observing systems exist in Japan and the USA, but the data are not yet freely available for operational use. The European data are operationally assimilated at Météo-France (Poli et al. 2007).

To supplement, or in some cases replace, radiosonde wind measurements, wind profilers are being deployed to measure wind profiles in the troposphere. Wind profilers are upward looking highly sensitive Doppler radars that are specifically designed to measure vertical profiles of wind speed and direction above the profiler site. Operational networks have been established in the central USA, Europe and Japan (Fig. 4). These data are assimilated operationally at several NWP centres

(e.g. Andersson and Garcia-Mendez 2002). Lidars are another technology that can be used to measure wind profiles from the ground.

There are proposals to fly a constellation of balloons in the lower stratosphere (e.g. GAINS, Global Air Ocean In-situ System). Winds can be derived from the balloon trajectories, and temperatures measured in situ. As well as giving useful observations of the lower stratosphere, the balloons could also be able to deploy dropsondes, targeting regions of particular interest.

## *4.2 Development of the Space Component of the GOS*

A constellation of new operational polar-orbiting satellites will progressively become available during the next 10 years, in particular the American National Polar-orbiting Operational Environmental Satellite System (NPOESS). These systems, preceded by the already existing AQUA and SSMIS and forthcoming NPOESS Preparatory Project (NPP) missions, provide unprecedented capability in terms of accurate description of various geophysical parameters. These data will be complemented by research missions that will further improve the observation of the hydrological cycle of the atmosphere by providing accurate description of clouds and rain profiles (e.g. GPM, EARTHCARE, MEGHA-TROPIQUES) and, as mentioned above, for the first time global wind profiling information (ADM-Aeolus) will also become available (Stoffelen et al. 2005; Tan et al. 2007). It is also possible that high spectral infrared sounders will also become available on geostationary orbits, providing high vertical and temporal resolution atmospheric profiles.

To be more specific and according to the latest CBS OPAG ET-EGOS (Expert Team on the Evolution of the Global Observing System) (WMO 2004), the vision for the evolved GOS at the 2015 horizon and beyond suggests (see also the chapter *Research Satellites*, Lahoz):

- Six operational Geostationary satellites (GEOs) with onboard multispectral imagers (Infrared/Visible – IR/VIS), some with hyperspectral sounders (IR);
- 4 operational low earth orbiting (LEO) satellites providing a uniform data coverage with onboard multispectral imagers (Microwave/Infrared/Visible/Ultraviolet – MW/IR/VIS/UV), sounders (MW), radio-occultation (RO) capabilities, some with hyperspectral sounders (IR), conical scan MW or scatterometers and altimeters;
- In addition, several R&D satellites will complement the operational constellation. Further LEOs with active and passive microwave precipitation and cloud measurements, and two LEOs with soil moisture and ocean salinity capability will also become available within the next 10-year timeframe;
- Atmospheric composition missions, currently available with the Envisat-EOS satellites (as of 2009), will hopefully reach a more operational status towards and after 2015 (e.g. ESA Sentinels 4 and 5);

- Last but not least, a LEO with wind profiling capabilities will become available during this timeframe.

Moreover, the recent results obtained by a number of operational centres (e.g. Healy and Thépaut 2006) suggest that a GPS radio-occultation observing capability is now a high priority requirement, not only for NWP but also for reanalysis and climate applications (see chapter *Reanalysis: Data Assimilation for Scientific Investigation of Climate*, Rood and Bosilovich).

## 5 Concluding Remarks

We have already noted that we do not fully exploit all the information in conventional observations. Unused elements, such as reports of cloud type or rainfall, might contain enough information to help a forecaster identify locations and instances where the computer analysis is most likely in error. Methods for manual intervention that allow modification of the computer analyses have been developed for this purpose, although that is not without pitfalls. Satellite imagery also has a lot of information about cloud distribution and types. While the use of cloud data is not straightforward, the presence (or absence) of cloud allows one to draw some inferences about humidity (or liquid water content) and vertical velocity, which should be useful information for a data assimilation scheme. The importance of the humidity analysis, and hence the importance of all data that pertain to the hydrological cycle (see chapter *Land Surface Data Assimilation*, Houser et al.), is set to increase (Andersson et al. 2005) as the assimilation methods improve, and model resolution is increased. The use of cloud and rain information is currently the subject of active research (Chevallier et al. 2002).

Some precipitation data are beginning to be used in data assimilation systems, in mesoscale models for short range forecast models, as well as global models (Bauer et al. 2006). Hou et al. (2001) studied the impact of assimilating rainfall and total precipitable water information from the TRMM (Tropical Rainfall Measuring Mission) and SSM/I (Special Sensor Microwave/Imager). They showed that these data not only improve the hydrological cycle but also cloud and radiation information, as well as large-scale motion in the tropics.

Towards the end of 1999 a more advanced version of the variational analysis (4D-Var; see chapter *Variational Assimilation*, Talagrand) was developed at ECMWF and significant changes also occurred into the GOS mainly due to the launch of the first ATOVS instrument onboard of NOAA satellites. A comprehensive set of OSEs (Observing System Experiments; see chapter *Observing System Simulation Experiments*, Masutani et al.) was then performed (Bouttier and Kelly 2001; Kelly et al. 2004) to validate the new assimilation model and assess the impact of various components of the GOS within this new scheme. From the results, the necessity of using satellite data in NWP was clear. In fact, for the first time in

the Northern Hemisphere, satellite data had larger impact than radiosonde observations (particularly in summer). Nowadays, denying all satellite observations from the ECMWF assimilation system entails a skill reduction of about 3 days in the Southern Hemisphere and about 2/3 of a day in the Northern Hemisphere. This shows the progress that has been made in the past 10 years at better exploiting atmospheric information from these satellite observations.

To finish, and to give an idea of the size of the GOS, as of February 2009, ECMWF uses actively every day ~18 million observations, i.e., pieces of information (one wind component at one level; one radiosounding; one channel radiance, are one observation). Many more data are quality-controlled.

## References

Andersson, E., P. Bauer, A. Beljaars, F. Chevallier, E. Hólm M. Janisková, P. Kållberg, G. Kelly, P. Lopez, A. McNally, E. Moreau, A.J. Simmons, J.-N. Thépaut and A.M. Tompkins, 2005. Assimilation and modeling of the atmospheric hydrological cycle in the ECMWF forecasting system. *Bull. Amer. Meteorol. Soc.*, **86**, 387–402.

Andersson, E. and A. Garcia-Mendez, 2002. Assessment of European wind profiler data, in an NWP context. *ECMWF Tech. Memo.*, **372**, pp 14.

Bauer, P., P. Lopez, A. Benedetti, D. Salmond and E. Moreau, 2006. Implementation of 1D+4D-Var assimilation of precipitation-affected microwave radiances at ECMWF. I: 1D-Var. *Q. J. R. Meteorol. Soc.,* **132**, 2277–2306.

Bouttier, F. and G. Kelly, 2001. Observing-system experiments in the ECMWF 4D-Var data assimilation system. *Q. J. R. Meteorol. Soc.,* **127**, 1469–1488.

Chevallier, F., P. Bauer, J.-F. Mahfouf and J.-J. Morcrette, 2002. Variational retrieval of cloud profiles from ATOVS observations. *Q. J. R. Meteorol. Soc.,* **128**, 2511–2525.

Compo, G.P., J.S. Whitaker, and P.D. Sardeshmukh, 2006. Feasibility of a 100-Year Reanalysis Using Only Surface Pressure Data. *Bull. Amer. Meteorol. Soc.*, **87**, 175–190.

Eyre, J.E., 2000. Planet Earth seen from space: Basic Concepts. In *Proceedings of the ECMWF Seminar on the Exploitation of the New Generation of Satellite Instruments for Numerical Weather Prediction*, pp 5–19.

Healy, S.B. and J.-N. Thépaut, 2006: Assimilation experiments with CHAMP GPS radio occultation measurements *Q. J. R. Meteorol. Soc.*, **132**, 605–623.

Hou, A.Y., S. Zhang, A. da Silva, W. Olson, C. Kummerow and J. Simpson, 2001. Improving global analysis and short-range forecats using rainfall and moisture observations derived from TRMM and SSM/I passive microwave sensors. *Bull. Amer. Meteor. Soc.*, **82**, 659–679.

Järvinen, H., E. Andersson and F. Bouttier, 1999. Variational assimilation of time sequences of surface observations with serially correlated errors. *Tellus*, 51A, 469–488.

Kelly, G., A.P. McNally, J.-N. Thépaut and M.D.E. Szyndel, 2004. Observing Experiments of all main data types in the ECMWF operational system. In *Proceedings of 3rd WMO Workshop on the Impact of Various Observing Systems on Numerical Weather Prediction*, Alpbach, Austria, 9–12 March 2004, pp 63–94.

Labitzke, K.G. and H. van Loon, 1999. *The Stratosphere Phenomena, History and Relevance.* Springer, Berlin, 179 pp.

Langland, R.H., 2005. Issues in targeted observing. *Q. J. R. Meteorol. Soc.*, **131**, 3409–3425.

Poli, P., P. Moll, F. Rabier, G. Desroziers, B. Chapnik, L. Berre, S.B. Healy, E. Andersson and F.-Z. El Guelai, 2007. Forecast impact studies of zenith total delay data from European near real-time GPS stations in Météo-France 4DVar. *J. Geophys. Res.*, **112**, D06114, doi: 10.1029/2006JD007430.

Stoffelen, A., J. Pailleux, E. Källén, J. M. Vaughan, L. Isaksen, P. Flamant, W. Wergen, E. Andersson, H. Schyberg, A. Culoma, R. Meynart, M. Endemann and P. Ingmann, 2005. The atmospheric dynamics mission for global wind measurement. *Bull. Amer. Meteorol. Soc.*, **86**, 73–87.

Swinbank, R., 2003. Observing the atmosphere. In *Data Assimilation for the Earth System*. NATO Science Series: IV. Earth and Environmental Sciences 26, Swinbank, R., V. Shutyaev and W.A. Lahoz (eds.), Kluwer Academic Publishers, Dordrecht, The Netherlands, pp 137–148, 378pp.

Tan, D.G.H., E. Andersson, M. Fisher and L. Isaksen, 2007. Observing system impact assessment using a data assimilation ensemble technique: Application to the ADM-Aeolus wind profiling mission. *Q. J. R. Meteorol. Soc.*, **133**, 381–390.

Thépaut, J.-N., 2003. Assimilation of remote sensing observations in numerical weather prediction. In *Data Assimilation for the Earth System*. NATO Science Series: IV. Earth and Environmental Sciences 26, Swinbank, R., V. Shutyaev and W.A. Lahoz (eds.), Kluwer Academic Publishers, Dordrecht, The Netherlands, pp 225–240, 378pp.

Weissmann, M. and C. Cardinali, 2007. Impact of airborne Doppler lidar observations on ECMWF forecasts. *Q. J. R. Meteorol. Soc.*, **133**, 107–116.

WMO/CBS OPAG on *Integrated Observing Systems*: Final report of the 7th Session of the Expert Team on Observational Data Requirements and redesign of the Global Observing System, Geneva, Switzerland, 12–16 July 2004. *Available from WMO website.*

# Assimilation of Operational Data

**Erik Andersson and Jean-Noël Thépaut**

## 1 Introduction

In this chapter we focus on the observations that are available for operational, real-time applications in meteorology, i.e., for numerical weather prediction (NWP). Many in situ observations can be treated as point-wise measurements. Their influence on the analysis is expected to be localized and smoothed according to the specified background error covariance structures (chapters *Mathematical Concepts of Data Assimilation*, Nichols; *Error Statistics in Data Assimilation: Estimation and Modelling*, Buehner). Most remotely-sensed sounding data, on the other hand, are integrated measurements that cannot be treated as point-wise observations. This is an important distinction which needs to be accounted for by the analysis scheme. Therefore, we expand the discussion of observation operators to integrals, and examine how such data can be expected to influence the analysis. In operational meteorology, the most prominent examples of integral observations are measurements of infrared and microwave radiation from satellite instruments. Other, recent examples include ground-based GPS (Global Positioning Satellites) and radio-occultation data. The related issues of quality control and data thinning are also covered. Assimilation of time-sequences of observations is discussed. This chapter complements chapters *The Global Observing System* (Thépaut and Andersson) and *Research Satellites* (Lahoz).

## 2 Assimilation of Radiance Observations

Many operational satellite instruments measure infrared or microwave radiation emanating from the atmosphere and the Earth's surface. These data provide information on the temperature and humidity of the atmosphere, the temperature and emissivity of the surface, as well as clouds and precipitation which all affect the

E. Andersson (✉)
European Centre for Medium-Range Weather Forecasts, ECMWF, Shinfield, Reading, UK
e-mail: jean-noel.thepaut@ecmwf.int

W. Lahoz et al. (eds.), *Data Assimilation*, DOI 10.1007/978-3-540-74703-1_11,

measured radiances. These data types have been particularly challenging to the design of data assimilation schemes, as the information from a single radiance measurement depends on so many atmospheric and surface variables. The data assimilation scheme must faithfully distribute the information along the path of the measurement (near-vertical in the case of nadir-sounding data and near-horizontal in the case of limb-sounding data) and also partition the information accurately between temperature, humidity and surface skin temperature, and other quantities as appropriate (Rodgers 1976, 1990). See also chapters in Part I, *Theory*, for a discussion of data assimilation schemes.

## *2.1 Constraints on the Inversion of Radiance Data*

As the radiance measurements generally provide information on broad vertical structures, the conversion from radiance to more detailed profiles of temperature and humidity is an ill-posed problem unless additional "background" information is used (Rodgers 1976; Eyre and Lorenc 1989). In other words, several distinctly different profiles of temperature and humidity may produce the same radiances. Some additional information is thus necessary for the retrieval of an unambiguous solution.

The 1D-Var (one dimensional variational) retrieval scheme (Eyre 1989) used profiles from a short range forecast as background, whereas other retrieval schemes used statistical background information (Reale et al. 1986) or libraries of representative atmospheric profiles (Chédin and Scott 1985; Fleming et al. 1986). Even with very sophisticated techniques it is unavoidable that errors in the selected background will contribute to the retrieval error (Flobert et al. 1991). The problem shows up as very systematic air-mass-dependent biases in the retrieved data (Andersson et al. 1991). The errors introduced by the retrieval process are characterized by horizontal correlations that vary with the meteorological conditions, and are therefore difficult to accurately account for in the analysis. This problem is fully eliminated by incorporating the retrieval process within the analysis. A combined retrieval/analysis approach enables a more accurate combination of the information contained in the background, in the radiances and in the conventional data (Andersson et al. 1994; McNally et al. 1999, 2000; Köpken et al. 2004). In this approach all data are analysed simultaneously in a single global inversion problem.

The presence of conventional data may with this methodology help the inversion of radiance data in their vicinity. In other words the radiance inversion to temperature and humidity is somewhat constrained by the in situ data. The more observational information that can be used in the retrieval/analysis procedure, including both point-wise and integral observations, the more accurate the analysis is likely to be, assuming the constraints are applied appropriately. In 3D/4D-Var (three/four dimensional variational assimilation; see chapter *Variational Assimilation*, Talagrand) the background term, $J_b$ constrains (smoothes) the analysis/retrieval in the horizontal and vertical, whereas in 1D-Var the background

term constrains the vertical structure only. The horizontal part of the $J_b$ constraint smoothes out the detrimental effect on the analysis of any small-scale noise in the measurements. The mass-wind coupling of $J_b$, which imposes approximate geostrophy in the extra-tropics, makes it possible also for wind observations to constrain the temperature (gradient) information inferred from radiances and in situ data.

## *2.2 Non-linear Dependence on the Background Humidity Field*

The observation operator $\mathcal{H}(\mathbf{x})$ for radiance data is a set of radiative transfer calculations. These are carried out using a fast radiative transfer code, e.g., RTTOV, Radiative Transfer for TOVS (Eyre 1991; Saunders et al. 1999). Saunders et al. show that the radiative transfer calculations are very nearly linear with respect to temperature and significantly non-linear with respect to humidity. Because of the non-linearity the sensitivity of the satellite radiance measurement (in a humidity sensitive channel) will vary significantly with the amount and the distribution of humidity in the atmosphere. For computational efficiency, the observation operators are often linearized. Most of the non-linear dependence of the full observation operators $\mathcal{H}(\mathbf{x})$ is nevertheless accounted for by linearizing around the current atmospheric state ($\mathbf{x}$) using what are called tangent-linear operators. In the incremental formulation of the variational problem (see the chapters *Mathematical Concepts of Data Assimilation*, Nichols; *Variational Assimilation*, Talagrand) this is done by using $\mathcal{H}(\mathbf{x})$ to evaluate the innovations $\mathbf{d}(= \mathbf{y} - \mathcal{H}(\mathbf{x}_b))$ and by linearizing $\mathcal{H}(\mathbf{x})$ around the background state $\mathbf{x}_b$ which accounts for the state-dependence much more accurately than linearizing around climatology would. The tangent-linear of $\mathcal{H}(\mathbf{x})$ (denoted **H)** is thus applied to small increments with respect to the already accurate $\mathbf{x}_b$ and linearization errors are largely avoided. To the extent that **H** varies with $\mathbf{x}_b$ we thus have a different **H** at every location. Linearization errors are further reduced in many operational applications by re-linearizing the observation operators around a preliminary analysis approximately half-way through the minimization (Rabier et al. 2000), or even more frequently (Trémolet 2005), to further incorporate non-linear effects.

The non-linear aspects of the observation operators have important effects on the analysis. For example, the information content in radiance data, which contributes to reduced analysis error, depends on **H** through the term $\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H}$ ($T$ being the transpose; **R** being the observation error covariance). This term represents the radiance information in terms of analysed quantities (i.e., the temperature and humidity components of $\mathbf{x}$). As **H** potentially is different for every background profile, the analysis thus takes account of the fact that the retrieval/analysis accuracy may vary with varying atmospheric conditions. The sensitivity of certain infrared measurements to changes in humidity varies strongly over the globe due to the distribution of humidity in the atmosphere. This has implications on the extent to which such channels contribute to the humidity retrieval/analysis accuracy, which in turn determines the relative weight given to the radiance data, the background information and the in situ data, respectively.

Such state-dependent data weighting could be incorporated in a data assimilation scheme with separate retrieval and analysis steps only by passing the full covariance matrix of retrieval error for each retrieved profile from the retrieval scheme to the analysis. However, such an approach would be difficult in practice due to the large data volume involved.

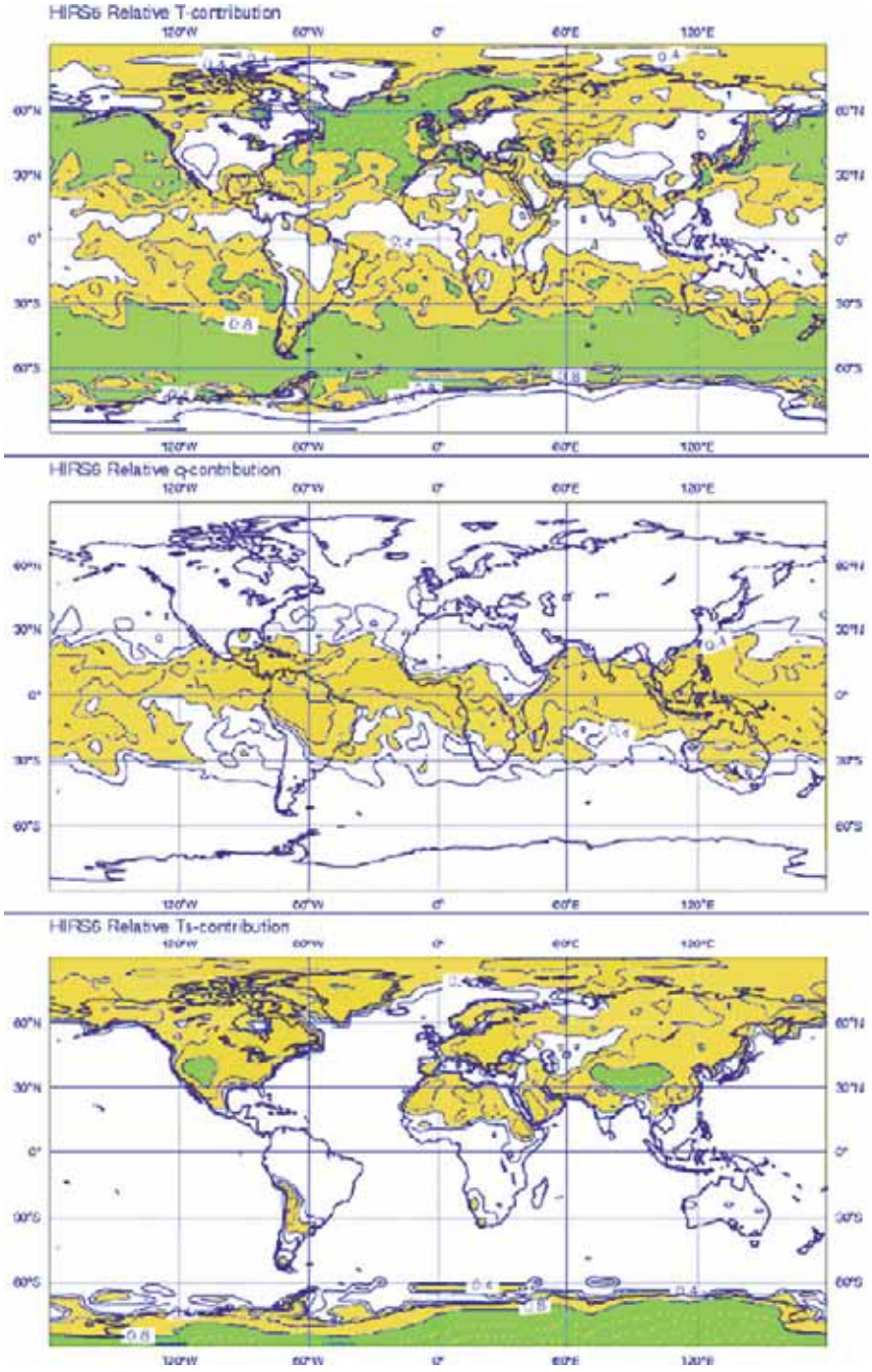
## *2.3 Temperature/Humidity Partitioning of Radiance Increments*

Radiance measurements in sounding channels may be sensitive to some or all of air temperature, humidity, ozone, several trace gases and clouds, as well as the temperature and emissivity of the surface. So far, it has been operational practice to seek to remove or mask the effects of clouds and the surface through careful data selection and screening of the data. Alternatively, surface effects can partly be accounted for by allowing the surface skin temperature (and, optionally, other quantities) at each satellite data point to vary during the minimization by including them in an extension of the control vector. Similar techniques are used in bias estimation – see chapter *Bias Estimation*, Ménard.

A measurement in a humidity sensing channel indicating a higher brightness temperature than that computed from the background implies either that the background temperature is too low or that the background humidity is too high. Such a measurement should result in analysis increments in both temperature and humidity. The ambiguity between the two quantities (in the absence of other observations) is resolved by the background term $J_b$: the observed radiance increment will be partitioned between temperature and humidity analysis increments depending on the relative magnitude of temperature and humidity background errors in **B** (the background error covariance matrix – see chapter *Mathematical Concepts of Data Assimilation*, Nichols). The partitioning also depends on the atmospheric transmission resulting in varying sensitivity of the measurement to changes in humidity, and this is encapsulated in the **H** operator as discussed above. A measurement in a given channel may result primarily in humidity analysis increments in some atmospheric conditions and result primarily in temperature increments in other atmospheric conditions.

Diagnostic calculations which help one understand these effects are presented in Fig. 1, where the effective background error in terms of each radiance channel, i.e., the term $\mathbf{HBH}^T$ has been estimated. The figure shows maps of the relative contributions from temperature, humidity and surface temperature background errors, respectively, to the HIRS channel-6 background error. The figure shows that

**Fig. 1** (continued) Relative contribution to the HIRS channel 6 background error from respectively air temperature (*top panel*), humidity (*middle panel*) and surface skin temperature (*bottom panel*). The contours are 0.2, 0.4, 0.6 and 0.8 with shading starting at 0.4 (*yellow*); values greater than 0.8 are *shaded green*. The sum of the three charts is equal to one everywhere, by construction. From Andersson et al. (2000) (© Royal Meteorological Society)


HIRS6 Relative T-contribution
HIRS6 Relative q-contribution
HIRS6 Relative Ts-contribution

humidity background errors dominate in the tropics whereas temperature background errors dominate in the mid and high latitudes. As $\mathbf{HBH}^T$ is one of the terms that determine the magnitude of the analysis increments we can expect that HIRS channel-6 measurements will affect predominantly the humidity analysis in the tropics and predominantly the temperature analysis in mid and high latitudes.

### *2.4 Passive Tracer Analysis*

In four-dimensional assimilation schemes, the time variation recorded by observations can be explored to great effect. For example, through the adjoint (chapter *Variational Assimilation*, Talagrand) of the humidity advection over the assimilation time window, 4D-Var can fit a time-sequence of humidity data either by modifying the humidity field itself, or by adjusting the advecting wind. Frequent observations in a water vapour sounding channel can thus affect the 4D-Var analysis of the wind field. This is an important result which holds for any passive tracer quantity. It has provided the motivation to work on the assimilation of frequent water vapour radiance data from geostationary satellites in an attempt to improve the analysis of not only the humidity field but also the tropical wind field (Munro et al. 1999; Köpken et al. 2004), and furthermore to develop assimilation of ozone data to improve the stratospheric wind analysis (Riishøjgaard 1996; Hólm et al. 1999).

## 3 Assimilation of Hourly Surface Pressure Measurements

It has long been recognized that surface pressure tendency observations or, equivalently, time series of frequent surface pressure observations, provide important information on the intensity and motion of mid latitude storms. Such observations are essential for the subjective analysis of weather maps and for short-range forecasting. These observations have, however, not been used in objective analysis until the advent of 4D-Var. This is because of difficulties accounting for the temporal information in the data in static analysis schemes. In a static scheme all observations used in an analysis are assumed to refer to the analysis time: all observations belonging to a given 6-h analysis period would thus refer to the central time of that period. From frequently reporting stations, only the observation closest to the centre of each 6-h period could be used.

Unlike static data assimilation schemes, four-dimensional data assimilation schemes compare the observations to a time-sequence of model states (the model trajectory) that extends over the assimilation window. The benefits are that the observations are used at the appropriate time, and many more observations can be used from frequently reporting stations. This enables effective use of time sequences of surface pressure observations, for example.

### *3.1 Synoptic Analysis of Rapidly Developing Storm Using Surface-Pressure Data from a Single Station*

Synopticians rely heavily on surface pressure tendency data to improve the analysis of rapidly developing mid latitude storms which are highly baroclinic systems. An observed time sequence of surface pressure observations might, for example, indicate that a particular storm is not developing rapidly enough in the model. The ideal analysis correction in this situation should consist in small adjustments to the three-dimensional structure of the storm at initial time, such as to increase the baroclinicity and so the intensification rate of the storm. The ideal analysis increment would be an unstable perturbation with respect to the given atmospheric situation over a finite time interval (the assimilation window).

Rabier et al. (1996) showed that 4D-Var will tend to create analysis increments that correct the part of the background error which projects onto the fast-growing singular vectors of the model. Thereby, 4D-Var implicitly increases the effective background errors in the unstable directions (Thépaut et al. 1993, 1996; Rabier et al. 2000), which facilitates such flow-dependent analysis corrections. The 4D-Var system, therefore, provides the mechanisms necessary for effective use of surface pressure tendency data. Long before the time of 4D-Var, Bengtsson (1980) studied what should be the optimal analysis response to pressure tendency data in a baroclinic system within a four-dimensional data assimilation system. In his theoretical work, Bengtsson demonstrated that it is desirable to update the flow in the mid and upper troposphere in response to surface pressure tendency observations. The following example from Järvinen et al. (1998) shows 4D-Var results with respect to a real storm that was developing more quickly in reality than in the model as it was approaching Ireland from the North Atlantic: the so called “Irish Christmas Storm”, 1997.

When the storm was approaching Ireland, the observations were indicating a quicker pressure fall (20 hPa in 5 h) than the model prediction (16.5 hPa in 5 h). The assimilation experiment using hourly surface pressure data produced an analysis with a 3.5 hPa deeper cyclone than the assimilation without the additional data. In this situation, where the model had under-predicted the intensification of a quickly developing baroclinic system, we focus on the upper-air analysis increments created by 4D-Var in response to the surface pressure data. To this end, an experiment was run using only the surface-pressure time sequence from one single Irish SYNOP station, Malin Head. The reported pressure at Malin Head fell by 20.1 hPa in 5 h, whereas the trajectory model integration from the background produced a pressure fall of 15.8 hPa in the same period, i.e., an under-prediction of 4.3 hPa, indicative of a less intense deepening of the storm in the model than in reality.

The 4D-Var correction of the initial state, i.e., the analysis increment, is shown in Fig. 2a. Its evolution 3 and 6 h into the assimilation window is shown in Fig. 2b, c, respectively. These are results from an assimilation of hourly pressure observations from Malin Head. The figures show that the analysis increment in the time-sequence experiment intensifies rapidly over the 6-h assimilation period. We

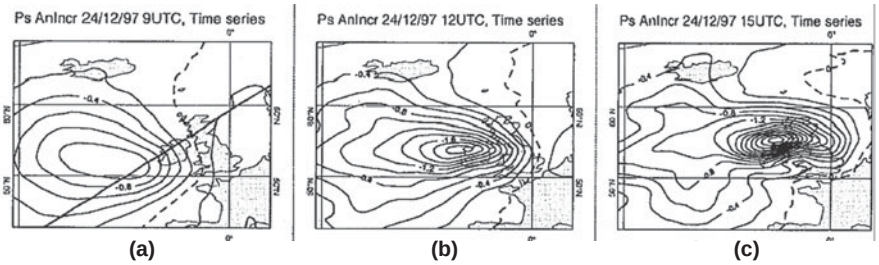



**Fig. 2** Surface pressure (hPa) analysis increment (**a**, *left hand panel*) resulting from a time series of hourly surface pressure data at Malin Head, Ireland. The evolution of the increment 3 and 6 h into the 6-h assimilation window is shown in (**b**, *middle panel*) and (**c**, *right hand panel*) (From Järvinen et al. 1998)

can conclude that the 4D-Var analysis increments have successfully destabilized the atmosphere with respect to the baroclinic development of the storm over the 6-h time interval.

The vertical cross-sections of the analysis increments show that the analysis has cooled the cold air to the west of the storm and sharpened the thermal gradient in the frontal zone. The tropopause height in the cold air mass has been lowered and the vertical winds in the frontal region have been strengthened. It is of interest to note that the maximum temperature increments, created by the time sequence of surface pressure observation at Malin Head, are located in the mid troposphere 10° to the west of the station. In the absence of any flow-dependent effects the maximum temperature increment would occur in the lower troposphere directly above the station.

## *3.2 Background Errors in Observable Quantities*

The background error covariances **B** in a variational analysis are specified in terms of those quantities that lead to a compact formulation of the background term (the $J_b$ term of the penalty function in the variational analysis), viz., balanced vorticity, unbalanced temperature, divergence and surface pressure, and specific humidity. Because of the non-linearities in the observation operators, it is not immediately obvious how the magnitudes of these background errors can be expressed in terms of observable quantities such as radiances for comparison with the various observation errors.

Within the variational analysis, the background errors in terms of observed quantities ($\mathbf{HBH}^T$) are implied, but are not normally computed explicitly. They depend, in general, on the $J_b$ formulation and on the observation operators. In the case of radiance observations, this involves the Jacobian of the radiative transfer model which in turn depends on the atmospheric state. Efficient methods to diagnose the diagonal of $\mathbf{HBH}^T$ have been developed. One method is an adaptation of the randomization technique suggested by Fisher and Courtier (1995) to estimate the "effective" background error variances in a variational analysis system. By

multiplying each random vector with the tangent linear observation operators **H**, one can produce maps of background error standard deviations for observed quantities, i.e., estimates of the diagonal of $\mathbf{HBH}^T$.

The diagnosed background errors for observed quantities can be compared with statistics of observation-minus-background departures. If the values are not consistent, it is an indication that some aspects of the specified **B** matrix may need to be improved. Such an investigation was performed by Poli et al. (2007) with respect to the new observational data (namely zenith total delay) provided by ground-based GPS stations.

## 4 Variational Quality Control of Observations with Non-Gaussian Errors

The assimilation methods presented so far assume that the observations are affected by random errors that can be well approximated by a Gaussian frequency distribution. Furthermore, the observations are assumed to be bias-free and devoid of any serious error due to malfunction of instruments, incorrect readings, software errors, and other so-called gross errors. Several different methods have been developed to detect such errors and reject all data that have high probability of being in gross error. We have seen that the innovation vector $\mathbf{y} - \mathcal{H}(\mathbf{x}_b)(= \mathbf{d})$ is a measure of the departure of each observation against a common atmospheric state. This information is extremely useful to assess whether any gross errors contaminate the new data. The departures thus provide the basis for several successful quality control procedures: the first-guess check, buddy-checks, OI (optimal interpolation) checks, Bayesian methods and variational quality control.

Quality control is an inherently non-linear process. As we have already seen, the incremental formulation of 4D-Var has the advantage that non-linearities can be accounted for, and this can be explored also for the purpose of quality control.

### *4.1 Probability Density Function of Observation Error*

The quadratic form of the observation cost function, $J_o$, corresponds to an assumption of Gaussian distributions of observation error (Lorenc 1986). This can be seen from the definition

$$J_o = \ln p + c \tag{1}$$

where $p$ is the probability density function (PDF) of the error in the observations and $c$ is an arbitrary constant.

With a Gaussian PDF, i.e.,

$$p = \exp\left[0.5\ (\mathbf{H}\delta\mathrm{x} - \mathbf{d})^T \mathbf{R}^{-1} (\mathbf{H}\delta\mathrm{x} - \mathbf{d})\right] \tag{2}$$

we obtain the familiar quadratic expression for $J_o$ after insertion in Eq. (1).

Error distributions for real observations are rarely precisely Gaussian. Real distributions often have significantly wider tails than can be expected from purely Gaussian statistics. These tails of the distributions indicate the presence of gross errors in the data that should be excluded from the analysis. Non-Gaussian PDFs can be constructed to better reflect the real distributions. The corresponding variational cost function then becomes non-quadratic, and can be derived from Eq. (1). Ingleby and Lorenc (1993) proposed to model the PDF of observation error as a sum of two distributions: one which follows the normal Gaussian distribution, representing random errors, and one which is modelled by a flat "box car" distribution, representing the population of data affected by gross errors. Other models for the probability density of incorrect data have also been used (e.g. Huber 1977; Gelman et al. 1995).

The choice of a flat distribution is convenient as it corresponds to the assumptions that those data (the incorrect data) provide no useful information to the analysis. The corresponding modification to the cost function to take into account the non-Gaussian nature of gross errors has the effect of reducing the analysis weight given to data with large departures from the current value of **x** (or preliminary analysis). This is fundamentally different from the normal Kalman gain weights which are independent of the value of the observed departure.

## *4.2 Variational Quality Control*

The expression for the modified cost function $J_o^{\mathrm{QC}}$ and its gradient with variational quality control (VarQC) in the case that the gross-error PDF is represented by a flat (box-car) distribution becomes:

$$
\begin{aligned}
J_o^{QC} &= -\ln\left[\frac{\gamma + \exp(-J_o^N)}{\gamma + 1}\right], \\
\nabla J_o^{QC} &= \nabla J_o^N\left[1 - \frac{\gamma}{\gamma + \exp(-J_o^N)}\right], \\
&\text{with } \gamma \text{ defined as } \gamma = \frac{A\sqrt{2\pi}}{(1-A)2d},
\end{aligned}
\tag{3}
$$

where $A$ represents the prior probability of gross error and $d$ the width of the box-car function. The symbol $J_o^N$ is used here to denote the normal $J_o$ resulting from Gaussian PDFs. Figure 3 illustrates that $J_o^{QC}$ is near-quadratic for small observation departures, but flattens out for large departures and that the gradient $\nabla J_o^{QC}$ then drops towards zero. This is in contradistinction to the linearly-increasing gradient of the normal cost function $\nabla J_o$. A VarQC weight can be defined as the ratio between $\nabla J_o^{QC}$ and $\nabla J_o$. This weight is one for small observation departures, meaning that the observation is fully used; the weight approaches zero, i.e., the observation is rejected, for large observation departures. The interval within which the VarQC

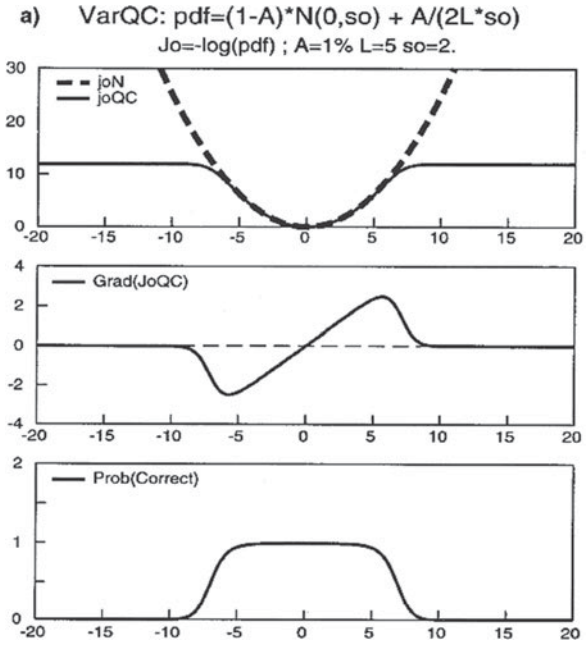



**Fig. 3** The observation cost function (*top panel*) for one single observation without (*dashed*) and with (*full line*) variational quality control (VarQC). The *middle panel* shows the gradient of the VarQC cost function. The *lower panel* shows the a posteriori probability that the observation is correct (one minus the probability of gross error). Arbitrary units

weight changes from nearly one to nearly zero is relatively narrow. The observations whose departures fall within this interval are partly used and partly rejected. It is these observations that will contribute to the non-quadraticity of the cost function, not the rejected ones.

VarQC does not reject data irrevocably. The VarQC weights can be recomputed at every iteration of the minimization. Rejected observations can thereby regain influence on the analysis during later iterations if supported by surrounding data. This "soft" QC approach is in contrast to traditional "hard" quality control procedures which detect and discard questionable data prior to the main analysis. There is also no need for a separate (often complicated) quality control (QC) decision making algorithm. All used data of all observed quantities are quality controlled simultaneously during the course of the main minimization.

In VarQC the weights given to observations are a function of the magnitude of the observed departure and may be zero. With purely Gaussian statistics, on the other hand the weight is never zero, corresponding to the assumption that every observation improves the analysis regardless of the distance between observation and analysis.

Non-Gaussian observation error statistics lead to a non-quadratic cost function. There may also be multiple minima in the VarQC cost function, each representing the rejection/non-rejection of individual data (see examples given in Dharssi et al. 1992). The technique used by Dharssi et al. was to set the observation errors to very large values initially and then gradually reduce them to their "true" value. The

approach taken by Andersson and Järvinen (1999) to limit this problem is instead to start VarQC from as good as possible an initial state. This is achieved by partially minimizing the cost function without quality control before VarQC is switched on. The approach relies on the various pre-analysis checks, such as the checks against climatology and against the background (Järvinen and Undén 1997) to remove the obviously wrong observations. Otherwise, the preliminary analysis without VarQC could be seriously corrupted by gross errors.

### *4.3 The Need for Realistic Background Error Specification*

The specified background error statistics are very important for the success of quality control procedures. The misfit between background and observations (the innovations) can be large either because the observations are wrong or because the background is wrong (or both). Observations may therefore be rejected for the wrong reason in areas where the background is very poor, unless the specified background error is high there (Dee et al. 2001). There is thus a danger that observations are rejected in the vicinity of intense storms, not because the observations are poor but because the background is poor.

The development of the 4D-Var system on longer assimilation periods (longer than the current 6 or 12 h) and the development of data assimilation ensemble techniques, will make the estimated background error at observation points more dependent on the atmospheric state (Thépaut et al. 1996; Rabier et al. 1998). As a consequence of these flow-dependent structures, it is expected that automatic quality control procedures (e.g. VarQC) will become more skilful at discriminating between good and incorrect data also in extreme weather situations and dynamically active areas, such as rapidly moving cyclones and troughs.

## 5 Impact of Observations on the Quality of Numerical Forecasts

The impact of observations in NWP data assimilation systems is usually evaluated through Observing System Experiments (OSEs; see chapter *Observing System Simulation Experiments*, Masutani et al.) where several components of the Global Observing System (GOS) are denied individually and the subsequent forecast skill degradation accordingly quantified (e.g. Bouttier and Kelly 2001). New complementary statistically based methods that evaluate the contribution of observations to a data assimilation system are emerging, for example by looking at their impact on the reduction of the analysis error variance (Desroziers et al. 2005), or by diagnosing their influence on and their consistency within the analysis (Cardinali et al. 2004; Chapnik et al. 2006). See also chapter *Evaluation of Assimilation Algorithms* (Talagrand).

A comprehensive description of the impact of observations on NWP is provided in the proceedings of the third WMO workshop on the impact of various observing

systems on Numerical Weather Prediction (WMO 2004). All the results compiled in this document suggest that:

- Conventional observations remain an important component of the GOS. Radiosondes contribute significantly to regional and also global NWP. Aircraft observations also provide valuable contribution to the GOS and most NWP centres are now using the high temporal resolution of these data, especially during flight ascents and descents. Last but not least, and despite the overwhelming volume of satellite data, surface observations (in particular surface pressure) over sea remain essential to anchor the surface pressure field. Surface data are important for regional and global NWP;
- The ability of global NWP systems to use satellite data has evolved remarkably over the last decades and these observations now constitute the backbone of the GOS for this application. Figure 4 is updated from Simmons and Hollingsworth (2002). This figure presents running annual-mean anomaly correlation of 500 hPa height for ECMWF's operational 3-, 5-, 7- and 10-day forecasts for the extra-tropical Northern and Southern Hemispheres for the period from January 1980 to February 2007. The first remark is the general upward trend of the curves (indicating a progressive improvement of the forecast quality over the covered period). A second striking feature is the higher rate of improvement in the forecasts in the Southern Hemisphere. In 27 years, the skill of medium range weather forecasts in the Southern Hemisphere has reached a level now comparable to the one in

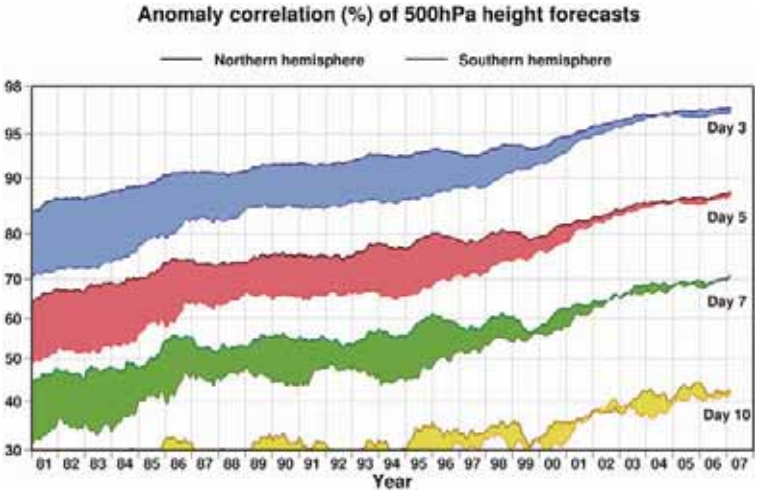



**Fig. 4** Anomaly correlation coefficients of 3-, 5-, 7- and 10-day ECMWF 500 hPa height forecasts for the extra-tropical Northern and Southern Hemispheres, plotted in the form of annual running means of archived monthly-mean scores for the period from January 1980 to November 2006. Values plotted for a particular month are averages over that month and the 11 preceding months. The *coloured shadings* show the differences in scores between the two hemispheres at the forecast ranges indicated (after Simmons and Hollingsworth 2002)

the Northern Hemisphere. Bearing in mind the relatively poor data coverage provided by the conventional observing system (in particular over the oceans), this result is a strong indication of an improved usage of satellite data in the ECMWF assimilation system. The latest OSEs performed at ECMWF suggest that satellite observations add 3 days of forecast skill in the Southern Hemisphere, against two-third to three-quarter of a day in the Northern Hemisphere. The impact of satellite data in regional NWP models remains more contrasted, and much work is needed to improve their usage at high resolution and over land and sea ice.

## 6 Final Remarks

Current data assimilation methods enable effective use of a much wider range of observations than was previously possible (chapter *The Global Observing System*, Thépaut and Andersson). In this chapter we have explored some of the new possibilities offered with respect to:

- Non-linear and multivariate observation operators;
- Non-Gaussian observation error distributions and quality control;
- Flow-dependent background error covariances.

These aspects are general features that presently play a role in the utilization of conventional as well as satellite measurements. The variety of satellite measurements will continue to increase very substantially in the coming years. Some of the future data types are likely to put even higher demands on the data assimilation scheme. Indirect, remotely-sensed measurements of the atmosphere are often ambiguous with respect to the analysed quantities. In a variational scheme the ambiguities in the inversion of the data are resolved statistically by reference to the background information and information from coincident observations. This process will work accurately only if the covariance statistics of background and observation errors are specified accurately. Continued progress in the formulation of the background term is therefore likely to be just as important to the accuracy of the data assimilation as the addition of new data.

A further challenge for the near future is to incorporate observations of clouds and precipitation, with a view to correct the diabatic processes in the assimilation. Satellite observations provide information on the location and intensity of convection, as well as estimates of precipitation rates. These are especially valuable for the tropical analysis. Some promising results have already been obtained and have led to operational implementation in some NWP centres. Bauer et al. (2006a, b) have indeed shown that the assimilation of radiances from microwave imager data could improve the model spin-up and to a lesser extent the analysis error in terms of tropical cyclone location. In the long term, the current data assimilation methods will need to be further enhanced to enable a full feedback from satellite observations in presence of clouds and rain on the temperature, humidity and wind corrections to the analysis.

## References


Andersson, E., M. Fisher, R. Munro and A. McNally, 2000. Diagnosis of background errors for radiances and other observable quantities in a variational data assimilation scheme, and the explanation of a case of poor convergence. *Q. J. R. Meteorol. Soc.,* **126**, 1455–1472.

Andersson, E., A. Hollingsworth, G. Kelly, P. Lönnberg, J. Pailleux and Z. Zhang, 1991. Global observing system experiments on operational statistical retrievals of satellite sounding data. *Mon. Weather Rev.*, **119**, 1851–1864.

Andersson, E. and H. Järvinen, 1999. Variational quality control. *Q. J. R. Meteorol. Soc.* **125**, 697–722.

Andersson, E., J. Pailleux, J.-N. Thépaut, J. Eyre, A.P. McNally, G. Kelly and P. Courtier, 1994. Use of cloud-cleared radiances in three/four-dimensional variational data assimilation. *Q. J. R. Meteorol. Soc.,* **120**, 627–653.

Bauer, P., P. Lopez, A. Benedetti, D. Salmond and E. Moreau, 2006a. Implementation of 1D+4D-Var assimilation of precipitation affected microwave radiances at ECMWF. I: 1D-Var. *Q. J. R. Meteorol. Soc.*, **132**, 2307–2332.

Bauer, P., P. Lopez, D. Salmond, A. Benedetti, S. Saarinen and M. Bonazzola, 2006b. Implementation of 1D+4D-Var assimilation of precipitation affected microwave radiances at ECMWF. II: 4D-Var. *Q. J. R. Meteorol. Soc*., **132**, 2307–2332.

Bengtsson, L., 1980. On the use of a time sequence of surface pressures in four-dimensional data assimilation. *Tellus*, **32,** 189–197.

Bouttier, F. and G. Kelly, 2001. Observing-system experiments in the ECMWF 4-DVAR assimilation system. *Q. J. R. Meteorol. Soc.*, **127**, 1469–1488.

Cardinali, C., S. Pezzulli and E. Andersson, 2004. Influence matrix diagnostic of a data assimilation system. *Q. J. R. Meteorol. Soc.*, **130**, 2767–2786.

Chapnik, B., G. Desroziers, F. Rabier and O. Talagrand, 2006. Diagnosis and tuning of observational error in a quasi operational data assimilation setting. *Q. J. R. Meteorol. Soc.*, **132**, 543–565.

Chédin, A. and N.A. Scott, 1985. Initialization of the radiative transfer equation inversion problem from a pattern recognition type approach. In *Advances in Remote Sensing Retrieval Methods.* Application to the satellites of the TIROS-N series, A. Deepak (ed.), Academic Press, New York, pp 495–515.

Dee, D.P., L. Rukhovets, R. Todling, A.M. da Silva and J.W. Larson, 2001. An adaptive buddy check for observational quality control. *Q. J. R. Meteorol. Soc.*, **127**, 2451–2471.

Desroziers, G., L. Berre, B. Chapnik and P. Poli, 2005. Diagnosis of observation, background and analysis-error statistics in observation space. *Q. J. R. Meteorol. Soc.*, **131**, 3385–3396.

Dharssi, I., A.C. Lorenc and N.B. Ingleby, 1992. Treatment of gross errors using maximum probability theory. *Q. J. R. Meteorol. Soc.*, **118**, 1017–1036.

Eyre, J.R., 1989. Inversion of cloudy satellite sounding radiances by nonlinear optimal estimation. *Q. J. R. Meteorol. Soc.*, **115**, 1001–1037.

Eyre, J.R., 1991. A fast radiative transfer model for satellite sounding systems. *ECMWF Tech. Memo*, **176**, *available from ECMWF*.

Eyre, J.R. and A.C. Lorenc, 1989. Direct use of satellite sounding radiances in numerical weather prediction. *Meteorol. Mag.*, **118**, 13–16.

Fisher, M. and P. Courtier, 1995. Estimating the covariance matrices of analysis and forecast error in variational data assimilation. *ECMWF Tech. Memo.*, **220**, *available from ECMWF*.

Fleming, H.E., M.D. Goldberg and D.S. Crosby, 1986. Minimum variance simultaneous retrieval of temperature and water vapor from satellite radiance measurements. In *Proceedings of 2nd Conference on "Satellite Meteorology – Remote Sensing and Applications*", Williamsburg, American Meteorological Society, Boston, pp 20–23.

Flobert, J.-F., E. Andersson, A. Chédin, A. Hollingsworth, G. Kelly, J. Pailleux and N.A. Scott, 1991. Global data assimilation and forecast experiments using the Improved Initialization Inversion method for satellite soundings. *Mon. Weather Rev.*, **119**, 1881–1914.

Gelman, A., J.B. Carlin, H.S. Stern and D.B. Rubin, 1995. *Bayesian Data Analysis*. Texts in Statistical Science, Chapman and Hall, London.

Healy, S.B., J.R. Eyre, M. Hamrud and J.-N. Thépaut, 2006. Assimilating GPS radio occultation measurements with two-dimensional bending angle observation operators. *EUMETSAT/ECMWF Fellowship Programme Research Reports*, **16**, p. 21.

Hólm, E.V., A. Untch, A. Simmons, R. Saunders, F. Bouttier and E. Andersson, 1999. Multivariate ozone assimilation in four-dimensional data assimilation. In *Proceeding SODA Workshop on "Chemical Data Assimilation"*, de Bilt, The Netherlands, 9–10 December 1998, pp 89–94.

Huber, P.J., 1977. *Robust Statistical Methods*. Society for Industrial and Applied Mathematics, Pennsylvania, USA.

Ingleby, N.B. and A.C. Lorenc, 1993. Bayesian quality control using multivariate normal distributions. *Q. J. R. Meteorol. Soc.*, **119**, 1195–1225.

Järvinen, H., E. Andersson and F. Bouttier, 1998. Variational assimilation of time sequences of surface observations with serially correlated errors. *ECMWF Tech. Memo.*, **266**, *available from ECMWF*.

Järvinen, H. and P. Undén, 1997. Observation screening and first guess quality control in the ECMWF 3D-Var data assimilation system. *ECMWF Tech. Memo*., **236**, *available from ECMWF*.

Köpken, C., G. Kelly and J.-N. Thépaut, 2004. Assimilation of Meteosat radiance data within the 4D-Var system at ECMWF: Assimilation experiments and forecast impact. *Q. J. R. Meteorol. Soc.*, **130**, 2277–2292.

Lorenc, A.C., 1986. Analysis methods for numerical weather prediction. *Q. J. R. Meteorol. Soc.*, **112**, 1177–1194.

McNally, A.P., E. Andersson, G.A. Kelly and R.W. Saunders, 1999. The use of raw TOVS/ATOVS radiances in the ECMWF 4D-Var assimilation system. *ECMWF Newsletter,* **83**, pp 2–7.

McNally, A.P., J.C. Derber, W. Wu and B.B. Katz, 2000. The use of TOVS level-1B radiances in the NCEP SSI analysis system. *Q. J. R. Meteorol. Soc.*, **126**, 689–724.

Munro, R., G. Kelly, M. Rohn and R. Saunders, 1999. Assimilation of geostationary water vapour radiance data at ECMWF. *Technical Proceeding of the 10th international*, Boulder, Colorado, 27 January–2 February 1999.

Poli, P., P. Moll, F. Rabier, G. Desroziers, B. Chapnik, L. Berre, S.B. Healy, E. Andersson and F.-Z. El Guelai, 2007. Forecast impact studies of zenith total delay data from European near real-time GPS stations in Meteo-France 4DVar. *J. Geophys. Res.,* **112**, 10.1029/2006JD007430.

Rabier, F., H. Järvinen, E. Klinker, J.F. Mahfouf and A. Simmons, 2000. The ECMWF operational implementation of four-dimensional variational assimilation. Part I: Experimental results with simplified physics. *Q. J. R. Meteorol. Soc*. **126**, 1143–1170.

Rabier, F., E. Klinker, P. Courtier and A. Hollingsworth, 1996. Sensitivity of forecast errors to initial conditions. *Q. J. R. Meteorol. Soc.*, **122**, 121–150.

Rabier, F., J.-N. Thépaut and P. Courtier, 1998. Extended assimilation and forecast experiments with a four-dimensional variational assimilation system. *Q. J. R. Meteorol. Soc.*, **124**, 1861–1887.

Reale, A.L., D.G. Gray, M.W. Chalfant, A. Swaroop and A. Nappi, 1986. Higher resolution operational satellite retrievals. Preprints, *2nd Conference on "Satellite Meteorology/Remote Sensing and Applications"*, Williamsburg, 13–16 May 1986, American Meteorological Society, Boston, pp 16–19.

Riishøjgaard, L.P., 1996. On four-dimensional variational assimilation of ozone data in weather prediction models. *Q. J. R. Meteorol. Soc*. **122**, 1545–1571.

Rodgers, C.D., 1976. Retrieval of atmospheric temperature and composition from remote measurements of thermal radiation. *Rev. Geophys. Space Phys.*, **14**, 609–624.

Rodgers, C.D., 1990. Characterization and error analysis of profiles retrieved from remote sounding measurements. *J. Geophys. Res.,* **95**, 5587–5595.

Saunders, R., M. Matricardi and P. Brunel, 1999. An improved fast radiative transfer model for assimilation of satellite radiance observations. *Q. J. R. Meteorol. Soc*. **125**, 1407–1426.

Simmons, A.J. and A. Hollingsworth, 2002. Some aspects of the improvement in skill of numerical weather prediction. *Q. J. R. Meteorol. Soc*. **128**, 647–677.

Thépaut, J.-N., P. Courtier, G. Belaud and G. Lemaître, 1996. Dynamical structure functions in a four-dimensional variational assimilation: A case study. *Q. J. R. Meteorol. Soc*. **122**, 535–561.

Thépaut, J.-N., R.N. Hoffman and P. Courtier, 1993. Interactions of dynamics and observations in a four-dimensional variational assimilation. *Mon. Weather Rev.*, **121**, 3393–3414.

WMO, 2004. *Proceedings of the 3rd WMO Workshop on the Impact of Various Observing Systems on Numerical Weather Prediction*, Alpbach 9–12 March 2004. Böttger, M. and Pailleux, J. (eds.) WMO/TD No. 1228.

# Research Satellites

**William Lahoz**

## 1 Introduction

Our knowledge of the Earth System ultimately comes from observations. Although observations have uncertainties and biases, they are the "truth" against which theories and models must be confronted and evaluated. Predictions of the variability of the Earth System require an understanding of the variability in the observations representing the "truth". This requires observations that are: (i) high quality, i.e., have small errors and biases; (ii) consistent, i.e., there is uniformity in the observing system characteristics; and (iii) long-term, i.e., the results have statistical significance. Depending on application, there may be further observational requirements. For example, for monitoring climate change, global coverage would generally be required; for studying high impact weather, high spatial and temporal resolution would generally be required.

## 2 Observations

Examples of observing platforms include ground-based instruments, sondes, balloons, aircraft (collectively known as in situ instruments) and remote sounding satellites, typically divided into operational satellites (focused on Numerical Weather Prediction, NWP) and research satellites (typically focused on research of the Earth System). However, the distinction between operational and research satellites is becoming blurred, as more research satellites are used operationally. Collectively, the satellite and in situ data form the Global Observing System (GOS). Currently, the observations of NWP centres such as the Met Office and the European Centre for Medium-Range Weather Forecasts (ECMWF) come from in situ instruments and satellite platforms, with the latter dominated by nadir-viewing satellites. This chapter focuses on research satellite data. It thus complements the chapters *The*

W. Lahoz (✉)
Norsk Institutt for Luftforskning, Norwegian Institute for Air Research, NILU, Kjeller, Norway
e-mail: wal@nilu.no

W. Lahoz et al. (eds.), *Data Assimilation*, DOI 10.1007/978-3-540-74703-1_12,

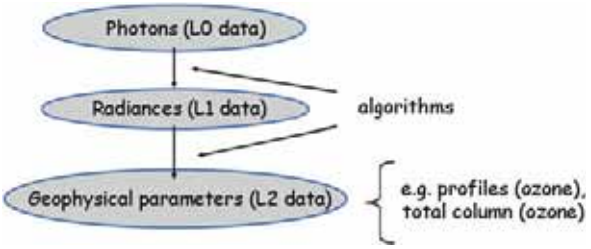



**Fig. 1** Schematic illustrating the transformation of level 0 (L0) data into level 1 (L1) data and level 2 (L2) data

*Global Observing System* (Thépaut and Andersson) and *Assimilation of Operational Data* (Andersson and Thépaut), which focus on the GOS and the assimilation of operational data, respectively.

It is important to realize that satellite instruments do not measure directly temperature, ozone or similar geophysical parameters (see chapter *The Global Observing System*, Thépaut and Andersson). What they measure is photon counts (*level* 0 data). Algorithms then transform the level 0 data into radiances (*level* 1 data). Subsequently, using retrieval techniques (Rodgers 2000), retrievals of profiles or total column amounts are derived (*level* 2 data) – see Fig. 1. It is the level 2 data that many Earth Observation (EO) scientists use as the starting point for their studies. Fields derived from manipulation of level 2 data, e.g., by interpolation to a common grid are termed *level* 3 data. Analyses derived from the assimilation of level 1 and/or 2 data are termed *level* 4 data. The use of level 4 data is becoming more common in the EO community.

Level 2 data from a satellite instrument is not a point measurement, but instead represents an observation which is representative of a finite volume in the atmosphere, the dimensions of this volume determined by the horizontal and vertical resolution of the measurement. The so-called averaging kernel (see also chapter *The Global Observing System*, Thépaut and Andersson) provides information on how measurements represent an "average" over a particular volume (Rodgers 2000).

Level 2 data (as well as the level 1 and level 0 data) have associated with them a number of errors, including *random* and *systematic* errors, and the error of *representativeness* (or representativity). Random errors (sometimes termed *precision*) have the property that averaging the data can reduce them. This is not the case of the systematic error or bias (sometimes termed *accuracy*). The error of representativeness is associated with the extent to which the "average" measurement represents a measurement within any point of the finite volume for which this "average" is appropriate.

## 3 Research Satellite Data

### *3.1 General Considerations*

Observing platforms have advantages and disadvantages. For example satellite observations have low spatial and temporal resolution but good global coverage,

whereas in situ data have high spatial and temporal resolution but poor global coverage. For observing instruments aboard satellite platforms, nadir sounders have good horizontal resolution but poor vertical resolution, whereas limb sounders have poor horizontal resolution but good vertical resolution (Fig. 2). Currently, the observations used by the Met Office and ECMWF for NWP come from: aircraft, nadir-viewing satellites, sondes, and the surface, with some observations from limb-viewing satellites (see, e.g., http://www.ecmwf.int/research/ifsdocs/CY28r1/index.html; and chapters *The Global Observing System*, Thépaut and Andersson; *Assimilation of Operational data,* Andersson and Thépaut). The use of such a mixture of observation platforms is typical for met agencies.

As discussed in chapter *The Global Observing System* (Thépaut and Andersson), broadly speaking satellite observations of the Earth/atmosphere fall into the following categories: (a) passive technologies; (b) active technologies; (c) and limb technologies. Passive and active technologies can be used to sense the atmosphere (*sounding instruments*) or the Earth's surface (*imaging instruments*). Limb technologies (i.e., those with a limb-viewing geometry) can be divided into limb passive sounders and GPS (Global Positioning System) radio occultation technologies.

Satellites can be divided into *operational* and *research* types. Operational satellites provide data in near-real-time (NRT; for ECMWF data up to 17 h is used for various purposes, but for the NWP four dimensional variational, 4D-Var, forecast the cut-off can be as short as 1 h or less), and are used in NWP. Currently, the vast majority of operational satellites have nadir-viewing geometries. Research satellites provide data off-line (typically more than 2 days after data acquisition), use both nadir- and limb-viewing geometries, and are mainly used by the scientific community in, e.g., studies of climate change and attribution, and studies of ozone depletion. Recently, research satellite data have begun to be of interest to the NWP community (a requisite is their availability in NRT). A particular example has been the production of operational ozone analyses since April 2002 by ECMWF using GOME and SCIAMACHY total column ozone data and MIPAS height-resolved ozone data.

Satellites can also be classified according to their orbits: (a) geostationary (GEO); (b) low Earth orbit (LEO) – also termed polar orbiting satellites. GEO and LEO satellites are discussed in chapter *The Global Observing System* (Thépaut

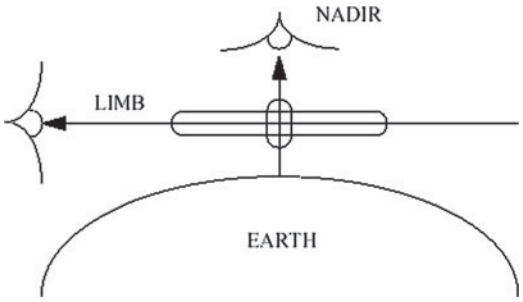



**Fig. 2** Schematic illustrating nadir- and limb-viewing geometries. The *ovals* represent the volume associated with typical horizontal and vertical resolutions. Based on Lahoz (2003)

and Andersson). Research satellites generally have LEO orbits, and can have sun-synchronous or non sun-synchronous orbits.

Sun-synchronous satellites have a fixed Equator crossing time, whereas non-sun-synchronous satellites do not. Sun-synchronous satellites (e.g. ESA's Envisat and NASA's EOS Aura) have the advantage that the instruments always face away from the sun, so that no periodic satellite yaw manoeuvre to avoid sunlight damaging the instruments is necessary. However, they have the disadvantage that they cannot observe the diurnal cycle at a particular location. For example, NO and $NO_2$, which play a role in determining the distribution of ozone, have strong diurnal cycles.

Non-sun-synchronous satellites (e.g. NASA's UARS) have the advantage of being able to observe the diurnal cycle at a particular location, but the disadvantage that a periodic satellite yaw manoeuvre is needed to avoid sunlight damaging the instruments. In the case of the MLS instrument aboard UARS, this had the effect of the instrument either having a North Looking configuration (34°S–80°N) or a South Looking configuration (80°S–34°N).

We now describe in more detail the various research satellites used for Earth Observation, EO.

## *3.2 Research Satellites*

Most of the Research & Development (R&D) space agencies in the world: NASA (USA), ESA (Europe), JAXA (Japan) and CSA (Canada) have launched research satellites over the last 10 years, and are involved in plans to launch research satellites over the next 5–10 years. The following list of research satellites is meant to be illustrative and is not exhaustive. Acronyms are identified in the *Appendix*. The websites mentioned below have been checked on July 2009.

*NASA.* NASA is involved with UARS and Earth Probe (EP) TOMS (there have been several TOMS instruments, of which EP is one example, all measuring total column ozone), and is involved in the EOS series of satellites (Terra, Aqua and Aura).

NASA's UARS was launched in September 1991, and ceased operations in December 2005. The longest lasting of the UARS atmospheric composition instruments, HALOE, made its last occultation sounding in November 2005. UARS carried a suite of instruments making measurements of the stratosphere and mesosphere that have been used to infer meteorological and chemical fields. A number of UARS limb and occultation sounder instruments (CLAES, HALOE, HRDI, ISAMS, MLS) have made measurements of temperature, ozone, water vapour, ClO and winds. The MLS instrument has also made measurements of upper troposphere water vapour. The UARS data have been extensively evaluated (UARS special issue in *J. Geophys. Res.*, 1996, Vol. **101**, 9539–10473), and have contributed to our understanding of many aspects of the atmospheric circulation and chemistry (e.g. the UARS special issue in *J. Atmos. Sci.*, 1994, Vol. **51**, 2781–3105).

EOS Terra was launched in December 1999. It carries on board 5 instruments: ASTER, CERES, MISR, MODIS and MOPITT. EOS Terra provides information on: (i) land surface, water and ice (ASTER); (ii) radiation (CERES); (iii) radiation and biosphere parameters (MISR); (iv) biological and physical processes on land and the ocean (MODIS); and (v) CO and $CH_4$ in the troposphere, where they are pollution markers (MOPITT). An example of the impact of EOS Terra comes from MOPITT, which has provided the first global CO (proxy for air pollution) measurements from space (http://www.atmosp.physics.utoronto.ca/MOPITT/home.html).

EOS Aqua (http://aqua.nasa.gov) was launched in May 2002. It carries on board 6 instruments: AMSR/E, MODIS, AMSU, AIRS, HSB and CERES. EOS Aqua provides information on: (i) clouds, radiation and precipitation (AMSR/E); (ii) clouds, radiation, aerosol and biosphere parameters (MODIS); (iii) temperature and humidity (AMSU, AIRS, HSB); and (iv) radiation (CERES). HSB was lost early in the mission. The NWP community has assimilated clear-sky radiances from AIRS that are sensitive to temperature and humidity. The EOS Aqua data have been described in the literature (EOS Aqua special issue in *IEEE*, 2003, Vol. **41**).

EOS Aura (http://aura.gsfc.nasa.gov) was launched in July 2004. It carries on board 4 instruments: EOS MLS, HIRDLS, TES and OMI. EOS Aura provides information on: (i) chemistry of the upper troposphere and lower stratosphere (UTLS), chemistry of the middle and upper stratosphere, upper troposphere water, and the impact of volcanoes on global change (EOS MLS); (ii) temperature and constituents in the upper troposphere, stratosphere and mesosphere (HIRDLS); (iii) global maps of tropospheric ozone and its photochemical precursors (TES); and (iv) maps of total column ozone (which continue the TOMS record), $NO_2$ and UV-B radiation (OMI). EOS MLS (which builds upon the experience with UARS MLS) and HIRDLS are limb sounders (post-launch problems have limited the measurement performance of HIRDLS); TES can be used in both nadir and limb sounder mode (although due to technical problems only the nadir mode is functional after May 2005); OMI is a nadir sounder. One of the many innovative aspects of EOS Aura is the near-real-time production of total column ozone data from OMI (http://www.temis.nl).

The EOS Aura data have been described in the literature (EOS Aura special issue in *IEEE*, 2006, Vol. **44**), and in a special issue on EOS Aura validation in *J. Geophys. Res.*, 2008, Vol. **113** (see also Schoeberl et al. 2008).

EOS Aqua and EOS Aura are part of the EOS "A-Train" (http://www.spacetoday.org/Satellites/TerraAqua/ATrain.html). The "A-Train" refers to the constellation of USA satellites and international Earth Science satellites that plan to fly together with EOS Aqua to enable co-ordinated science observations. These satellites have an afternoon Equator crossing time close to the mean local time of the "lead" satellite, EOS Aqua, which is 1:30 pm – thus the name, "A" being short for afternoon. The A-Train consists of, in temporal order of their afternoon Equator crossing time: EOS Aqua (1:30 pm); CloudSat (1:31 pm); CALIPSO (1:31:15 pm); PARASOL (1:33 pm); and EOS Aura (1:38 pm). Substantial scientific activities are being undertaken to exploit the *synergy* and *complementarity* between the instruments on board the A-train.

Cloudsat provides information on the altitude and properties of clouds; CALIPSO and PARASOL provide information on clouds and aerosol. The instruments on EOS Aqua and EOS Aura are described above.

OCO, which would have provided information on $CO_2$ and was launched in February 2009, unfortunately suffered a technical failure and was lost shortly after launch (Palmer and Rayner 2009). It was scheduled to have been part of the A-train (afternoon Equator crossing of 1:15 pm).

*ESA*. ESA is involved with the ERS-2 satellite, which carries the GOME instrument; the ODIN mission and Envisat. GOME is a nadir sounder that has been making measurements of total column ozone and $NO_2$ since 1995. Since June 2003 the ERS-2 satellite has experienced problems. Details of these problems and action taken by ESA for can be found in http://earth.esa.int/pub/GOME/YEARLY/anomalies_2007.html

ODIN, also involving the CSA (the Canadian Space Agency), CNES (the French Space Agency) and SNSB (the Swedish Space Agency), was launched in February 2001. It carries on board 2 instruments: OSIRIS and SMR. ODIN is providing information on ozone and $NO_2$ (total columns and profiles).

Envisat (http://envisat.esa.int) was launched in March 2002 (GMT). It carries on board 10 instruments: AATSR, ASAR, DORIS, GOMOS, LRR, MERIS, MIPAS, MWR, RA-2 and SCIAMACHY. Envisat provides information on: (i) temperature, ozone, water vapour and other atmospheric constituents using limb, nadir and occultation geometries (MIPAS, SCIAMACHY, GOMOS); (ii) aerosol (AATSR, MERIS); (iii) sea surface temperature (AATSR); (iv) sea colour (MERIS); (v) land and ocean images (ASAR); (vi) land, ice and ocean monitoring (RA-2); (vii) water vapour column and land surface parameters (MWR); and (viii) cryosphere and land surface parameters (DORIS). LRR is used to calibrate RA-2. The MIPAS instrument has suffered problems since 2004 which have affected its performance. The broad spectrum of information from Envisat reflects a paradigm that the Earth System should be treated as a whole, and that information from its various components should be integrated. However, the complexity and cost of Envisat mean it is unlikely that ESA will launch future missions of a size similar to Envisat.

One of the innovative aspects of the Envisat mission has been the use of data assimilation techniques to evaluate the atmospheric chemistry instruments (GOMOS, MIPAS, SCIAMACHY); see chapter *Constituent Assimilation* (Lahoz and Errera).

The Envisat data has been evaluated at a series of workshops organized by ESA. Examples include the Envisat Validation Workshop held at ESRIN on 9–13 December 2002 (http://envisat.esa.int/pub/ESA_DOC/envisat_val_1202/proceedings; ESA Special Publication SP-531); the Second Workshop on the Atmospheric Chemistry Validation of Envisat, ACVE-2, held at ESRIN on 3–7 May 2004 (http://envisat.esa.int/workshops/acve2; ESA Special Publication SP-562); and the Third Workshop on the Atmospheric Chemistry Validation of Envisat, ACVE-3, held at ESRIN on 4–7 December 2006 (ESA Special Publication SP-642). The data from MIPAS and SCIAMACHY have also been evaluated in special issues in *Atmos. Chem. Phys.*: (i) Geophysical Validation of

SCIAMACHY 2002–2004 (Eds. Kelder, Platt and Simon), http://www.atmos-chem-phys.net/special_issue19.html (2005); and (ii) MIPAS (Michelson Interferometer for Passive Atmospheric Sounding): Potential of the experiment, data processing and validation of results (Eds. Espy and Hartogh), http://www.atmos-chem-phys.net/special_issue70.html (2006). Data from the atmospheric chemistry instruments in Envisat have been used to study the unprecedented Antarctic ozone hole split of September 2002 (special issue in *J. Atmos. Sci.*, 2005, Vol. **62**).

Other current and future missions from ESA include GOCE (launched March 2009); SMOS, launched in November 2009; Cryosat-2 (a replacement to the aborted Cryosat mission), due for launch in 2010; SWARM, due for launch in 2011; ADM-Aeolus, due for launch in 2011; and EarthCARE (in collaboration with JAXA, the Japanese space agency), due for launch in 2013 (all these dates as of March 2010; see, e.g., http://www.esa.int/esaLP/LPearthexp.html). These six missions (Earth Explorers) are part of ESA's Living Planet Programme (http://www.esa.int/esaLP); GOCE, ADM-Aeolus and EarthCARE are core missions; Cryosat-2, SMOS and SWARM are opportunity missions. GOCE will measure the Earth's gravity field; SMOS will measure soil moisture over the Earth's land masses and salinity over the oceans; Cryosat-2 will measure cryosphere parameters; SWARM will measure the Earth's geomagnetic field and its temporal evolution; ADM-Aeolus will measure the line-of-sight wind in the troposphere and lower stratosphere; and EarthCARE will provide vertical profiles of clouds and aerosols, and radiances at the top of the atmosphere. All these six missions include novel measurements and/or novel measurement techniques.

In May 2006, six new Earth Explorer Missions were selected by ESA for further study: BIOMASS (global measurements of forest biomass); TRAQ (monitoring air quality and transport of long-range transport of pollutants); PREMIER (understanding processes that link trace gases, radiation, chemistry and climate in the atmosphere); FLEX (observing global photosynthesis through the measurement of fluorescence); A-SCOPE (improving understanding of the global carbon cycle and regional $CO_2$ fluxes); and $CoReH_2O$ (detailed observations of key snow, ice and water cycle characteristics). These missions were assessed at a User Consultation Meeting in January 2009: PREMIER, BIOMASS and $CoReH_2O$ were selected to proceed to the next phase of development and undergo feasibility (Phase A) studies. It is expected that only one of these missions will eventually fly.

The GMES (Global Monitoring for Environment and Security) programme, which aims for full operational provision of satellite data for GMES services, involves the use of existing and planned national space capabilities as well as the development of new infrastructure. GMES will be developed taking into account the activities of the Group on Earth Observations (GEO; http://www.earthobservations.org/). With its federating role, GMES will be the main European contribution to the global 10-year implementation plan for the Global Earth Observing System of Systems (GEOSS).

The GMES Space Component program is intended to meet the requirements of the three pilot services identified by the EC for early implementation (land monitoring, ocean monitoring and emergency management) and other services to be

deployed in the 2008–2020 period. The GMES Space Component programme is built around five concepts of space missions or GMES "Sentinels" (see below for details), plus access to existing and complementary missions from ESA Member States, EUMETSAT, Canada and third parties. The following complementary missions are considered candidates for GMES operational service contributions in order to get the programme started: SPOT-5 (CNES); TerraSAR-X (DLR/EADS Astrium, Germany); COSMO-SkyMed (ASI, Italy); RADARSAT-2 (CSA/MDA, Canada); Pleiades (CNES); Jason-2 (EUMETSAT/CNES/NOAA/NASA); MSG (EUMETSAT); MetOp (EUMETSAT); DMC – Disaster Monitoring Constellation (SSTL, UK); RapidEye (RapidEye AG, Germany) and EnMAP – Environmental Mapping and Analysis Program (hyperspectral mission from DLR).

The following members of the Sentinel family have been identified as core elements of the GMES Space Component:

- *Sentinel 1 – C-Band SAR mission*: This is a spacecraft in sun-synchronous orbit at a mean altitude of 693 km, with a 12 day repeat cycle, and a Synthetic Aperture Radar (SAR) operating in C-band. It will have four nominal operation modes: strip map (80 km swath, 5 × 5 m resolution); interferometric wide swath (250 km swath, 20 × 5 m resolution); extra wide swath (400 km swath, 25 × 100 m resolution); and wave (5 × 20 m resolution). Applications include: monitoring sea ice zones and the Arctic environment; surveillance of the marine environment; monitoring land surface motion risks; and providing mapping in support of humanitarian aid in crisis situations. It is due to provide continuity to data hitherto provided by ERS-2, RADARSAT and the Envisat missions. Launch is planned for 2011 with a 7 years design lifetime. For more details see ESA (2005).
- *Sentinel 2 – Superspectral imaging mission*: This is a spacecraft in sun-synchronous orbit at a mean altitude of 786 km, with a 10 days repeat cycle. It will have a filter-based pushbroom multi-spectral imager with 13 spectral bands (Visible-Near Infrared, VNIR; Shortwave-Infrared, SWIR). It will have three spatial resolutions: 10, 20, and 60 m, and field of view of 290 km. Applications include: generic land cover maps; risk mapping and fast images for disaster relief; and generation of leaf coverage, leaf chlorophyll content and leaf water content. It is due to provide enhanced continuity to data hitherto provided by SPOT and Landsat. Launch is planned for 2012 with a 7 years design lifetime. For more details see ESA (2007a).
- *Sentinel 3 – Ocean and global land mission*: This is a spacecraft in sun-synchronous orbit at a mean altitude of 814.5 km over the geoid, with a 27 days repeat cycle. It has three sets of instruments: (i) Ocean and Land Colour Instrument (OLCI), with 5 cameras, 8 bands (only visible) for open ocean (low resolution), 15 bands (only visible) for coastal zones (high resolution), and spatial sampling of 300 m at the sub-satellite point (SSP); (ii) Sea and Land Surface Temperature (SLST), 9 spectral bands, 0.5 km (Visible; SWIR) to 1 km (Microwave-Infrared, MWIR; Thermal Infrared, TIR) resolution, and swath of 180 rpm dual view scan (nadir and backwards); (iii) a RA package including a Ku/C Radar Altimeter (SRAL), a Microwave Radiometer (MWR) and Precise

Orbit Determination (POD). Applications include: sea/land colour data and surface temperature; sea surface and land ice topography; coastal zones, inland water and sea ice topography; and vegetation products. It will provide SAR mode data over sea ice and coastal regions (hitherto provided by RA-2 on Envisat), and wide-swath low/medium resolution data from optical and infrared radiometers (hitherto provided by AATSR and MERIS on Envisat and Vegetation on SPOT). Launch is planned for 2012 with a 7 years design lifetime. For more details see ESA (2007b).

- *Sentinel 4 – GEO Atmospheric mission*: This is a GEO (geostationary orbit) mission with a European focus (satellite located at 0° longitude). It will have a narrow field spectrometer covering UV (290–400 nm), visible (400–500 nm) and near-infrared (NIR; 750–775 nm) bands. Spatial sampling will be 5–50 km and spectral resolution will be between 0.06 and 1 nm (depending on band). Applications include: monitoring changes in atmospheric composition (e.g. ozone, $NO_2$, $SO_2$, BrO, formaldehyde and aerosol); and tropospheric variability. It will be embarked on MTG-S and operated by EUMETSAT. Launch is planned for after 2017.
- *Sentinel 5 – LEO Atmospheric mission*: This is a LEO (low Earth orbit) mission with global coverage at a reference altitude of about 817 km. It will have a wide swath pushbroom spectrometer suite covering UV (270–400 nm), visible (400–500, 710–750 nm), near-infrared (NIR; 750–775 nm), and shortwave-infrared (SWIR; 2,305–2,385 nm) bands. Spatial sampling will be 5–50 km and spectral resolution will be between 0.05 and 1 nm (depending on band). Applications include: monitoring changes in atmospheric composition at high temporal, i.e., daily, resolution (e.g. ozone, $NO_2$, $SO_2$, BrO, formaldehyde and aerosol); and tropospheric variability. It will be embarked on post-EPS and operated by EUMETSAT. Launch is planned for after 2017. A Sentinel 5 precursor is planned for the period 2013–2019 to fill the data gap between the expected end of the Envisat and EOS Aura missions (before 2014) and the expected launch dates of MTG-S (2017) and post-EPS (2020). This data gap affects in particular short-wave measurements with sufficient quality for tropospheric applications. The Sentinel 5 precursor would be desirable for two reasons: to provide continuity of data, and to provide transition into operational implementation (e.g. Sentinel 5 precursor data could be combined with IASI and GOME-2 data). See ESA (2007c) for more details on Sentinels 4 and 5.

The actual implementation of the missions will be according to a flexible architecture that may lead to grouping some of them on single platforms. The whole programme spans the 2006–2018 timeframe and will be implemented in two Segments. Segment-1 (planned for 2006–2013) will have two funding lines: a joint ESA/EC funding line where ESA-procurement rules apply but with modification to account for EU financial regulations; an ESA-only funding line where the geographical return targets can be applied. Segment-2 (planned for 2009–2018) is expected to be co-funded by the EC and ESA. The information on GMES Sentinels presented above was valid as of December 2008; as often happens with satellite

missions, this could change. For updates on information on ESA missions see: http://news.eoportal.org/.

*Other agencies*: The Japanese Space Agency, JAXA (http://www.jaxa.jp/index_e.html), is involved with the ADEOS missions: ADEOS (launched 1996) and ADEOS-II (launched December 2002). The ADEOS mission carried several instruments on board, including ADEOS TOMS (which measured total column ozone) and ILAS (a limb instrument which measured temperature, ozone, water vapour and other atmospheric constituents). The ADEOS mission only lasted for 10 months.

The ADEOS-II mission carried on board 5 instruments: AMSR, GLI, SeaWinds, POLDER and ILAS-II. ADEOS-II provides information on: (i) water column, precipitation, and ocean and ice parameters (AMSR); (ii) land, ice and biosphere parameters (GLI); (iii) winds over the ocean (SeaWinds); (iv) radiation parameters (POLDER); and (v) temperature, ozone and other atmospheric constituents (ILAS-II). The ADEOS-II mission only lasted until October 2003. The ILAS-II products have been evaluated in several papers appearing in a special section of *J. Geophys. Res.* (Vol. **111**, 2006).

TRMM (http://trmm.gsfc.nasa.gov) is a joint project of Japan (JAXA) and the USA (NASA). It was launched in November 1997. It can observe the rainfall rate in the tropics and its horizontal and vertical distribution, which were not possible by the other measuring methods. TRMM data can help understand global change and implement environmental policies. TRMM Microwave Imager (TMI) data are also used operationally at ECMWF (and possibly other NWP centres) thanks to their availability in real time.

GOSAT is a JAXA satellite in collaboration with the Japanese Ministry of the Environment (MOE) and NIES (National Institute for Environmental Studies); it was launched in January 2009 (http://www.jaxa.jp/press/2009/01/20090124_ibuki_e.html). It is designed to observe the global distribution of $CO_2$ and $CH_4$, and is expected to contribute to international efforts to prevent global warming by acquiring information on absorption and emission levels of greenhouse gases.

GCOM are a series of JAXA satellites designed to study the global water cycle (GCOM-W) and the global carbon cycle (GCOM-C). GCOM-W will carry the AMSR-2 instrument. It will fly in a sun-synchronous orbit with equatorial crossing time of 1:30 pm, close to that of the EOS Aqua satellite (see above); it is designed to extend the measurements of the AMSR-E instrument on the EOS Aqua platform. There will be three consecutive generations of satellites, with 1 year overlap; the first satellite is due to be launched in 2012. GCOM-C will carry the SGLI instrument. It will also fly in a sun-synchronous platform with equatorial crossing time of 10:30 am. It is due to be launched in 2014.

The Canadian Space Agency, CSA (http://www.space.gc.ca/index.html), is involved with the SCISAT-1 platform, a solar occultation sounder. It includes ACE and MAESTRO. It was launched in 2003, and the mission has been extended to 2009. The ACE data products have 14 baseline species, including: ozone, water vapour, methane, $N_2O$ and $NO_2$.

The validation of SCISAT-1 data is taking place in two phases. In the initial phase, several papers were published in a *Geophys. Res. Lett.* special issue (Vol. **32**, 2005); the second phase is ongoing. There are eight validation groups, covering, e.g., ozone and methane. There is a special issue in *Atmos. Chem. Phys.* (papers appearing in 2008–2009) describing the validation of SCISAT-1/ACE products (http://www.atmos-chem-phys.net/special_issue114.html, eds. A. Richter, T. Wagner). Plans for ACE-II are being considered.

The CSA SWIFT instrument was a candidate for launch in the time frame of 2015 and beyond but, unfortunately, the instrument has recently been shelved (Richard Ménard, personal communication). It was intended to measure stratospheric winds and ozone (see http://swift.yorku.ca). For illustrative purposes of how to quantify future additions to the GOS, we can mention an Observing System Simulation Experiment (OSSE; see chapter *Observing System Simulation Experiments*, Masutani et al.) carried out to evaluate the incremental benefit to the GOS of SWIFT measurements (Lahoz et al. 2005). It was found that SWIFT measurements would benefit the GOS, and be useful for scientific studies of, e.g., stratospheric variability.

The Chinese Space Agency SBUS and TOU instruments are essentially copies of SBUV (SBUS) and TOMS (TOU). These missions are coordinated by the Chinese National Satellite Meteorological Centre (NSMC; http://www.fas.org/spp/guide/china/agency/nsmc.htm). They will be flown on the FY-3 (Fengyun-3) platform, which is a polar orbiter, with a 10:10 am Equator crossing time. The first FY-3 satellite was launched in May 2008 (see http://www.sinodefence.com/space/spacecraft/fengyun3.asp); the next satellite is due to be launched in the timeframe of 2009 or later; there is the possibility of a third satellite. (According to the China Meteorological Administration, CMA, a further satellite series, FY-4, is planned for launch by 2013 and beyond.) NSMC has requested that NOAA help them with the data processing algorithms.

Other space agencies likely to increase their profile over the next few years, and not mentioned above, include the Argentinian Space Agency (CONAE), the Russian Federal Space Agency, and the National Space Agency of Ukraine. A list of space agencies is provided in: http://en.wikipedia.org/wiki/List_of_space_agencies.

## *3.3 Benefits of Research Satellites*

Research satellites provide several benefits. Because they often have both limb- and nadir-viewing instruments, they allow the combination of limb/nadir geometries to provide better atmospheric analyses (see, e.g., Struthers et al. 2002), and provide information on tropospheric ozone (see, e.g., Lamarque et al. 2002), which is very difficult to measure from space. Because they have instruments which focus on measurements of ozone and of photochemical species which affect the ozone distribution, they provide information for studying stratospheric ozone depletion, and information that helps develop coupled climate/chemistry models, and a chemical forecasting capability (important for UV and pollution forecasting) – see chapter

*Inverse Modelling and Combined State-Source Estimation for Chemical Weather* (Elbern et al.). Increased interest by the operational centres in ozone and chemical forecasting makes research satellites more attractive to them. An example of this interest has been the development of algorithms at ECMWF to assimilate limb radiances sensitive to ozone and humidity (see Lahoz et al. 2007b for a summary).

The combined use of research and operational satellite data also provides opportunities for synergy. For example, different viewing geometries (limb and nadir) can be used with techniques such as data assimilation to improve the representation of the atmosphere, and partition information between the stratosphere and troposphere. Synergy between research and operational satellites, and the potential benefits to the operational agencies accruing from this synergy, can make it attractive to use research satellites in an operational capability. This can happen in a number of ways: (i) one-off use of research satellite data (e.g. measurement of a key photochemical species such as ozone, or of a novel geophysical parameter such as stratospheric winds); (ii) regular use of research satellite data (e.g. a satellite series that can extend the time record of key geophysical parameters such as ozone and water vapour); and (iii) use of the research satellite instrument design in future operational missions.

Finally, it is worth insisting on the *complementarity* of the research and the operational approach to satellite data, in particular in the sense that often research and development instruments are precursors of operational instruments. Thus, operational centres exercise the science on research satellites to improve their readiness when operational satellites come by. The best illustration of this is provided by the AIRS and IASI instruments. The science community, in particular at the operational centres, was able to use AIRS to prepare for the assimilation of data from multi-spectral sounders, and this minimized the delay when IASI started to provide data to NWP centres (see chapter *The Global Observing System*, Thépaut and Andersson).

## *3.4 Research Satellites and the Global Climate Observing System*

Global monitoring of climate requires products derived from satellite measurements (GCOS-92 2004). A GCOS (Global Climate Observing System) Implementation Plan (GIP) has been proposed (GCOS-92 2004; GCOS-107 2006). This aims to set up an observing system that provides information of the *Essential Climate Variables* (ECVs – see Table 1 below) and their associated products that are needed to assist signatory countries of the UNFCCC (United Nations Framework Convention on Climate Change; "Parties") in meeting their responsibilities under the UNFCCC. The proposed system will provide information to: (i) characterize the state of the global climate system and its variability; (ii) monitor forcing of the climate system, including both natural and anthropogenic contributions; (iii) support attribution of the causes of climate change; (iv) support prediction of global climate change; (v) enable down-scaling of global climate change information to regional

**Table 1** Earth system domain and ECVs

| Domain | Essential climate variables |
|---|---|
| Atmospheric (over land, sea and ice) | Precipitation; Earth radiation budget (including solar irradiance); upper-air temperature; wind speed and direction; water vapour; cloud properties; $CO_2$; ozone; aerosol properties |
| Oceanic | Sea-surface temperature; sea level; sea ice; ocean colour (for biological activity); sea state; ocean salinity |
| Terrestrial | Lakes; snow cover; glaciers and ice caps; albedo; land cover (including vegetation type); fraction of photosynthetically active radiation (faPAR); leaf area index (LAI); biomass; fire disturbance; soil moisture |

and local scales; and (vi) enable characterization of extreme events important in impact assessment and adaptation and the assessment of risk and vulnerability.

The GIP is intended to describe a feasible and cost-effective path toward an integrated observing system that depends on both in situ and satellite measurements. The ECVs largely dependent on satellite observations, and identified by the GIP, are given in Table 1 above – see also the recent GCOS report (GCOS-129 2009), which discusses ECVs and the notion of *Fundamental Climate Data Records* (FCDRs).

The GIP recognizes that addressing a number of GCOS issues requires research satellites:

- To provide intermittent, supplemental detail to sustained observations through (often challenging) new measurements;
- To seek improved and more effective ways of fully meeting observation targets and creating the required satellite records;
- To develop new observational capabilities to cover some of the ECVs for which a data record cannot at present be initiated due to an absence of proven capability.

Finally, the GIP indicates that Parties that support the space agencies should: (a) ensure continuity and overlap of key satellite sensors; (b) record and archive all satellite *metadata* (i.e., information on the satellite data); (c) maintain currently adopted data formats for all archived data; (d) provide data service systems that ensure accessibility; and (e) reprocess all data relevant to climate for inclusion in integrated climate analyses and reanalyses.

## *3.5 Capacity Report for Satellite Missions*

As part of the CAPACITY (Composition of the Atmosphere: Progress to Applications in the user CommunITY) study (http://www.knmi.nl/capacity), a report was written on existing and planned satellite missions (Kerridge et al. 2005). This report considered capabilities and limitations. We provide below a summary of this report – acronyms are identified in Appendix. The ESA CAMELOT

(Composition of the Atmospheric Mission concEpts and sentineL Observation Techniques) study (Levelt et al. 2009) is the follow-on study to the CAPACITY study.

### 3.5.1 Capabilities

*Cloud and aerosol*: Imagers on the operational satellites MSG and MetOp/NPOESS, and on the research satellites ERS-2, Envisat, EOS Terra and EOS Aqua, PARASOL, EarthCARE and the Sentinel 3 satellite, will provide geographical coverage on tropospheric clouds and aerosol, together with other physical properties (e.g. optical depth, size parameter, phase, and liquid water content). Radar and lidar instruments on Cloudsat, CALIPSO and EarthCARE will provide vertical profile information on clouds and aerosol along the sub-satellite track, although the design lifetimes of such active instruments are relatively short (~3 years).

Ice water content is a significant meteorological variable but will not be determined with sufficient accuracy by passive imagers. Visible and infrared wavelengths are insensitive to the size distribution of particles in the cirrus range. Extinction efficiencies of these size components typically peak in the sub-mm or THz regions, which are not measured by planned missions. Nadir-viewing imagers and spectrometers offer little if any information on either stratospheric aerosols or polar stratospheric clouds (PSCs).

*Water Vapour*: Water vapour soundings adequate for NWP will be performed in cloud-free scenes by MetOp/NPOESS. The operational system will not provide useful water vapour data above the tropopause, and vertical resolution in the upper troposphere will not be sufficient for future research applications.

*Ozone*: MetOp/NPOESS (GOME-2/OMPS) should provide adequate observations to monitor stratospheric ozone and total column ozone. Tropospheric ozone retrievals have been demonstrated for GOME (aboard the ERS-2 platform) and simulations indicate that nadir-FTIR observations from IASI/CrIS may add significant value to height-resolved ozone information from GOME-2/OMPS in the troposphere. Ozone observations by MetOp/NPOESS will not have sufficient vertical resolution in the UTLS for future research applications.

A ground pixel size smaller than that of GOME-2 or OMPS to allow more frequent sounding of the lower troposphere between clouds would be desirable for future research applications and air quality forecasting. The operational system will provide UV/Visible observations at only two local times (9:30 am, GOME-2; 1:30 pm, OMPS). Ozone observations at additional local times might be desirable for air quality forecasting.

*Trace gases other than ozone*: MetOp and NPOESS UV/Visible sensors should provide slant columns of several tropospheric trace gases in addition to ozone: $NO_2$, $SO_2$, $H_2CO$ (formaldehyde) and BrO. Typically, nadir observations contain no height-resolved information. Limb observations of the stratosphere made simultaneously by OMPS will allow slant-column information from nadir observations to be assigned to the troposphere. For GOME-2, a chemistry-transport model, CTM (with or without assimilation of OMPS limb data) will be needed to represent stratospheric

distributions of these trace gases and enable assignment of slant-column information to the troposphere. A ground pixel size smaller than that of GOME-2 or OMPS, to allow more frequent sounding of the lower troposphere between clouds, would be desirable for future research applications and air quality forecasting. As for ozone, the MetOP/NPOESS system will provide UV/Visible observations at only two local times (9:30 am, GOME-2; 1:30 pm, OMPS). For air quality forecasting, observations at additional local times would be desirable for trace gas pollutants with short photochemical lifetimes. Similar considerations hold for volcanic emissions of $SO_2$.

MetOp/NPOESS FTIR sensors will observe several trace gases in addition to water vapour and ozone, e.g., methane ($CH_4$) and CO. Height-assignment and height-resolution of these types of constituent observations is intrinsically limited, so they will best be exploited through data assimilation (see chapter *Constituent Assimilation*, Lahoz and Errera). For trace gases other than water vapour, sensitivity of the FTIR technique is lowest in the boundary layer, where temperature contrast with the surface is lowest. Because the MetOp/NPOESS system will have FTIR sensors operating concurrently in at least two different orbits, such observations will be made at four local times per day (Equator crossing times: 1:30 am, 9:30 am, 1:30 pm and 9:30 pm). Given the comparatively long photochemical lifetimes of methane and CO, this temporal sampling should be sufficient for most applications.

FTIR spectrometers on MetOp, NPOESS and GOSAT will also observe $CO_2$. Because $CO_2$ is close to being a uniformly mixed gas in the troposphere, extremely stringent observational requirements would need to be imposed to quantify perturbations in $CO_2$ mixing ratio at the amplitudes and spatial and temporal scales required for future research applications.

For future research on biogenic emission and uptake of trace gases such as $CO_2$, methane and $N_2O$, there will be a demand for remote-sensing measurements on a very fine spatial scale (tens of metres). This is not attainable from satellites but might be attainable from aircraft or balloons.

### 3.5.2 Limitations

A number of limitations of the currently planned suite of missions were identified:

- The absence of UV/Visible and infrared solar occultation missions for monitoring of stratospheric trace gas and aerosol profiles beyond MAESTRO and ACE on SCISAT-1, which are unlikely to be functioning beyond 2010 (see also Sect. 3.2);
- Requirements for sounding tropospheric trace gases will be addressed by MetOp/NPOESS (see Sect. 3.5.1). To comply better with quantitative requirements, the following would be desirable:
  - Nadir FTIR: spectral resolution similar to TES, i.e., higher than that of CrIS or IASI, to target additional tropospheric trace gases (e.g. non-methane hydrocarbons);
  - Nadir UV/Visible: observations later in the day than GOME-2 (Equator crossing time 9:30 am) and OMPS (Equator crossing time 1:30 pm) for early

morning air quality forecast and for detection of afternoon pollution episodes; ground-pixel size smaller than OMPS (50 × 50 km) to observe the boundary layer more frequently in between clouds; spectral coverage and resolution comparable to GOME-2 (to achieve photometric precision on, e.g., $NO_2$).

- MetOP/NPOESS will address requirements for sounding tropospheric aerosol. To comply better with quantitative requirements, height-resolution would also be desirable, for which the spectral coverage of GOME-2 and OMPS does not extend far enough into the near-infrared. This will be provided by the CALIPSO, ADM-Aeolus and EarthCARE lidars, although only along sub-satellite tracks and only for limited time periods (dictated by laser lifetimes and low orbit heights);
- MetOp/NPOESS will not address requirements for sounding UTLS trace gases and aerosol, with the exception of stratospheric ozone (GOME-2 and OMPS) and aerosol (OMPS). The ODIN, Envisat and EOS Aura limb sounders are currently addressing these requirements, but none of these are likely to be functioning beyond 2010 (see also Sect. 3.2).

### 3.5.3 Conclusions

It was found that the suitability of existing instrument technology depends on a number of factors including: (i) theme and application to be addressed; (ii) scope of the satellite mission, including restriction on number of platforms, orbits, number and types of sensors and systems; and (iii) importance and priority of particular observations, i.e., what is the effect of not achieving particular observational requirements. The CAPACITY study showed that, while many measurements are made and applications are addressed to various extents, there is scope for improving current techniques and bringing new types of sensor and observations to the available complement of instruments.

In order to define future satellite missions, the potential performance of integrated observing systems, which include satellite and in situ measurements, and a number of analysis tools, such as specialized retrieval schemes and data assimilation systems, must be assessed. An example of such a tool is OSSEs (see chapter *Observation System Simulation Experiments*, Masutani et al.). The relative timescale of the planned future missions from ESA, NASA and elsewhere is also important so that *complementarity* between missions can be assured and relevant *synergies* exploited.

## 4 Data Assimilation of Research Satellites

In the 1990s, following years of development of meteorological data assimilation by the NWP community, the data assimilation methodology (e.g. Kalnay 2003) began to be applied to constituents (including aerosol), with a strong focus on stratospheric

ozone (Rood 2003, 2005). Research satellites have measured most of the constituent data assimilated.

The assimilation of stratospheric constituents from research satellites is discussed in chapter *Constituent Assimilation* (Lahoz and Errera) – see also Lahoz et al. (2007a). The assimilation of tropospheric constituents from research satellites is discussed in chapters *Constituent Assimilation* (Lahoz and Errera) and *Inverse Modelling and Combined State-Source Estimation for Chemical Weather* (Elbern et al.); the assimilation of data from research satellites into operational systems is discussed in chapters *The Global Observing System* (Thépaut and Andersson) and *Assimilation of Operational Data* (Andersson and Thépaut). In this section we provide an introduction to constituent data assimilation, with a focus on the stratosphere.

Because of its comparatively later application, constituent data assimilation is less mature than meteorological data assimilation (i.e., NWP). Nevertheless, there has been substantial progress over the last 15 years, with the field evolving from initial efforts to test the methodology to later efforts focusing on products for monitoring ozone and other constituents. More recently, the production of ozone forecasts by a number of operational centres (e.g. ECMWF, Dethof 2003) has become routine. A notable feature of the application of the data assimilation methodology to stratospheric constituents has been the strong interaction between the NWP and research communities, for example, in the EU-funded ASSET project (Lahoz et al. 2007b).

The main aims for assimilating ozone in the stratosphere include the development of ozone and UV-forecasting capabilities; the need to monitor stratospheric ozone to track the evolution of the stratospheric composition, mainly ozone and the gases that destroy it (WMO 2006), and assess compliance with the Montreal protocol; and the need to evaluate the performance of instruments measuring ozone, especially those providing long-term datasets (e.g. TOMS, GOME). The assimilation of ozone is also important for technical reasons, including: the constraints ozone observations provide on other constituents; the use of assimilation techniques to evaluate models and ozone observations; the development of computer code to assimilate instrument radiances sensitive to temperature and constituents; and the dynamical information provided by ozone tracer distributions. Other stratospheric constituents besides ozone that are of interest in this regard include $H_2O$, $N_2O$, $CH_4$, $NO_2$, $HNO_3$, ClO, BrO and aerosol (see IGACO 2004 for a more complete list).

In NWP, the main motivation for stratospheric constituent assimilation has been the use of constituent information (in particular, water vapour and stratospheric ozone) to improve the weather forecast. Historically, two approaches have been used for stratospheric constituent data assimilation. One has done assimilation as part of an NWP system, used for operational weather forecasting; the other has done assimilation in a standalone chemical model, either a CTM or a photochemical box model, often with a more sophisticated representation of chemical processes. Whereas the aim of the NWP approach has been to improve weather forecasts, the aims of the

**Table 2** Selected assimilated stratospheric chemistry research satellite observations, 1978-present

| Satellite/instrument | Availability | Constituents |
|---|---|---|
| TOMS (several satellites) (McPeters et al. 1998) | 1978-present | Total column ozone |
| SBUV/2 (several satellites) (Miller et al. 2002) | 1978-present | Ozone layers |
| HIRS channel 9 (several satellites) (Joiner et al. 1998) | 1978-present | Radiances sensitive to ozone |
| LIMS (Gille and Russell 1984) | 1978–1979 | Ozone, $H_2O$, $HNO_3$ and $NO_2$ profiles |
| UARS CLAES (Roche et al. 1993) | 1991–1993 | $CH_4$, $NO_2$ profiles |
| UARS MLS (Waters 1998) | 1991–1997 | Ozone profiles |
| UARS HALOE (Russell et al. 1993) | 1991–2005 | Ozone, $N_2O$, $CH_4$, $H_2O$, HCl profiles |
| ATMOS (four space shuttle missions) (Gunson et al. 1996) | April 1985; March 1992; April 1993; November 1994 | O3, NO, $NO_2$, $N_2O_5$, $HNO_3$, $HO_2NO_2$, HCN, $ClONO_2$, HCl, $H_2O$, CO, $CO_2$, $CH_4$, and $N_2O$ profiles |
| CRISTA (two space shuttle missions) (Offermann et al. 1999) | November 1994; August 1997 | Ozone, $CH_4$, $N_2O$, CFC-11, $HNO_3$, $ClONO_2$ and $N_2O_5$ profiles |
| ERS-2 GOME (Burrows et al. 1999) | 1995-present | Total column ozone and $NO_2$, ozone profiles |
| ODIN SMR (Murtagh et al. 2002) | 2001-present | Ozone and $N_2O$ profiles |
| Envisat MIPAS (Fischer et al. 2000) | 2002-present | Ozone, $H_2O$, $NO_2$, $HNO_3$, $N_2O$, and $CH_4$ profiles; radiances sensitive to humidity and ozone |
| Envisat SCIAMACHY (Bovensmann et al. 1999) | 2002-present | Total column ozone, ozone profiles |
| Envisat GOMOS (Bertaux et al. 2000) | 2002-present | Ozone, $NO_2$, $NO_3$ profiles |
| ADEOS ILAS-II (Nakajima et al. 2006) | 2002–2003 | Ozone profiles |
| EOS Aura MLS (Waters et al. 2006) | 2004-present | Ozone profiles |
| EOS Aura OMI (Levelt et al. 2006) | 2004-present | Total column ozone |

chemical model approach are broader, and include providing chemical forecasts and analyses of chemical constituents.

Table 2 above provides selected stratospheric constituent research satellite observations for the period 1978 to the present, that have been assimilated by

NWP-based or chemical model data assimilation systems. References describing the satellites/instruments are provided.

## 5 Future Prospects

There are a number of activities, all involving the use of data from research satellites that are likely to become important in the future. These include:

- The operational use of research satellite data by significant numbers of operational centres. Examples of data used include ozone (already assimilated operationally at ECMWF), stratospheric water vapour, $CO_2$ and aerosols – see chapters *The Global Observing System* (Thépaut and Andersson), *Assimilation of Operational Data* (Andersson and Thépaut);
- Allied with the opportunity to assimilate data from limb sounders, the assimilation of limb radiances by research and operational groups. A lot of work has been done on developing fast and accurate forward models and the interface between the forward model and the assimilation. Progress is more advanced in the case of IR radiances than in the case of UV/Visible radiances, mainly due to the increased importance of scattering effects for the latter two;
- Chemical forecasting and air quality studies, including tropospheric pollution forecasting, and estimation of sources and sinks of pollutants and greenhouse gases – see chapters *Constituent Assimilation* (Lahoz and Errera), *Inverse Modelling and Combined State-Source Estimation for Chemical Weather* (Elbern et al.);
- An Earth System approach to environmental and associated socio-economic issues. This approach would incorporate the biosphere and the carbon cycle, and the coupling of all components of the Earth System. An example of activities bringing together the various components of the Earth System is the EU-funded GEMS project (Hollingsworth 2005; Hollingsworth et al. 2008).

## References

Bertaux, J.-L., E. Kyrölä and T. Wehr, 2000. Stellar occultation technique for atmospheric ozone monitoring: GOMOS on Envisat. *Earth. Observ. Q.*, **67**, 17–20.

Bovensmann, H., J.P. Burrows, M. Buchwitz, et al., 1999. SCIAMACHY: Mission objectives and measurement modes. *J. Atmos. Sci.*, **56**, 127–150.

Burrows, J.P., M. Weber, M. Buchwitz, et al., 1999. The Global Ozone Monitoring Experiment (GOME): Mission concept and first scientific results. *J. Atmos. Sci.*, **56**, 151–175.

Dethof, A., 2003. Assimilation of ozone retrievals from the MIPAS instrument onboard ENVISAT. *ECMWF Tech. Memo.*, **428**.

European Space Agency, ESA, 2005. Mission Requirements Document for the European Radar Observatory Sentinel-1, E. Attema, ES-RS-ESA-SY-0007.

European Space Agency, ESA, 2007a. GMES Sentinel-2 Mission Requirements Document, ESA Sentinel-2 Team, EOP-SM/1163/MR-dr.

European Space Agency, ESA, 2007b. Sentinel-3 Mission Requirements Document, M.R. Drinkwater and H. Rebhan, EOP-SMO/1151/MD-md.

European Space Agency, ESA, 2007c. GMES Sentinels 4 and 5 Mission Requirements Document, EOP-SMA/1507.

Fischer, H., C. Blom, H. Oelhaf, et al., 2000. Envisat MIPAS – An instrument for atmospheric chemistry and climate research. Readings, C and R.A. Harris (eds.), ESA Publication SP-1229, The Netherlands.

GCOS-92, 2004. Implementation Plan for the Global Observing System for Climate in support of the UNFCCC. GCOS-92, WMO/TD No. 1219, October 2004.

GCOS-107, 2006. Systematic observation requirements for satellite-based products for climate. Supplemental details to the satellite-based component of the "Implementation Plan for the Global Observing System for Climate in support of the UNFCCC". GCOS-107, WMO/TD No. 1338, September 2006.

GCOS-129, 2009. Progress report on the implementation of the Global Observing System for Climate in support of the UNFCCC 2004–2008. GCOS-129, WMO/TD No. 1489, GOOS-173, GTOS-70, August 2009.

Gille, J.C. and J.M. Russell, 1984. The limb infrared monitor of the stratosphere: Experiment description, performance, and results. *J. Geophys. Res.*, **89**, 5125–5140.

Gunson, M.R., M.M. Abbas, M.C. Abrams, et al., 1996. The Atmospheric Trace Molecule Spectroscopy (ATMOS) experiment: Deployment on the ATLAS Space Shuttle missions. *Geophys. Res. Lett.*, **23**, 2333–2336.

Hollingsworth, A., 2005. Global Earth-system Modelling using Space and in situ data. *ECMWF Seminar Proceedings,* September 2005, Reading. Available from http://www.ecmwf.int.

Hollinsgworth, A., R.J. Engelen, C. Textor, et al., 2008. The Global Earth-system Monitoring using Satellite and in-situ data (GEMS) Project: Towards a monitoring and forecasting system for atmospheric composition. *Bull. Amer. Meteorol. Soc.*, doi:10.1175/2008BAMS2355.1.

IGACO, 2004. The Changing Atmosphere. An Integrated Global Atmospheric Chemistry Observation theme for the IGOS partnership. ESA SP-1282, Report GAW No. 159 (WMO TD No. 1235), September 2004; Implementation up-date, December 2004. Available from: http://www.igospartners.org/docsTHEM.htm.

Joiner, J., H.-T. Lee, L.L. Strow, et al., 1998. Radiative transfer in the 9.6 μm HIRS ozone channel using collocated SBUV-determined ozone abundances. *J. Geophys. Res.*, **103**, 19213–19230.

Kalnay, E., 2003. *Atmospheric Modeling, Data Assimilation and Predictability*, Cambridge University Press, Cambridge, 364pp.

Kerridge, B.J., W.J. Reburn, V.L. Jay, et al., 2005. ESA Capacity Study. Report for WP2200, May 2005. Available from http://www.knmi.nl/capacity/FinalDocs/.

Lahoz, W.A., 2003. Research Satellites. In *Data Assimilation for the Earth System*. NATO Science Series: IV. Earth and Environmental Sciences 26, Swinbank, R., V. Shutyaev and W.A. Lahoz (eds.), Kluwer Academic Publishers, Dordrecht, The Netherlands, pp 241–250, 378pp.

Lahoz, W.A., R. Brugge, D.R. Jackson, S. Migliorini, R. Swinbank, D. Lary and A. Lee, 2005. An Observing System Simulation Experiment to evaluate the scientific merit of wind and ozone measurements from the future SWIFT instrument. *Q. J. R. Meteorol. Soc.*, **131**, 503–523.

Lahoz, W.A., Q. Errera, R. Swinbank and D. Fonteyn, 2007a. Data assimilation of stratospheric constituents: A review. *Atmos. Chem. Phys.*, **7**, 5745–5773.

Lahoz, W.A., A.J. Geer, S. Bekki, N. Bormann, S. Ceccherini, Q. Errera, H.J. Eskes, D. Fonteyn, D.R. Jackson, B. Khattatov, S. Massart, V.-H. Peuch, S. Rharmili, M. Ridolfi, A. Segers, O. Talagrand, H. Thornton, A.F. Vik and T. von Clarmann, 2007b. The Assimilation of Envisat data (ASSET) project. *Atmos. Chem. Phys.*, **7**, 1773–1796.

Lamarque, J.-F., B.V. Khattatov and J.C. Gille, 2002. Constraining tropospheric ozone column through data assimilation. *J. Geophys. Res.*, **107**, 10.1029/2001JD001249.

Levelt, P.F., G.H.J. van den Oord, M.R. Dobber, et al., 2006. The ozone monitoring instrument. *IEEE Trans. Geosci. Remote Sensing*, **44**, 1093–1101.

Levelt, P., P. Veefkind and the CAMELOT Team, 2009. ESA CAMELOT Study: Challenges in future operational missions for GMES atmospheric monitoring, sentinel 4 and 5. *Geophys. Res. Abs.*, **11**, EGU2009-8911.

McPeters, R.D., P.K. Bhartia, A. Krueger, et al., 1998. Earth Probe Total Ozone Mapping Spectrometer (TOMS) Data Products User's Guide. NASA Reference Publication 1998-206895.

Miller, A.J., R.M. Nagatani, L.E. Flynn, et al., 2002. A cohesive total ozone data set from the SBUV(/2) satellite system. *J. Geophys. Res.*, **107**, 10.1029/2001JD000853.

Murtagh, D., U. Fisk, F. Merino, et al., 2002. An overview of the Odin atmospheric mission. *Can. J. Phys.,* **80**, 309–319.

Nakajima, H., T. Sugita, T. Yokota, et al., 2006. Characteristics and performance of the Improved Limb Atmospheric Spectrometer-II (ILAS-II) on board the ADEOS-II satellite. *J. Geophys. Res.*, **111**, 10.1029/2005JD006334.

Offermann, D., K.-U. Grossman, P. Barthol, et al., 1999. The cryogenic infrared spectrometer and telescopes for the atmosphere (CRISTA) experiment and middle atmosphere variability. *J. Geophys. Res.*, **104**, 16311–16327.

Palmer, P.I. and P. Rayner, 2009. Launch failure. *Nat. Geosci.*, **2**, April 2009.

Roche, A.E., J.B. Kumer, J.L. Mergenthaler, et al., 1993. The Cryogenic Limb Array Etalon Spectrometer (CLAES) on UARS: Experiment description and performance. *J. Geophys. Res.*, **98**, 10763–10775.

Rodgers, C.D., 2000. *Inverse Methods for Atmospheric Sounding: Theory and Practice*. World Scientific, Singapore, 238pp.

Rood, R.B., 2003. Ozone assimilation. In *Data Assimilation for the Earth System.* NATO Science Series: IV. Earth and Environmental Sciences 26, Swinbank, R., V. Shutyaev and W.A. Lahoz (eds.), Kluwer Academic Publishers, Dordrecht, The Netherlands, pp 263–277, 378pp.

Rood, R.B., 2005. Assimilation of stratospheric meteorological and constituent observations: A Review. *SPARC Newsletter*, **25**, July 2005.

Russell III, J.M., L.L. Gordley, J.H. Park, et al., 1993. The Halogen occultation experiment. *J. Geophys. Res.*, **98**, 10777–19797.

Schoeberl, M.R., A.R. Douglass and J. Joiner, 2008. Introduction to special section on Aura Validation. *J. Geophys. Res.*, **113**, 10.1029/2007JD009602.

Struthers, H., R. Brugge, W.A. Lahoz, A. O'Neill and R. Swinbank, 2002. Assimilation of ozone profiles and total column measurements into a global general circulation model. *J. Geophys. Res.*, **107**, 10.1029/2001JD000957.

Waters, J.W., 1998. Atmospheric measurements by the MLS experiments: Results from UARS and plans for the future. *Adv. Space Res.*, **21**, 1363–1372.

Waters, J.W., L. Froidevaux, R.S. Harwood, et al., 2006. The Earth Observing System Microwave Limb Sounder (EOS MLS) on the Aura satellite. *IEEE Trans. Geosci. Remote Sensing*, **44**, 1075–1092.

WMO: Scientific Assessment of Ozone Depletion, 2006. World Meteorological Organization, Global Ozone Research and Monitoring Project, Report No. 50.

# Part III
# Meteorology and Atmospheric Dynamics

# General Concepts in Meteorology and Dynamics

**Andrew Charlton-Perez, William Lahoz, and Richard Swinbank**

## 1 Introduction

The aim of this chapter is to give a general overview of the atmospheric circulation, highlighting the main concepts that are important for a basic understanding of meteorology and atmospheric dynamics relevant to atmospheric data assimilation.

## 2 The Atmospheric Circulation

### *2.1 General Details*

The main driver of the atmospheric circulation is differential heating from solar radiation. The tropics, where the sun is almost overhead in the middle of the day, are heated most strongly, resulting in high temperatures. On the other hand, at high latitudes the sun is much lower in the sky during the day, resulting in lower temperatures. The time-mean global atmospheric circulation acts to reduce the gradient in temperature between the Equator and Pole by the redistribution of warm and cold air masses. However, a key additional determinant of the dynamical processes and flows which are observed is the presence of a gradient in angular momentum due to the approximately spherical surface of the Earth. The latitudinal temperature differences are modulated by the seasonal cycle. In summer, the Earth is tilted toward the sun, leading to warmer temperatures, and the Summer Pole is in continuous daylight. By contrast, the mid latitudes receive less heat and light from the sun in winter, and the Winter Pole is in continuous darkness.

Looking at the Earth from outer space, the atmosphere forms a very thin blue layer. While the radius of the Earth is about 6,370 km, 90% of the mass of the atmosphere is no more than 16 km above the surface of the Earth. Thus, the horizontal length scales of the atmospheric circulation can be very much larger than the

A. Charlton-Perez (✉)
Department of Meteorology, University of Reading, Reading, UK
e-mail: a.j.charlton@reading.ac.uk

W. Lahoz et al. (eds.), *Data Assimilation*, DOI 10.1007/978-3-540-74703-1_13,

height scales. Many of the most important features of the atmospheric circulation occur on global scales, but are confined to a few km in the vertical. Meteorologists also often use pressure, instead of height, as a vertical coordinate, since it decreases monotonically with height and allows the equations of motion for the atmosphere to be somewhat simplified. Pressure is measured in Pascals (Pa) and typically expressed as hectoPascals (hPa). The pressure at the Earth's surface is approximately 1,000 hPa and around 100 hPa at 16 km in the mid latitudes. The e-folding length scale (or scale height) for pressure in the atmosphere is usually taken to be around 7 km.

Figure 1 is a zonal-mean latitude-pressure cross-section of the climatological thermal structure of the atmosphere for January and July, derived from long-term measurements of temperature from weather balloons and satellites. Starting at the

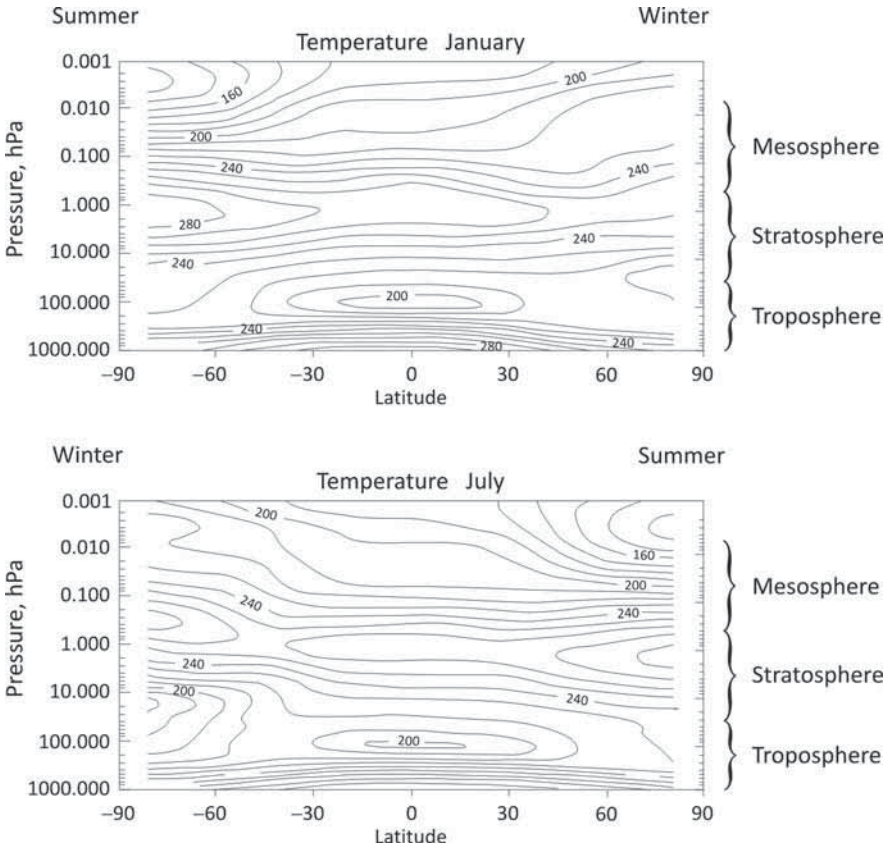



**Fig. 1** Monthly zonal mean climatology of temperature, K. The vertical coordinate is pressure (hPa). *Top*: January; *Bottom*: July. The winter and summer hemispheres are indicated. The troposphere, stratosphere and mesosphere are indicated. Based on COSPAR (COmmittee on SPAce Research) International Reference Atmosphere, CIRA, material. See also Fleming et al. (1988)

surface the temperature decreases rapidly until we reach the *tropopause*; the part of the atmosphere below the tropopause is referred to as the *troposphere* (or sometimes the *lower atmosphere*). The main reason for the decrease in temperature with height – the *lapse rate* – is the reduction in pressure with height; when an air parcel is displaced upwards, it expands and cools adiabatically.

Weather systems, such as storms, fronts and tropical cyclones, are all confined to the troposphere. A key part of the structure of these systems, and the troposphere in general, is motion of air parcels in the vertical. Vertical motion helps to redistribute atmospheric constituents throughout the troposphere, leading to reductions in the gradients of many constituents. Where the air is moist, vertical motion can also lead to condensation of water vapour contained in the air parcel, since the amount of water vapour which can exist in an air parcel is strongly dependent upon pressure (from the *Clausius-Clapeyron* relation). Condensation of water vapour has two effects: it forms clouds and heats the air parcel through latent heat release from the change of water phase. Latent heating can act to reduce the lapse rate that would otherwise be observed and is an important driver of atmospheric motion, particularly in the tropics. For a more detailed explanation of this, and other issues covered in this chapter, the reader is recommended to consult an atmospheric dynamics text book, such as Holton (2004).

There is an abrupt change in the temperature profile at the tropopause, the boundary between the *troposphere* and *stratosphere*. In the stratosphere, the temperature structure is determined by the balance between radiative heating and cooling processes caused by absorption and emission of long-wave radiation by carbon dioxide and absorption of solar radiation by ozone. This is in contrast to the troposphere, which, apart from the effect of clouds is almost transparent to incoming solar radiation and primarily heated by contact with the Earth's surface. The balance between heating and cooling processes changes with altitude, largely as a result of the increase in ozone concentration. This leads to an increase of temperature (a negative lapse rate) through the stratosphere. The resulting high static stability inhibits vertical motion. The altitude of the tropopause varies with latitude, from as low as about 10 km (approximately 300 hPa) at high latitudes to around 16 km (approximately 100 hPa) in the tropics. As a result of the high tropical tropopause, at around 16 km the atmosphere is coldest at the Equator and warmest at the Summer Pole.

At around 50 km (approximately 1 hPa), where the effect of the ozone heating fades away, we reach the *stratopause*. The stratopause is the boundary between the *stratosphere* and *mesosphere*. In the mesosphere, the lapse rate becomes positive once again. The stratosphere and mesosphere, taken together, are sometimes referred to as the *middle atmosphere*. The top of the mesosphere is at around 80 km (approximately 0.01 hPa), where the *mesopause* is located. For a comprehensive description of the dynamical meteorology of the middle atmosphere, the reader is referred to the excellent book by Andrews et al. (1987).

Above the mesosphere, the *thermosphere* is the outermost layer of the neutral atmosphere, and temperatures once more increase with altitude. At this level the atmosphere is so thin that molecules collide only rarely. Because molecular collisions are infrequent, the conventional definition of temperature based on the ideas

of Maxwell and Boltzmann is difficult to apply. One manifestation of this is the breakdown at these heights of the (commonly made) assumption of *local thermal equilibrium* (*LTE*). Furthermore, with the breakdown of turbulent mixing, the different atmospheric constituents start to settle out under the influence of gravity. The nature of the atmosphere thus becomes quite different from that found in the lower and middle atmosphere, and is beyond the scope of this book. For more information about the *upper atmosphere*, see, e.g., Rees (1989).

## 2.2 Influence of Rotation

The Earth's rotation has a major impact on the observed atmospheric flow. The acceleration of the westerly ($u$) and southerly ($v$) wind components (i.e., the zonal and meridional components of the momentum equations, respectively), is written as (Andrews et al. 1987):

$$\frac{Du}{Dt} - \left(f + \frac{u\tan\phi}{a}\right)v + \frac{\Phi_\lambda}{a\cos\phi} = F_x \tag{1}$$

$$\frac{Dv}{Dt} + \left(f + \frac{u\tan\phi}{a}\right)u + \frac{\Phi_\phi}{a} = F_y \tag{2}$$

where $D/Dt$ is the *total derivative*; $a$ is the Earth's radius; $\lambda$ and $\phi$ are longitude and latitude, respectively; $f = 2\Omega\sin\phi$ is the *Coriolis parameter* ($\Omega$ is the Earth's rotation rate, $= 7.29 \times 10^{-5}$ s$^{-1}$); $\Phi$ is the *geopotential* (involving integration of the acceleration due to gravity, $g$, from the ground to a height $z$); the subscripts denote differentiation with respect to the subscript; and $F_x$ and $F_y$ are unspecified horizontal components of friction, or other non-conservative mechanical forcing. For more detail see also the chapter *The Role of the Model in the Data Assimilation System* (Rood).

These equations are derived from Newton's second law of motion; they equate the acceleration of air parcels with the balance of forces acting on them. The first term on the left hand side of both Eqs. (1) and (2) is the acceleration and the other two terms represent the *Coriolis* force and the pressure gradient force. The term on the right hand side of Eqs. (1) and (2) represents the frictional force. The Coriolis force is a "virtual" force which is a consequence of the Earth's rotation.

If the Earth were not rotating, the Coriolis force would be zero and air would ascend in the tropics, where it is heated most strongly, move toward the poles, descend at higher latitudes as it cooled, and return equatorwards at low levels. However, since the actual atmospheric circulation is strongly influenced by the Earth's rotation the Coriolis force has an important part to play. At the Equator, the surface is moving eastwards at about 185 ms$^{-1}$ as a result of the Earth's rotation. Imagine an air parcel starting at the Equator and moving toward one of the poles. As it moves away from the Equator, it would also move closer to the Earth's axis of rotation. In order to conserve *angular momentum* about this axis (defined as moment of inertia about this axis $\times$ angular rotation), the parcel would start to move

eastwards even faster as the moment of inertia decreased. In our normal frame of reference, which rotates with the Earth, the polewards-moving parcel would appear to be being accelerated to the east (In standard meteorological convention, we would refer to this as westerly acceleration, since wind directions are determined by where the wind comes from.). The Coriolis force allows this acceleration to be represented in the momentum equations.

By comparing the typical sizes of various terms in the momentum equations, it is possible to derive several approximate balances which govern atmospheric flow on different spatial scales.

In the horizontal, an important balance exists between the pressure gradient and Coriolis forces. Making the approximation that the variation of the Coriolis effect in latitude is linear (allowing us to remove the terms dependent on tan $\phi$ in Eqs. 1 and 2), and using height, z, as the vertical coordinate and the Cartesian coordinates $x$ and $y$ as the horizontal coordinates, the acceleration of westerly ($u$) and southerly ($v$) wind components (see above) can be written in the following manner (Andrews et al. 1987) – compare with Eqs. (1) and (2):

$$\frac{Du}{Dt} - fv + \frac{1}{\rho}\frac{\partial p}{\partial x} = F_x, \tag{3}$$

$$\frac{Dv}{Dt} + fu + \frac{1}{\rho}\frac{\partial p}{\partial y} = F_y, \tag{4}$$

where $\rho$ is the air density. For large-scale atmospheric flow ($\sim$1,000 km), the two largest terms in Eqs. (3) and (4) are the Coriolis term and the pressure gradient term; both are typically an order of magnitude larger than the acceleration and frictional terms. Equations (3) and (4) can be simplified to a balance between the Coriolis and pressure gradient forces, an approximation known as *geostrophic balance*. From the geostrophic approximation a simple estimate of the flow can be made (the *geostrophic wind*) which depends simply on location and the local pressure gradient. In vector notation, the horizontal geostrophic wind $\mathbf{u}_g$ is given by (where $\mathbf{k}$ is a unit vertical vector):

$$\mathbf{u}_g = \frac{1}{f}\mathbf{k} \times \frac{1}{\rho}\nabla p. \tag{5}$$

The geostrophic wind blows anticlockwise around cyclones (areas of low pressure) in the Northern Hemisphere, and clockwise around anticyclones – rather than blowing directly from high pressure to low pressure, as one would expect in the absence of the Coriolis effect. In the Southern Hemisphere the circulation of the wind is in the opposite sense, as the sign of $f$ is reversed.

When typical sizes of the terms in the vertical component of the momentum equation are compared (see Eq. 6 below), the most important terms are the pressure gradient term and the gravity term. Thus, where $w$ is the vertical wind component and the Cartesian coordinate $z$ is height (see chapter *The Role of the Model in the Data Assimilation System*, Rood):

$$\frac{Dw}{Dt} - \frac{u^2 + \mathrm{v}^2}{a} = -\frac{1}{\rho}\frac{\partial p}{\partial z} - g + 2\Omega u\cos(\phi) + \nu\nabla^2(w) \rightarrow \tag{6a}$$

$$-\frac{1}{\rho}\frac{\partial p}{\partial z} = -\frac{RT}{p}\frac{\partial p}{\partial z} \approx g \tag{6b}$$

In this case, the weight of the parcel acts downwards, and this is balanced by the vertical difference in pressure (or buoyancy). When these two forces are equal and opposite, the air parcel is in *hydrostatic balance*; the atmosphere is usually very close to this condition (Eq. 6b).

A further approximate property of the large-scale flow can be derived by combining geostrophic and hydrostatic balance. At a constant altitude, the geostrophic wind is proportional to the pressure gradient – see Eq. (5). In the Northern Hemisphere westerly winds (a positive value of $u$) would be in balance with pressure that decreases to the north ($f$ is positive in the Northern Hemisphere). By contrast, in the Southern Hemisphere westerly winds would be in balance with pressure that decreases to the south ($f$ is negative in the Southern Hemisphere). Considering hydrostatic balance, the variation of pressure with height depends on the temperature (Eq. 6b). Normally, the temperature decreases to the north in the Northern Hemisphere, so the north-south pressure gradient will increase with altitude. As a result, the strength of the westerly (geostrophic) wind will also increase with height. A similar argument shows that when the temperature decreases to the south in the Southern Hemisphere the westerly wind will again increase with height. Algebraically, this relationship can be derived by taking the derivative of Eq. (5) with respect to height, then replacing the vertical and horizontal derivative of pressure with the horizontal derivative of temperature using hydrostatic balance. This procedure leads to Eq. (7), the *thermal wind* relationship:

$$\frac{\partial \mathbf{u}_g}{\partial z} = \frac{R}{Hf}\mathbf{k} \times \nabla T, \tag{7}$$

where $R$ is the gas constant, $H$ is the density height scale (cf. the e-folding length scale for pressure), $T$ is temperature and $\mathbf{k}$ is a unit vertical vector.

Figure 2 shows the zonal-mean latitude-pressure cross-section of westerly winds for January and July. The troposphere is dominated by westerly jets that reach their peak values close to the tropopause, the *tropospheric jets*. Comparing these cross-sections with Fig. 1, it is evident that the increase of westerlies with height is correlated with the latitudinal temperature gradient in exactly the way that one would expect from the thermal wind relationship – see Eq. (7); the tropospheric jet is much weaker in the Summer Hemisphere, where the temperature gradients are weaker. In the stratosphere, the westerlies continue to increase with height, again reflecting the cold polar temperatures, to form the *polar night jets*. In the summer stratosphere, the temperature gradients are reversed, and the westerly polar night jet is replaced by a summer easterly jet.

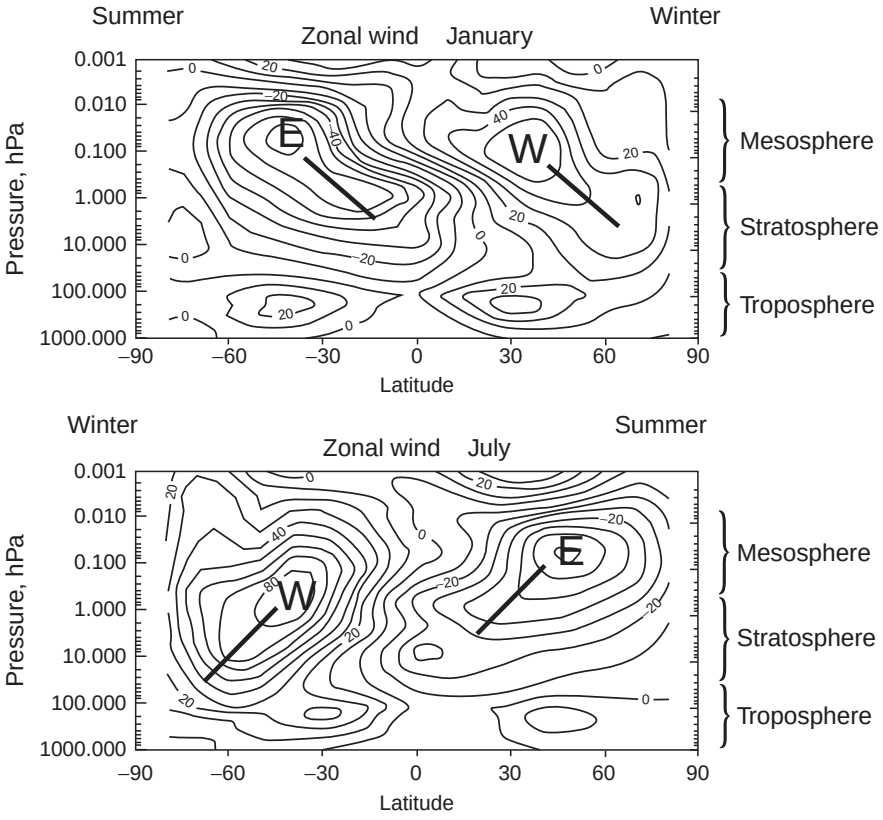



**Fig. 2** Monthly mean zonal mean climatology of zonal wind, $ms^{-1}$! *Top*: January; *Bottom*: July. The vertical coordinate is pressure (hPa). *Black lines* indicate the tilt of the jets. W indicates a westerly jet; E indicates an easterly jet. The troposphere, stratosphere and mesosphere are indicated. Based on CIRA material. See also Fleming et al. (1988)

As can be seen, the westerly and easterly stratospheric jets are significantly stronger (in magnitude) than the westerly tropospheric jets. Another feature to note is that the Southern Winter westerly jet is stronger than its Northern Winter counterpart. This can be ascribed to temperatures being generally colder at the South Pole than at the North Pole and, consequently, to a stronger temperature contrast between the South Pole and the Equator than between the North Pole and the Equator. Another feature to note is that the winter jet tilts upwards and equatorwards, whereas the summer jet tilts upwards and polewards.

The observed meridional, i.e., southerly wind is an order of magnitude smaller than the zonal, i.e., westerly, wind. The observed vertical wind is at least one further order of magnitude smaller still than the meridional winds. However, since the largest gradients of angular momentum, energy and moisture tend to occur in the meridional direction, understanding of the meridional and vertical circulation is key

to our understanding of the global circulation and climate. In the next section, we discuss the broad features of the meridional circulation and the forces which drive it.

## 3 General Circulation in the Troposphere

### *3.1 The Thermally-Driven Circulation in the Tropics*

In the tropics, the Coriolis effect is small so, to a first approximation, the large-scale atmospheric circulation is a direct response to thermal driving. The strongest ascent is where the solar heating is strongest. At the solstice seasons, there is a strong direct circulation, known as the *Hadley cell*, with ascent in the Summer Hemisphere and upper-level flow into the Winter Hemisphere, and a much weaker partner cell in the Winter Hemisphere. Air descends in the subtropics, forming a subtropical high-pressure belt. This circulation pattern is discussed in more detail by many other authors, e.g., Peixoto and Oort (1992). At the equinox seasons, there is a pair of Hadley cells of similar magnitude.

The Hadley cell is a *direct circulation*, i.e., it involves ascent over a warmer region and descent over a colder region; the Polar cell is also a direct circulation. In an *indirect circulation* (the *Ferrel cell* is an example), the reverse takes place, i.e., ascent over a colder region and descent over a warmer region (see Sect. 3.4).

The ascent in the Hadley circulation does not occur equally at all latitudes; there are differences between land and sea that result from the different heat capacities of the different surface types. The line along which the air from the two hemispheres converges and ascends is known as the *Inter-tropical Convergence Zone* (*ITCZ*). This is an area of intense rainfall, since the warm tropical air can hold a lot of water vapour, which condenses and rains out as the air ascends. By contrast, the regions where the descent occurs in the subtropics are characterized by low rainfall (for example, the Sahara and other deserts), since the predominant motion here is the descent of dry air from aloft.

A simple model of the Hadley cell which describes the main features of the circulation can be derived using only a few basic physical principles (Held and Hou 1980). The Held-Hou model assumes that angular momentum is conserved in the poleward flow aloft, that winds and temperatures are in thermal wind balance and that there is no net input or output of energy to the atmosphere integrated over the entire Hadley cell. With good agreement with the observations, the model predicts: (i) the existence of an eastward jet at the poleward edge of the Hadley cell; (ii) a very flat temperature profile in the deep tropics; (iii) the latitudinal extent of the Hadley cell; and (iv) the strength of the Hadley cell. Several additions can be made to the most simple formulation of the Held-Hou model, in order to understand other aspects of the structure of the Hadley circulation. The addition of a simple representation of moisture to the model increases the strength and decreases the width of the upwelling part of the flow and decreases the strength and increases the width of the downwelling part of the flow. The incorporation of off-equatorial heating

into the model, to mimic the seasonal cycle of solar heating, produces a circulation with a similar seasonal bias as seen in observations, between a broad intense cross-equatorial winter cell and a narrow, weak summer cell. For more details, see the excellent detailed text by Vallis (2007).

## *3.2 Angular Momentum Balance*

More insight into the structure of the tropospheric circulation can be gained by considering the angular momentum budget of the lower atmosphere. We have seen that, for the most part, surface winds in the tropics are equatorward and easterly. If we consider the effect on the solid Earth, the winds would exert a westward torque, tending to reduce the rate of rotation of the solid Earth, and at the same time the atmospheric flow would become more easterly. However, in the long term, both the atmosphere and the solid Earth are neither accelerating not decelerating. Thus, both easterly and westerly winds must exist over different parts of the globe.

The direct Hadley circulation transports westerly angular momentum into the subtropics. Atmospheric waves, in which there is a correlation between zonal and meridional wind components, are an efficient mechanism to transport westerly angular momentum into the mid latitudes to form mid latitude jet streams. In turn, the westerly angular momentum is transported downwards to drive the mid latitude westerlies. The transport of angular momentum from source regions in low latitudes to sink regions in mid latitudes maintains the observed atmospheric circulation. This provides an *acceleration* of the zonal mean wind at mid latitudes and a *deceleration* of the zonal mean wind at low latitudes. It can be shown that this transport of angular momentum poleward is effected by *eddies*, i.e., deviations from the zonal mean (*stationary eddies*) or the time mean (*transient eddies*).

Much of the exchange of angular momentum between atmosphere and solid Earth occurs as a result of turbulent mixing processes in the atmospheric boundary layer. These frictional processes occur on a scale that cannot be resolved by atmospheric models. The torque acting on the atmosphere as a result is generally referred to as *friction torque*. In addition, if we consider a mountain range, it will often be the case that the atmospheric pressure at a given height on one side of the mountain will be different from the corresponding pressure on the opposite side. This pressure difference is an additional mechanism for the transfer of angular momentum between the atmosphere and solid Earth, which is referred to as *mountain torque*. Both torques tend to act in the same sense.

As shown by, for example, Swinbank (1985), these torques cannot entirely explain the angular momentum balance of the atmosphere. This gap in the angular momentum budget indicates that *gravity wave drag* is also important, and the process needed to be included in atmospheric models. Gravity (buoyancy) waves occur at intermediate scales between the small-scale boundary layer turbulence and the large-scale mountain drag. The gravity waves are generated by flow over orography and other processes such as convective storms, and they carry momentum to upper levels, where the waves break and exert a drag on winds in the free atmosphere.

We now summarize the features of the meridional and vertical distributions of *angular momentum fluxes* (a flux involves transport of a quantity such as momentum):

- The annual mean angular momentum transport is almost symmetric about the Equator, and peaks at approximately 25°–30° for both hemispheres;
- Transport is dominated by transient eddies, e.g., mid latitude cyclones, large-scale travelling waves;
- The annual mean transport due to the *mean meridional circulation* (i.e., the atmospheric circulation in the meridional, north–south, plane) is small;
- In general, the total and transient angular momentum fluxes are largest in the Winter Hemisphere. This is consistent with the notion of a more vigorous and disturbed flow in winter;
- The most striking asymmetry occurs for the stationary eddies. The Northern Hemisphere stationary eddy flux is a significant fraction of the transient eddy flux. The stationary eddy flux is almost absent in Southern Hemisphere winter;
- The stationary eddy flux in the Northern Hemisphere is associated with large-scale stationary waves generated by flow over the Rocky Mountains and Tibet, and by land-sea contrasts. Similar features are not found in the Southern Hemisphere;
- There is a compensating asymmetry in the winter transient fluxes. The Southern Hemisphere has the largest values, so the sum of the transient and stationary values is almost identical for both hemispheres;
- Most of the transport is achieved in the upper troposphere, but angular momentum is added or removed from the atmosphere at the ground. This indicates that *vertical transport* is needed.

## *3.3 Rossby Waves and Mid Latitude Systems*

Before discussing the extra-tropical meridional circulation (Sect. 3.4), it is useful to outline the main features of *Rossby waves* and their relationship with weather systems in the mid latitudes. An excellent description of Rossby waves and their properties can be found in Holton (2004, Chap. 7); here we reproduce some of the more basic arguments.

The principles behind the existence and propagation of Rossby waves can be most easily understood by considering a barotropic fluid of constant depth on a rotating sphere. If the fluid is initially at rest, as we move across the sphere from Equator to Pole there will be a background gradient of planetary vorticity, simply equal to the variation of the Coriolis parameter with latitude. The propagation of a Rossby wave in this idealized fluid can be illustrated by considering a chain of fluid parcels at rest in the middle of the fluid (Fig. 3).

Now imagine that a fluid parcel is displaced poleward (upward in Fig. 3, top panel) from its resting position. By conservation of angular momentum, the planetary vorticity at the parcel's new location is larger than at its original location.

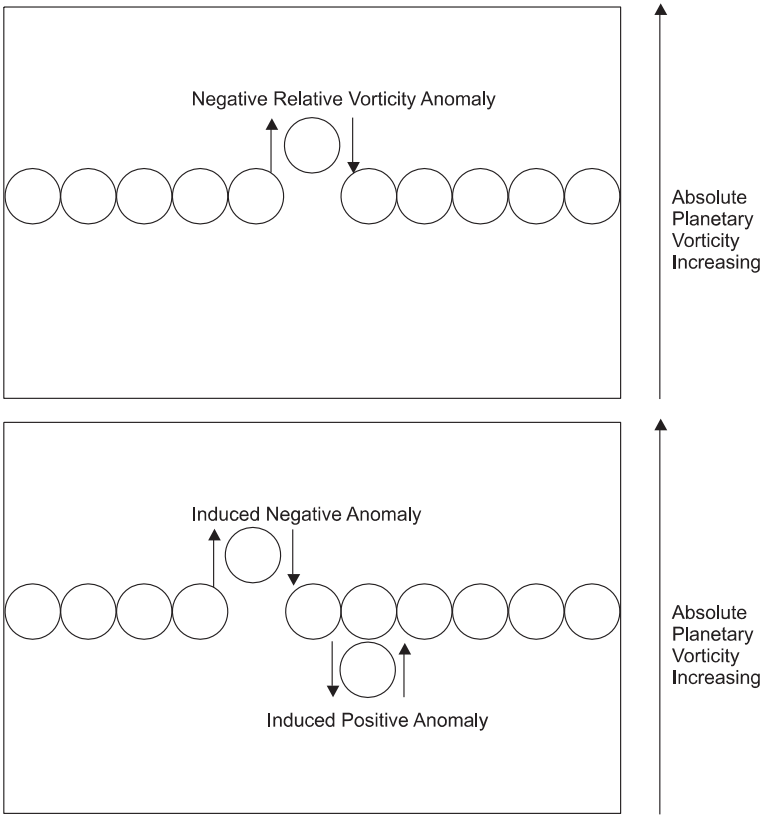



**Fig. 3** Schematic of the development of a negative relative vorticity anomaly and the associated meridional flow. *Top*, initial meridional displacement due to the anomaly; *bottom*, subsequent development of vorticity anomalies and westward (leftward in panel) propagation of the fluid displacements. See text for details

In order for the parcel to conserve its absolute vorticity (the sum of planetary and relative vorticity) its relative vorticity must be reduced, the parcel becomes a local negative or anticyclonic relative vorticity anomaly. The presence of the local relative vorticity anomaly generates meridional flow as indicated by the small arrows in Fig. 3 (top panel) causing northward (upward in Fig. 3, top panel) displacement of the fluid parcel to the west (leftward in Fig. 3, top panel) of the displaced parcel and southward displacement of the fluid parcel to the east. The advection of the fluid parcels north and south of their original position causes the generation of similar local relative vorticity anomalies, causes further meridional displacement of fluid parcels and eventually leads to westward propagation of the fluid displacements (Fig. 3, bottom panel) – the essence of the Rossby wave.

In a more complex baroclinic fluid like the atmosphere, Rossby wave propagation comes about through similar processes which involve conservation of potential vorticity on isentropic surfaces. In the atmosphere, Rossby waves can be excited either by the meridional displacement of parcels as illustrated in Fig. 3, or by local sources of relative vorticity, one example being through vortex stretching caused by latent heating in the tropics from anomalous moist convection. Particularly important in the troposphere and stratosphere are Rossby waves which have fixed phase relative to the ground (for this reason the waves are often called stationary waves). These are typically waves with small wavenumber/large wavelength and can be understood in the simplest terms as occurring due to a balance between westward phase propagation of the wave and eastward phase advection by the typically eastward mean flow.

Rossby waves also have an important part to play in *baroclinic instability*, the process which is thought to be responsible for the synoptic scale weather systems seen in mid latitudes. Again, an excellent in depth treatment of this topic can be found in Holton (2004, Chap. 8). Baroclinic instability requires the presence of a background vertical shear or gradient in the horizontal wind. A very simple description of baroclinic instability can be made by considering a model troposphere with constant shear and static stability, two, rigid, flat bounding surfaces at the ground and the tropopause and a purely zonal basic state flow. At the ground, the potential vorticity gradient is negative with largest potential vorticity values at the Equator. At the tropopause, the potential vorticity gradient is positive with largest potential vorticity at the poles. One can imagine Rossby waves which occur on the two rigid surfaces at the ground and the tropopause. These waves have phase propagation in opposite directions due to the change in sign of the background potential vorticity gradient. A key part of the instability, that the amplitude of waves grows spontaneously once initiated, is that the waves at the ground and tropopause can interact. This interaction takes the form of a weak induced velocity from one wave on the structure of the other. If the phase of the waves on the two surfaces is displaced by $^1/_4$ of a wavelength, optimal conditions for growth occur, where the induced meridional flow from each wave tends to enhance the amplitude of the individual peaks and troughs of the other wave. The presence of a vertically sheared horizontal flow allows the two Rossby waves to maintain their $^1/_4$ wavelength separation despite their opposite phase velocities, since the strong winds at tropopause level tend to advect phase features toward the east. This counter-propagating Rossby wave theory of baroclinic instability is described in rigorous mathematical detail in an excellent series of papers by Heifetz et al. (2004a, b) and Methven et al. (2005a, b).

Further analysis of baroclinically unstable structures shows that their phenomenology corresponds strongly to the structure of mid latitude weather systems. However, calculations of so called “baroclinic lifecycles” show that perturbations to a baroclinically unstable state rapidly evolve from their linear growth phase to a highly distorted non-linear breaking phase where the structure of the baroclinic wave rapidly breaks down. This breakdown often occurs at the end of the mid latitude storm tracks giving us the variety of typical synoptic systems often encountered in the mid latitudes.

## *3.4 The Extra-Tropical Meridional Circulation*

In the extra-tropics the diagnosis and understanding of the meridional circulation is more complicated than that in the tropics, because of the presence of Rossby waves. Rossby waves can cause significant meridional fluxes of mass and other quantities and have a significant role to play in the meridional circulation.

We begin by discussing the meridional circulation derived from standard diagnostics of the flow. Figure 4 shows the meridional circulation of the atmosphere in an *Eulerian frame*, where an observer is fixed to the rotating surface of the Earth and watches the flow relative to their location. Using these diagnostics, between around 35° and around 65° latitude in both hemispheres there is an indirect circulation, known as a Ferrel cell, i.e., ascent occurs at higher latitudes and descent at lower latitudes. This is exactly the opposite from what one would expect: apparently, ascent occurs in cold air and descent in warm air. However, this picture is misleading, because it ignores mass fluxes due to eddy motions, focusing instead on the zonal (i.e., longitudinal) average part of the flow. In fact, much of the ascent at mid latitudes occurs in the warm sector of mid latitude cyclones, which is generally warmer than its surroundings.

Finally, it should be pointed out that Fig. 4 also shows a small, direct, circulation at the poles, with ascent at lower latitudes and descent near the poles. These Polar cells act to warm the atmosphere at the poles and balance radiative cooling. Associated with this cell, in the lower troposphere, air moves from high latitudes to mid latitudes in a generally easterly sense.

To further understand the Ferrel cell we consider the angular momentum budget in its upper branch. As discussed in Sect. 3.2, angular momentum is transported to mid latitudes by eddies, leading to a convergence of atmospheric angular momentum in this region. This flux convergence is balanced by the Coriolis term, which removes westerly angular momentum aloft, and provides a low-level source of westerly angular momentum that drives the surface westerlies. The Ferrel cell reflects the effect of the atmospheric circulation in transferring angular momentum vertically at mid latitudes.

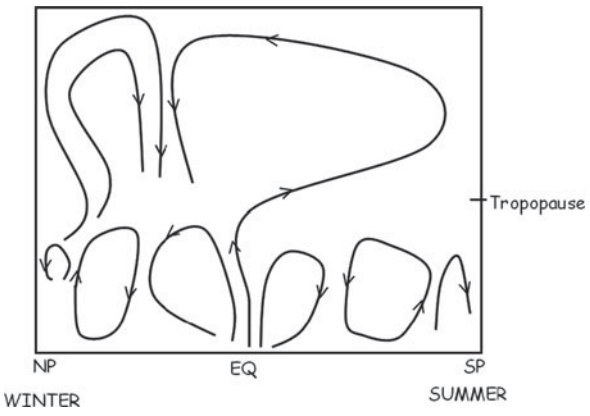



**Fig. 4** Eulerian picture of the atmospheric circulation. NP and SP stand for North Pole and South Pole, respectively. Northern Winter conditions are assumed. See text for details

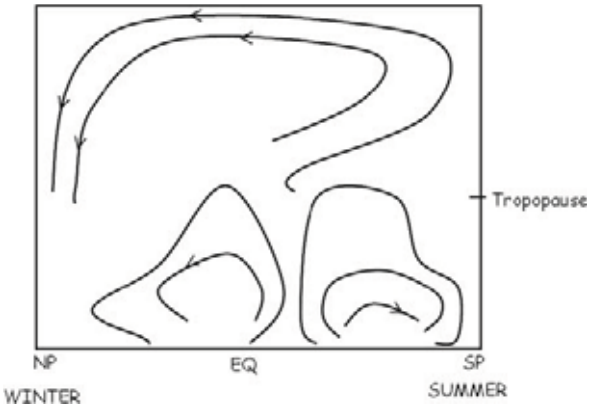



**Fig. 5** Lagrangian picture of the atmospheric circulation. NP and SP stand for North Pole and South Pole, respectively. Northern Winter conditions are assumed. See text for details

The meridional circulation may also be considered in a *Lagrangian framework* (in which an air parcel is followed), analogous to the transport properties of the circulation (Fig. 5). In general, this shows rising motion over the Summer Pole and descending motion over the Winter Pole. The Lagrangian average does not show an indirect circulation, since it takes into account eddy properties. The *Transformed Eulerian Mean* (*TEM*) framework, introduced by Andrews and McIntyre (1976) takes into account the effect of eddies in an Eulerian framework. Both the Lagrangian and TEM frameworks show that the tropospheric meridional circulation for constituents is a directly, thermally driven circulation. Throughout much of the troposphere, there is a general poleward motion, deeper in the tropics than in the mid latitudes. The circulation is closed by a more intense and shallow equatorward circulation confined to a thin layer close to the ground. These mean pictures are only revealed when many years of observational data are considered and averaged, in order to remove day-to-day fluctuations in the flow. For individual assimilation cycles, one should not expect to see this mean picture emerge in all cases. The middle atmosphere also has an important meridional circulation which is discussed at greater length in Sect. 4.

## *3.5 Other Tropical Circulations*

In both the tropics and extra-tropics, there are important features of the flow which result from asymmetries in the longitudinal direction. In this section we describe two well known features of the tropical flow, the monsoon and Walker circulations.

*Monsoon circulations.* The distribution of land and ocean can alter the general pattern of the atmospheric circulation. Over the ocean, the effect of solar heating is distributed over the top few metres of sea water, so the oceans exhibit a much weaker diurnal temperature cycle than land. A major consequence of the land-ocean heating contrast is the existence of tropical monsoon circulations. Monsoon circulations lead

to major seasonal variations in the low-level flow, which is generally from land to ocean in the winter season and from ocean to land in the summer season, after the onset of the monsoon circulation. According to the most widely used definitions, three major monsoon circulations occur around the tropics, in West Africa, Australia and in southern and south-eastern Asia (the Asian summer monsoon). Of these, the most well-known is the Asian summer monsoon. In Northern Summer strong heating occurs over the Indian subcontinent, enhanced by the Himalayas. The strong heating leads to convection, and a thermally driven convective cell results, with flow into the convecting region at low level and out of the convecting region at upper levels (see, e.g., http://www.wrh.noaa.gov/twc/monsoon/monsoon_what_is.php). Since the monsoon circulations occur at low latitudes, the effects of rotation are small and geostrophic balance does not hold, the dominant driver of motion being horizontal pressure gradients.

The onset of the Asian summer monsoon is around mid June, with variations of 2 weeks around this date. The monsoon is not steady, and has active and break periods. The year-to-year variations in onset and the number of break periods of the monsoon can significantly affect the total rainfall. This has an important consequence for populations dependent on this rainfall. However, despite this variability, the Asian summer monsoon circulation is remarkable in its regularity over many decades. The strong interactions between ocean and atmosphere in the tropics are currently thought to help to regulate the monsoon circulation.

*The Walker circulation*. In the longitudinal direction preferred regions for deep convection exist where temperatures are high and there is a ready supply of moisture. Strong large-scale ascent occurs over Africa, South America and Indonesia. Convective rainfall is particularly strong over Indonesia and the other islands of South East Asia because of the ready supply of water from the surrounding seas; this region is sometimes referred to as the *maritime continent*. Between the areas of deep convection and ascent, there may be areas of descent. As in the case of monsoons, this asymmetry in deep convection leads to a thermally driven circulation, in this case along the Equator. The major east–west circulation in the Pacific region – with ascent over the maritime continent and descent over the eastern Pacific is known as the *Walker circulation*. In its normal phase, the Walker circulation has ascent over the maritime continent, with westerly flow at upper levels, descent over the eastern Pacific and a return, easterly flow at low levels. Fluctuations in the strength and structure of the Walker circulation are typically measured by considering the sea-level pressure difference between Darwin in northern Australia and Tahiti in the mid Pacific. Variations in this pressure gradient, which is associated with a change in the region of preferred convection, are known as the *Southern Oscillation* (SO).

Variations in the Walker circulation are part of the *El-Niño Southern Oscillation* (*ENSO*), a major coupled oscillation of the atmosphere and ocean. Normally, the sea surface temperature (SST) on the Equator is several degrees colder in the East Pacific that in the West Pacific. Every few years, changes in the ocean lead to a warming in the sea temperatures off the coast of South America (since this typically occurs around Christmas time, the phenomenon was dubbed *El Niño*). The main area of convective activity that is normally centred over the western

Pacific moves eastward and as a result the Walker circulation is modified (see, e.g., http://www.pmel.noaa.gov/tao/elnino/nino-home.html). Conditions in which there is an enhanced warming of the waters over the maritime continent and western Pacific, and cold SSTs over the eastern Pacific are termed *La Niña*. The main features of *ENSO* are summarized as follows:

- Normally, SSTs in the western tropical Pacific (∼30°C) are warmer than in the eastern tropical Pacific (∼23°C). These temperatures are maintained by westward (easterly) low-level winds, which through *Ekman pumping* cause upwelling of cooler waters in the eastern Pacific. There is strong convection over the warm western Pacific, associated with Walker circulation;
- The atmosphere-ocean system over the Pacific can become unstable if disturbed, e.g., by sequences of westerly wind bursts. In this situation, warm surface water spreads out into the central and eastern Pacific, and convection moves eastward. The low-level wind in the western Pacific reverses to become eastward (westerly); the thermocline rises in the western Pacific and deepens in the eastern Pacific. These changes reinforce each other and are a *positive feedback*;
- The anomalous El Niño is terminated by upwelling oceanic *Kelvin waves* which propagate from the western to the eastern Pacific. The upwelling Kelvin wave is itself a response to oceanic Rossby waves which are generated by the same low-level wind anomalies which gave rise to the initial El Niño anomaly. This is a *negative feedback*. Hence perturbations to the coupled atmosphere-ocean system in the Pacific both help to create and destroy the El Niño. A simple model which captures many aspects of this behaviour is known as the Delayed Action Oscillator model (Suarez and Schopf 1994).

Typical time-scales for oscillations in the ENSO system are between 2 and 7 years. Hence, the onset of El Niño or La Niña conditions is relatively slow, occurring over several months. Changes to the Walker circulation can also result in changes to atmospheric circulation over large parts of the globe. The influences are often called *teleconnections* since they refer to correlations of climate anomalies over large distances, typically thousands of kilometres. In the case of ENSO, its influence is spread remotely through the generation and propagation of Rossby waves by anomalous latent heating in the middle troposphere by convection in the mid Pacific.

# 4 General Circulation in the Middle Atmosphere: The Brewer–Dobson Circulation

## *4.1 Introduction to the Middle Atmosphere*

The middle atmosphere, encompassing the stratosphere and the mesosphere, shows a series of major contrasts with the troposphere. These are the large-scale wind

and temperature distributions, which include a large annual cycle in the middle atmosphere which is absent in the troposphere (Figs. 1 and 2); the absence of synoptic-scale disturbances in the middle atmosphere, e.g., there are no weather fronts in the middle atmosphere; and the dominant role of diabatic processes in the middle atmosphere.

Although the middle atmosphere contains only a small fraction of the atmosphere (about 10%), it is important for humankind because: it contains most of atmospheric ozone, which cuts out harmful UV radiation; and its temperature is sensitive to changes in the concentration of greenhouse gases (e.g. water vapour, methane and carbon dioxide), so middle atmosphere temperature trends could provide an early signature of climate change. During wintertime, variability in the stratospheric circulation has been shown to have an impact on the tropospheric flow on time-scales from 10 to 60 days (see, e.g., Charlton et al. 2003).

The Lagrangian, meridional circulation in the middle atmosphere is known as the *Brewer–Dobson circulation*. The Brewer–Dobson circulation is characterized by diabatic heating (rising motion) over the Summer Pole; cross-hemispheric motion across the Equator; diabatic cooling (sinking motion) over the Winter Pole (Fig. 5). The Lagrangian picture provides a more accurate representation of the mean advective transport of stratospheric tracers, showing the wintertime descent over the Pole observed in tracer data from instruments aboard NASA's UARS (Upper Atmosphere Research Satellite) and ESA's Envisat satellites (e.g. Lahoz et al. 1996, 2007b).

The temperature and wind distributions in the atmosphere, including the middle atmosphere, have been discussed in Sects. 2.1 and 2.2. We now discuss the winter and summer stratosphere; later in this section we discuss the distributions of humidity and ozone in the middle atmosphere.

## *4.2 Winter and Summer Stratosphere*

In the Winter Hemisphere stratosphere, the temperature and geopotential height decrease toward the Pole, with eastward (westerly) winds and a strong cyclonic vortex centred on or near the Pole – the *polar vortex*. The winter polar vortex is stronger in the Southern Hemisphere than in the Northern Hemisphere, as the former has generally colder temperatures. In the Summer Hemisphere stratosphere, the temperature and geopotential height increases toward the Pole, with westward (easterly) winds and an anticyclonic vortex centred on or near the Pole – the *summertime high*. Observations of the stratospheric flow show that the wintertime, cyclonic vortex is periodically disturbed by the growth and decay of large, planetary scale structures. In Northern Hemisphere winter, there is a quasi-permanent anticyclone feature near the International Date Line, called the *Aleutian high.* In contrast the summertime anticyclonic vortex is almost undisturbed.

Perturbations of the stratospheric flow are caused by the growth and upward propagation of stationary Rossby waves (see Sect. 3.3). A key property of Rossby waves is that their vertical propagation is governed by the mean flow according to

the *Charney-Drazin theory* (Andrews et al. 1987). Waves can only propagate upward where the mean flow is westerly and less than a critical value, $U_c$. The critical velocity for upward propagation is also inversely related to horizontal wavenumber, and $U_c$ is much smaller for waves of higher wavenumber. The strength of the stratospheric jet is such that only waves with wavenumber 1 and 2 can readily propagate into the stratosphere. The Charney-Drazin theory therefore provides a first-order explanation of observed stratospheric variability. The wintertime stratosphere is dominated by large-scale quasi-stationary disturbances to the mean flow because only the largest, and hence stationary, Rossby waves can propagate in the strong wintertime jet. The summertime stratosphere is almost undisturbed, because the zonal mean flow is easterly, and below the threshold for Rossby wave propagation, hence it is largely free of planetary waves.

Further insight can be gained by considering the distribution of planetary wave activity in the Northern and Southern Hemispheres. In the extra-tropics, the Northern Hemisphere has a largely wavenumber two topography due to the presence of the Himalayas and Rocky Mountain chains, providing strong forcing for large-scale Rossby waves. In contrast, the Southern Hemisphere is relatively flat in the extra-tropics and has much lower stationary wave amplitudes than the Northern Hemisphere. Since Rossby waves are the prime reason for variability in the stratosphere, the contrast in planetary wave amplitudes in each hemisphere is strongly correlated with the strength of the corresponding wintertime polar vortex and temperature near the pole. The contrasts in stratospheric wintertime behaviour in the two hemispheres are largely the result of these differences in tropospheric dynamical behaviour, rather than any intrinsic difference in the radiative properties of the two hemispheres. It should also be noted, however, that many of the simplifying assumptions that are used to derive the Charney-Drazin theory are violated in the real atmosphere and the reasons that the theory appears to work in the simplest terms are not well understood.

A further important property of Rossby waves in the stratosphere is that when they encounter their critical velocity they can "break". The interaction of Rossby waves with critical lines is a complex subject which we do not cover in detail here. A key point, however, is that when Rossby waves break, they cause easterly acceleration of the mean flow. Rapid deceleration of the mean flow can lead to dramatic changes to the stratospheric flow known as *stratospheric sudden warmings*.

*Stratospheric sudden warmings.* Stratospheric sudden warmings (SSWs) are disturbances to the wintertime polar vortex caused by the breaking of transient Rossby waves. Although some variability of this kind is almost ubiquitous during Northern Hemisphere winter, it is customary when examining SSWs to focus on events in which the zonal mean Pole-to-Equator temperature gradient is significantly disturbed. When planetary waves break in the stratosphere an anomalous meridional circulation (of the kind discussed in previous sections) can occur, with enhanced descent and adiabatic warming over the Pole. Extremely rapid temperature changes at the Pole, of the order 80 K in 5 days, have been observed during SSWs, giving them the "sudden" part of their name.

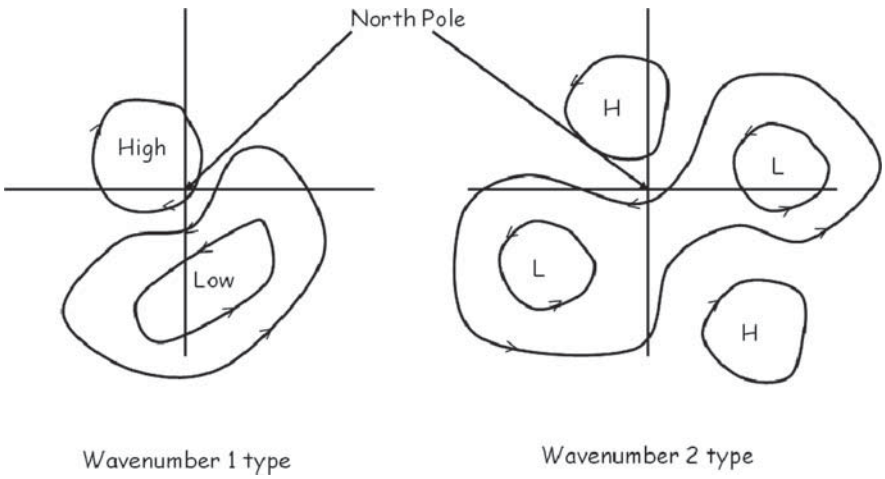



**Fig. 6** Schematic of a wavenumber-1 type warming (*left hand plot*) and a wavenumber-2 type warming (*right hand plot*). The "High" (H) and "Low" (L) labels indicate relatively high and low geopotential height (analogously, pressure) centres. The *arrows* provide a sense of the wind circulation

SSWs can be classified into major and minor events. In a *major warming* a reversal of the zonal mean zonal wind (westerlies become easterlies) occurs from the upper stratosphere down to 10 hPa and poleward of 60°N (Northern Hemisphere), 60°S (Southern Hemisphere). In a *minor warming* these conditions are not satisfied, but there is still weakening of the winds and an increase of the temperatures.

Major SSWs tend to occur with two different synoptic evolutions or "habits" (Charlton and Polvani 2007; see Fig. 6). In the *vortex displacement* or wavenumber 1 type of SSW, the polar vortex is displaced from the Pole toward eastern Eurasia by growth of the Aleutian anticyclone. In the *vortex splitting* or wavenumber 2 type of SSW the vortex remains close to the Pole, but elongates and splits into two similarly sized pieces. The vortex displacement type also has a strong westward tilt in the vertical, whereas the vortex splitting type has little tilt in the vertical (Matthewman et al. 2009).

Major SSWs are relatively rare phenomena; there are typically 6 events per decade in the Northern Hemisphere. Most events occur during the months of January and February, although events in December are not uncommon. There can also be distinct decadal variability in the dynamical activity of the vortex, for example during the period March 1992–November 1998 there were no major SSW events, while the period November 1998 to present has had an anomalously large number of major SSW events. Our relatively short records of stratospheric variability make it difficult to fully determine the behaviour in the Southern Hemisphere, but it is thought that there has been only one major SSW since the early 1960s, in September 2002. The current generation of general circulation models (GCMs) which resolve the stratosphere are able to produce SSWs with similar dynamical properties to those

in observations, but generally have a lower frequency of SSWs than those seen in observations (Charlton et al. 2007).

At the end of the winter, the *final warming* occurs. At this time, the polar vortex breaks down completely and there is a transition from winter to summer conditions. This event is the result of both radiative and dynamical influences since the increased solar heating over the Pole and eventual reversal of the Pole-to-Equator temperature gradient will lead to a reversal of the zonal mean stratospheric jet. The timing of the final warming is variable in the Northern Hemisphere, occurring any time between March and May; the timing is more regular in the Southern Hemisphere, where the polar vortex commonly breaks down during the period October/November.

The presence of major warmings and the timing of the final warming have a bearing on the temperature distribution in the winter stratosphere. As we shall see later, this temperature distribution has a bearing on the conditions that can cause ozone loss in the Arctic or the Antarctic via heterogeneous chemistry (see chapter *Introduction to Atmospheric Chemistry and Constituent Transport*, Yudin and Khattatov).

## *4.3 Humidity*

Water vapour plays an important role in the radiation budget of the atmosphere, especially in the *upper troposphere/lower stratosphere* (*UTLS*) region. It also provides information on the atmospheric circulation (as it is a tracer on seasonal time-scales); it is a source of $HO_x$ ($=OH+HO_2$, involved in the catalytic destruction of ozone – see chapter *Introduction to Atmospheric Chemistry and Constituent Transport*, Yudin and Khattatov); and it is a constituent of the *Polar Stratospheric Clouds* (*PSCs*) involved in polar ozone loss (Dessler 2000).

A feature of the middle atmosphere that distinguishes it from the troposphere is its dryness. The typical humidity profile increases from minimum values of about 2.5 ppmv (parts per million by volume) at the *hygropause* (located near the tropopause) to about 8 ppmv in the lower mesosphere, above which values decrease – see, e.g., Lahoz et al. (2007a) and chapter *Constituent Assimilation* (Lahoz and Errera). These humidity values in the middle atmosphere are orders of magnitude smaller than typical values in the lower troposphere. We now explain this vertical variation in the humidity.

It has been understood since the late 1940s (Brewer 1949) that the only possible explanation for the extreme dryness in the middle atmosphere is that air must experience very cold temperatures as it enters the stratosphere from the troposphere. The resulting very small saturation mixing ratios mean that almost all of the moisture condenses and precipitates out of the air. The cold temperatures required for saturation mixing ratios of a few ppmv are about 185 K (about –88°C), and occur near the

tropical tropopause; this indicates that air must enter the stratosphere preferentially through the tropical tropopause.

Reasoning this way, Brewer deduced the existence of a mean meridional circulation in the stratosphere with ascent across the tropical tropopause, poleward transport, and descent across the extra-tropical tropopause; this motion is the Brewer–Dobson circulation and is well represented by the Lagrangian picture described above (Fig. 5).

Water vapour is much more homogeneously distributed in the middle atmosphere than in the troposphere, but it does increase gradually with height because chemical reactions involving methane oxidation are a source of water vapour. The longer air is in the stratosphere, the more likely it will be that methane will have been oxidized to form water vapour. Thus, air descending in the wintertime polar vortex is relatively moist, and the isolation of the polar vortex results in a coherent mass of air located at high latitudes. In the mesosphere, photolysis is a sink of water vapour. At high latitudes in the Antarctic winter lower stratosphere, temperatures are low enough to cause water to freeze out and reduce water vapour mixing ratios.

Recent balloon observations from Boulder, Colorado (USA) suggest that water vapour in the stratosphere is slowly increasing (Fig. 7). However, recent water vapour observations from the UARS HALOE instrument suggest, on the contrary, that water vapour in the stratosphere is decreasing (Randel et al. 2006; Fig. 8). Owing to these discrepancies, the water vapour variability in the stratosphere is a current topic of research.

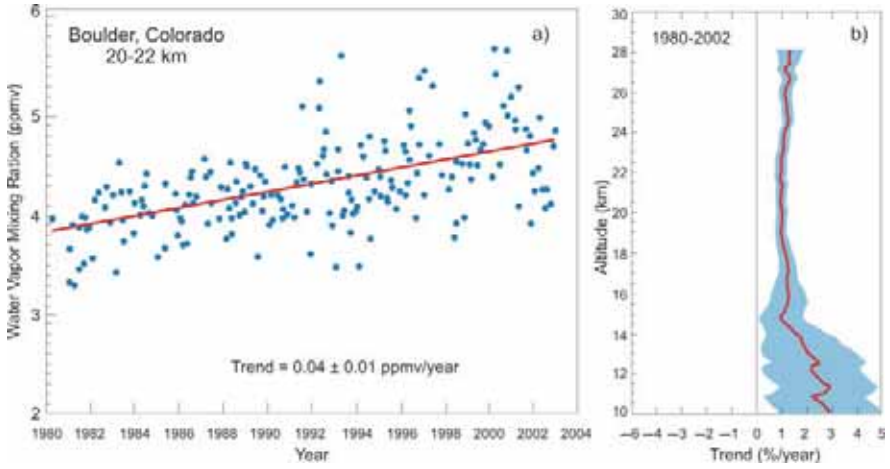



**Fig. 7** Water vapour observations over Boulder, Colorado (USA), based on balloon data from the period 1980–2002. (**a**) water vapour measurements over 20–22 km; (**b**) vertical profile of the water vapour trend (%/year) for the period 1980–2002. Source: http://commons.wikimedia.org/wiki/Image:BAMS_climate_assess_boulder_water_vapor_2002.png

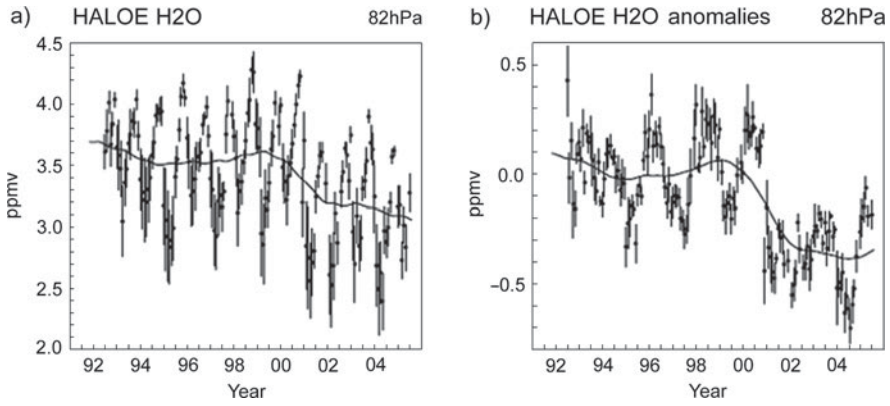



**Fig. 8** (**a**) Time series of near-global mean (60°N–60°S) water vapour at 82 hPa derived from HALOE, HALogen Occultation Experiment, data (1992–2005). The circles show monthly mean values, and *error bars* show the monthly standard deviation; (**b**) Deseasonalized near-global mean water vapour anomalies at 82 hPa. In both panels the *solid lines* represent running Gaussian-weighted means of the invididual points, using a Gaussian half-width of 12 months. With permission from Randel et al. (2006)

## *4.4 Ozone*

Most of the ozone in the atmosphere (∼90%) resides in the stratosphere, with only about 10% residing in the troposphere. Ozone is more abundant in the stratosphere because its main source is photochemical reactions in the tropical middle stratosphere; the largest values of ozone mixing ratio occur there (at heights ∼30 km; pressures ∼10 hPa). However, the largest *ozone column* amounts (an integration of ozone from the ground upwards) are found at high latitudes, because the stratospheric contribution to the column is larger at high latitudes and, there, the troposphere has a smaller vertical extent.

Ozone plays a major role in determining the middle atmosphere temperature structure by absorbing solar radiation. It also plays a significant role by absorbing and emitting thermal infrared radiation. The strong peak in solar heating around the stratopause is chiefly due to absorption of solar radiation by ozone. To a large extent, this explains the existence of a warm stratopause. Note that the heating peak occurs above the ozone mixing ratio peak and not at the mixing ratio peak. This is because solar radiation is absorbed by ozone on its way down through the atmosphere and thus less radiation reaches 10 hPa (roughly 30 km) than 1 hPa (roughly 50 km). However, the ozone distribution and solar heating cannot explain the existence of a warm stratopause near the Winter Pole, where there is no sunlight. This shows that dynamical processes are also important in determining the temperature structure of the middle atmosphere (Shine 1987).

The ozone created photochemically in the tropical middle stratosphere is transported by the stratospheric mean meridional circulation from the tropical region poleward and downward to the high latitude lower stratosphere (see, e.g., Fig. 5).

Ozone is destroyed in the wintertime lower stratosphere, the result being the well-known *ozone hole* (Farman et al. 1985). The amount of ozone loss is greater in the Antarctic than in the Arctic. This is because the conditions required for an ozone hole: low temperatures (no sunlight during winter); isolation (formation of winter polar vortex) are more prevalent in the Antarctic than in the Arctic, with the former generally having lower temperatures and having a stronger and longer lasting winter polar vortex. These conditions (low temperatures, isolation) allow the formation of PSCs, on which *heterogeneous* (surface chemistry) reactions can take place to liberate chlorine compounds which can be photolysed to release chlorine atoms when the sun returns in spring. These chlorine atoms are then involved in catalytic reactions that destroy ozone, forming the ozone hole. We thus see how wintertime conditions at high latitudes affect the ozone distribution.

### *4.5 Interaction Between Dynamics, Radiation and Chemistry*

Most of the stratosphere is close to radiative equilibrium, although major disturbances to this equilibrium can occur, for example during SSWs, as previously discussed. This approximate radiative equilibrium explains most features of the stratospheric zonal mean temperature structure (Fig. 1): temperature increases systematically from Winter Pole to Summer Pole; the warm stratopause coincides with a peak in solar radiative heating due to absorption by ozone. Note, however, that as mentioned above, radiative processes on their own cannot explain all of the temperature structure, and appeal must be made to dynamical processes.

In the Lagrangian picture (Fig. 5) diabatic processes are associated with vertical motion, in particular diabatic heating with ascent and diabatic cooling with descent. Note that in the *downward control* picture, Haynes et al. (1991), vertical motion is a consequence of wave breaking in the upper middle atmosphere. The wintertime Antarctic stratosphere is colder than its Arctic counterpart. Thus, the former is closer to radiative equilibrium, so that diabatic descent rates in the Antarctic winter stratosphere tend to be smaller than those in the Arctic winter stratosphere.

Temperature is also important for chemical processes in the stratosphere. Chemical reactions depend on temperature (see chapter *Introduction to Atmospheric Chemistry and Constituent Transport*, Yudin and Khattatov). Also, as seen above, temperature plays a key role in setting up the conditions required for the formation of PSCs, the presence of heterogeneous chemistry processes, and the eventual formation of the ozone hole.

## 5 Conclusions

This chapter provides a general overview of the atmospheric circulation from the ground to about 80 km, highlighting the main concepts that are important for a basic understanding of meteorology and dynamics. Meteorology and dynamics, together

with radiation and chemistry, help determine the distribution of the key atmospheric species ozone and water vapour. These species are of interest to the data assimilation community, both operational and research.

## References

Andrews, D.G., J.R. Holton, and C.B. Leovy, 1987. *Middle Atmosphere Dynamics,* Academic Press, New York, 489 pp.

Andrews, D.G. and M.E. McIntyre, 1976. Planetary waves in horizontal and vertical shear: The generalized Eliassen-Palm relation and the mean zonal circulation. *J. Atmos Sci.*, **33**, 2031–2048.

Brewer, A.W., 1949. Evidence for a world circulation provided by measurements of helium and water vapour distribution in the stratosphere. *Q. J. R. Meteorol. Soc.*, **75**, 351–363.

Charlton, A.J., A. O'Neill, D.B. Stephenson, et al., 2003. Can knowledge of the state of the stratosphere be used to improve statistical forecasts in the troposphere? *Q. J. R. Meteorol. Soc.*, **129**, 3205–3224.

Charlton, A.J. and L.M. Polvani, 2007. A new look at stratospheric sudden warmings. Part I: Climatology and modeling benchmarks. *J. Clim.*, **20**, 449–469.

Charlton, A.J., L.M. Polvani, J. Perlwitz, et al., 2007. A new look at stratospheric sudden warmings. Part II. Evaluation of numerical model simulations. *J. Climate*, **20**, 471–488, doi:10.1175/JCLI3994.1.

Dessler, A.E., 2000. *The Chemistry and Physics of Stratospheric Ozone,* Academic Press, New York, 214 pp.

Farman, J.C., B.G. Gardiner, and J.D. Shanklin, 1985. Large losses of total ozone in Antarctica reveal seasonal ClOx/NOx interaction. *Nature*, **315**, 207–210.

Fleming, E.L., S. Chandra, M.R. Schoeberl, and J.J. Barnett, 1988. Monthly Mean Global Climatology of Temperature, Wind, Geopotential Height and Pressure for 0–120 km. NASA Technical Memorandum 100697, February 1988. Available from http://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/19880013119_1988013119.pdf.

Haynes, P.H., C.J. Marks, M.E. McIntyre, T.G. Shepherd, and K.P. Shine, 1991. On the "Downward Control" of extratropical diabatic circulations by eddy-induced mean zonal forces. *J. Atmos. Sci.*, **48**, 651–678.

Heifetz, E., C.H. Bishop, B.J. Hoskins, and J. Methven, 2004a. The counter-propagating Rossby wave perspective on baroclinic instability. Part I: Mathematical basis. *Q. J. R. Meteorol. Soc.,* **130**, 211–231.

Heifetz, E., J. Methven, B.J. Hoskins, and C.H. Bishop, 2004b. The counter-propagating Rossby wave perspective on baroclinic instability. Part II: Application to the Charney model. *Q. J. R. Meteorol. Soc.,* **130**, 233–258.

Held, I.M. and A.Y. Hou, 1980. Nonlinear axially symmetric circulations in a nearly inviscid atmosphere. *J. Atmos. Sci.*, **37**, 515–533.

Holton, J.R., 2004. *An Introduction to Dynamic Meteorology,* 4th edition, Elsevier Academic Press, London, 553 pp.

Lahoz, W.A., Q. Errera, R. Swinbank, and D. Fonteyn, 2007a. Data assimilation of constituents: A review. *Atmos. Chem. Phys.*, **7**, 5745–5773.

Lahoz, W.A., A.J. Geer, and Y. Orsolini, 2007b. Northern hemisphere stratospheric summer from MIPAS observations. *Q. J. R. Meteorol. Soc.*, **133**, 197–211.

Lahoz, W.A., A. O'Neill, A. Heaps, et al., 1996. Vortex dynamics and the evolution of water vapour in the Stratosphere of the Southern Hemisphere. *Q. J. R. Meteorol. Soc.*, **122**, 423–450.

Matthewman N.J., J.G. Esler, A.J. Charlton-Perez, and L.M. Polvani, 2009. A new look at stratospheric sudden warmings. Part III. Polar vortex evolution and vertical structure. *J. Climate*, DOI: 10.1175/2008JCLI2365.1.

Methven, J., E. Heifetz, B.J. Hoskins, and C.H. Bishop, 2005a. The counter-propagating Rossby wave perspective on baroclinic instability. Part III: Primitive equation disturbances on the sphere. *Q. J. R. Meteorol. Soc.,* **131**, 1393–1424.

Methven, J., B.J. Hoskins, E. Heifetz, and C.H. Bishop, 2005b. The counter-propagating Rossby wave perspective on baroclinic instability. Part IV: Nonlinear life cycles. *Q. J. R. Meteorol. Soc.,* **131**, 1425–1440.

Peixoto, J.P. and A.H. Oort, 1992. *Physics of Climate,* American Institute of Physics, New York, 565 pp.

Randel, W.J., F. Wu, H. Vömel, G.E. Nedoluha, and P. Forster, 2006. Decreases in stratospheric water vapor after 2001: Links to changes in the tropical tropopause and the Brewer–Dobson circulation. *J. Geophys. Res.*, **111**, D12312, doi:10.1029/2005JD006744.

Rees, M.H., 1989. *Physics and Chemistry of the Upper Atmosphere*, Cambridge University Press, Cambridge, 304 pp.

Shine, K.P., 1987. The middle atmosphere in the absence of dynamical heat fluxes. *Q. J. R. Meteorol. Soc.*, **113**, 603–633.

Suarez, M.J. and P.S. Schopf, 1994. A delayed action oscillator for ENSO. *J. Atmos. Sci.,* **45**, 3283–3287.

Swinbank, R., 1985. The global atmospheric angular momentum balance inferred from analyses made during the FGGE, *Q. J. R. Meteorol. Soc*., **111**, 977–992.

Vallis, G.K., 2007. *Atmospheric and Ocean Fluid Dynamics*, Cambridge University Press, Cambridge, 745 pp.

# The Role of the Model in the Data Assimilation System

**Richard B. Rood**

## 1 Introduction

The chapters in Part I, *Theory*, describe in some detail the theory and methodology of data assimilation. This chapter will focus on the role of the predictive model in an assimilation system. There are numerous books on atmospheric modelling, their history, their construction, and their applications (e.g. Trenberth 1992; Randall 2000; Jacobson 2005). This chapter will focus on specific aspects of the model and modelling in data assimilation.

The chapter is outlined as follows:

- Definition and Description of the Model;
- Role of the Model in Data Assimilation;
- Component Structure of an Atmospheric Model;
- Consideration of the Observation-Model Interface;
- Physical Consistency and Data Assimilation;
- Summary.

## 2 Definition and Description of the Model

Dictionary definitions of *model* include:

- "A work or construction used in testing or perfecting a final product";
- "A schematic description of a system, theory, or phenomenon that accounts for its known or inferred properties and may be used for further studies of its characteristics".

R.B. Rood (✉)
University of Michigan, Ann Arbor, MI, USA
e-mail: rbrood@umich.edu

W. Lahoz et al. (eds.), *Data Assimilation*, DOI 10.1007/978-3-540-74703-1_14,

In atmospheric modelling a scientist is generally faced with a set of observations of variables, for instance, wind, temperature, water, ozone, etc., as well as either the knowledge or expectation of correlated behaviour between the different variables. A number of types of models could be developed to describe the observations. These include:

- Conceptual or heuristic models which outline in the simplest terms the processes that describe the interrelation between different observed phenomena. These models are often intuitively or theoretically based. An example would be the tropical pipe model of Plumb and Ko (1992), which describes the transport of long-lived tracers in the stratosphere;
- Statistical models which describe the behaviour of the observations based on the observations themselves. That is the observations are described in terms of the mean, the variance, and the correlations of an existing set of observations. Johnson et al. (2000) discuss the use of statistical models in the prediction of tropical sea surface temperatures;
- Physical models which describe the behaviour of the observations based on first principle tenets of physics (chemistry, biology, etc.). In general, these principles are expressed as mathematical equations, and these equations are solved using discrete numerical methods. Detailed discussions of modelling include Trenberth (1992), Randall (2000), and Jacobson (2005).

In the study of geophysical phenomena, there are numerous subtypes of models. These include comprehensive models which attempt to model all of the relevant couplings or interactions in a system and mechanistic models which have prescribed variables, and the system evolves relative to the prescribed parameters. All of these models have their place in scientific investigation, and it is often the interplay between the different types and subtypes of models that leads to scientific advance.

Models are used in two major roles. The first role is *diagnostic*, in which the model is used to determine and to test the processes that are thought to describe the observations. In this case, it is determined whether or not the processes are well known and adequately described. In general, since models are an investigative tool, such studies are aimed at determining the nature of unknown or inadequately described processes. The second role is *prognostic*; that is, the model is used to make a prediction.

In all cases the model represents a management of complexity; that is, a scientist is faced with a complex set of observations and their interactions and is trying to manage those observations in order to develop a quantitative representation. In the case of physical models, which are implicitly at focus here, a comprehensive model would represent the cumulative knowledge of the physics (chemistry, biology, etc.) that describe the observations. It is tacit, that an accurate, validated, comprehensive physical model is the most robust way to forecast; that is, to predict the future.

The *physical principles* represented in an atmospheric model, for example, are a series of *conservation equations* which quantify the conservation of momentum, mass, and thermodynamic energy. The equation of state describes the relation

between the thermodynamic variables. Because of the key roles that phase changes of water play in atmospheric energy exchanges, an equation for the conservation of water is required. Models which include the transport and chemistry of atmosphere trace gases and aerosols require additional conservation equations for these constituents. The conservation equations for mass, trace gases, and aerosols are often called *continuity equations*.

In general, the conservation equation relates the time rate of change of a quantity to the sum of the quantity's production and loss. For momentum the production and loss follow from the forces described by Newton's Laws of Motion. Since the atmosphere is a fluid, either a *Lagrangian* or an *Eulerian* description of the flow can be used (see chapter *General Concepts in Meteorology and Dynamics*, Charlton-Perez et al.). The Lagrangian description follows a notional fluid parcel, and the Eulerian description relies on spatial and temporal field descriptions of the flow at a particular point in the domain. Data assimilation can be performed in either the Lagrangian or Eulerian framework. In this chapter the Eulerian framework will be the primary focus. Holton (2004) provides a thorough introduction to the fundamental equations of motions and their scaling and application to atmospheric dynamics.

In order to provide an overarching background, it is useful to consider the elements of a modelling, or simulation, framework described in Fig. 1. In this framework are six major ingredients. The first are the boundary and initial conditions. For an atmospheric model, boundary conditions include topography, sea surface temperature, land type, vegetation, etc.; boundary conditions are generally prescribed from external sources of information.

The next three items in the figure are intimately related. They are the representative equations, the discrete and parametrized equations, and the constraints drawn from theory. The representative equations are the continuous forms of the conservation equations. The representative equations used in atmospheric modelling are approximations derived from scaling arguments (see Holton 2004); therefore, even the equations the modeller is trying to solve have a priori simplification which can be characterized as errors. The continuous equations are a set of non-linear partial differential equations. The solutions to the representative equations are a balance amongst competing forces and tendencies.

The discrete and parametrized equations arise because it is not possible to solve the representative equations in analytical form. The strategy used by scientists is to develop a numerical representation of the equations. One approach is to develop a grid of points which covers the spatial domain of the model. Then a discrete numerical representation of those variables and processes which can be resolved on the grid is written. Processes which take place on spatial scales smaller than the grid are parametrized. These approximate solutions are, at best, discrete estimates to solutions of the analytic equations. The discretization and parametrization of the representative equations introduce a large source of error. This introduces another level of balancing in the model; namely, these errors are generally managed through a subjective balancing process that keeps the numerical solution from producing obviously incorrect estimates.

Simulation Framework
(General Circulation Model, "Forecast")

| | | |
|---|---|---|
| Boundary Conditions | Emissions, SST, topography, ... | $\varepsilon$ |
| Representative Equations | $DA/Dt = P - LA$ | $\varepsilon$ |
| Discrete/Parametrize | $(A_{n+\Delta t} - A_n)/\Delta t = \ldots$ | $(\varepsilon_d, \varepsilon_p)$ |
| Theory/Constraints | $\partial u_g/\partial z = -(\partial T/\partial y)R/(Hf_0)$ | Scale Analysis |
| Primary Products (*i.e.* A) | T, u, v, ) , $H_2O$, $O_3$ ... | $(\varepsilon_b, \varepsilon_v)$ |
| Derived Products (F(A)) | Pot. Vorticity, v*, w*, ... | Consistent |

$(\varepsilon_b, \varepsilon_v)$ = (bias error, variability error)

**Fig. 1** A schematic description of the conceptual elements of an atmospheric model formulation. The boundary conditions include, for example, emissions of trace gases, sea surface temperature (SST), and topography. There are a set of partial differential equations that are the "Representative Equations", i.e., the conservation principles important in the physics (and chemistry, biology, . . .) of the atmosphere. Here, there is a generic variable $A$, and its change with respect to time, $t$, is equal to its Production, $P$, minus Loss, which is proportional to a Loss Frequency ($L$) and the amount of $A$. These partial differential equations are, usually, specified following scale analysis and approximation for the particular application targeted by the model. The Representative Equations are represented as numerical approximations ("Discrete/Parametrize"), where the index, $n$, represents a step in time of increment $\Delta t$. The "Theory/Constraints" are important to robust model formulation. Here, the *geostrophic approximation* is used as an example. It is important that the numerical methods represent the theoretical constraints that are obtained, for instance, by scale analysis. The "Primary Products" are those products for which there is a prognostic equation. The "Derived Products" are either diagnosed from the primary products or as a function of the primary products. Here potential vorticity and the residual circulation are used as examples. $\varepsilon$ represents the error that is present at all stages of the model formulation

While all of the terms in the analytic equation are potentially important, there are conditions or times when there is a dominant balance between, for instance, two terms. An example of this is *thermal wind balance* in the middle latitudes of the atmosphere (see Holton 2004; see chapter *General Concepts in Meteorology and Dynamics*, Charlton-Perez et al.). It is these balances, generally at the extremes of spatial and temporal scales, which provide the constraints drawn from theory. Such constraints are generally involved in the development of conceptual or heuristic models. If the modeller implements discrete methods which consistently represent the relationship between the analytic equations and the constraints drawn from theory, then the modeller maintains a substantive scientific basis for the interpretation of model results.

The last two items in Fig. 1 represent the products that are drawn from the model. These are divided into two types: *primary products* and *derived products*. The primary products are variables such as wind, temperature, water, ozone – parameters that are most often, explicitly modelled; that is, an equation is written for them. The primary products might also be called the resolved or prognostic variables. The derived products are of two types. The first type is those products which are diagnosed from model state variables, often in the parametrized physical processes. The second type follows from functional relationships between the primary products;

for instance, potential vorticity (Holton 2004). A common derived product is the budget – the sum of the different terms of the discretized conservation equations. The budget is studied, explicitly, on how the balance is maintained and how this compares with budgets derived directly from observations or reanalysis (see chapter *Reanalysis: Data Assimilation for Scientific Investigation of Climate*, Rood and Bosilovich).

In some cases the primary products can be directly evaluated with observations, and errors of bias and variability are estimated. If attention has been paid in the discretization of the analytic equations to honour the theoretical constraints, then the derived products will behave consistently with the primary products and theory. They will have errors of bias and variability, but when a budget is formed from the sum of the terms in the conservation equations, it will balance. That is, the discrete form of the conservation equation is solved. In this case the correlative relation between variables is represented and there is a "physical" consistency.

## 3 The Role of the Model in Data Assimilation

*Data assimilation* is the melding of observational information with information provided by a model (Daley 1991; Kalnay 2003; Swinbank et al. 2003). In assimilation for Earth system science, all types of models, conceptual, statistical, and physical, are used. Models are used in both their prognostic and diagnostic roles. First and foremost in data assimilation, the model provides an estimate of the expected value of the state variables that are observed and assimilated. The discussion, which follows, centres on this role of *state estimation*.

The focus here is on physically based models of the atmosphere formulated in an Eulerian description of the fluid dynamics. Outside of the atmospheric model (or more generally geophysical models) there are other models in the data assimilation system. Notably, because of the complexity of expressing error covariances, these are generally modelled. Also, there are forward and inverse models which transfer quantities between observed quantities, for example radiances observed by a satellite instrument, and geophysical quantities, for example corresponding temperature estimates. These types of models are discussed elsewhere in the book; see, e.g., the chapters in Part II (*Observations*) and Part IV (*Atmospheric Chemistry*), and the companion chapters in Part III (*Meteorology and Atmospheric Dynamics*).

A schematic of an assimilation system is given in Fig. 2. This is a *sequential* assimilation system where a forecast is provided to a statistical analysis algorithm that calculates the merger of model and observational information. Some assimilation methods cycle back and forth between these steps to assure maximum coherence. In this figure, errors are specified based on external considerations and methods. There is a formal interface between the statistical analysis algorithm and the model prediction which performs a quality assessment of the information prior to the merger. This interface might also include a balancing process called *initialization* (see Lynch 2003; see chapter *Initialization*, Lynch and Huang). The figure

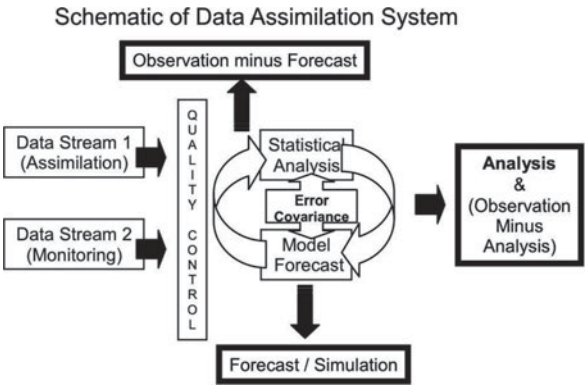



**Fig. 2** A schematic of a Data Assimilation System. This is a sequential assimilation system where a "Model Forecast" is provided to a "Statistical Analysis" algorithm that calculates the merger of model and observational information using "Error Covariance" information. In this figure, errors are specified based on external considerations and methods. There is a formal interface between the statistical analysis algorithm and the model prediction which performs a quality assessment ("Quality Control") of the information prior to the merger. This interface might also include a balancing process called initialization, which is not explicitly shown. There are two input streams for the observations, "Data Stream 1" and "Data Stream 2". The first of these streams represent the observations that will be assimilated with the model prediction. The other input stream represents observations that will not be assimilated. This second stream of observations could be, for example, a new type of observation whose error characteristics are being determined relative to the existing assimilation system. The products from the system are discussed more fully in the text

shows, explicitly, two input streams for the observations. The first of these streams represent the observations that will be assimilated with the model prediction. The other input stream represents observations that will not be assimilated. This second stream of observations could be, for example, a new type of observation whose error characteristics are being determined relative to the existing assimilation system.

From a functional point of view, the model provides a *short-term forecast* of the expected values of the state variables. This forecast is often called the first-guess, the background, or the prior. The background and the observations are mapped to the same space-time domain where they are compared. The model-provided background is used in the data quality control algorithm, as an objective assessment of the quality of the assimilation system, and as a crucial element of the statistical analysis (see, for example, Dee et al. 2001) – see also chapter *Error Statistics in Data Assimilation: Estimation and Modelling* (Buehner). In addition, there may be a formalized process to balance the spatial and temporal attributes of the features represented (or not represented) in both the model and the observations – initialization. In the statistical analysis, observation-based corrections to the background are determined based on the error characteristics of both the observations and the modelled forecast. These corrections are applied to the background and replace the existing values in the model. These new, corrected values provide the initial conditions for the next model forecast.

The specification of model-error covariances and their evolution with time is a difficult problem. In order to get a handle on these problems it is generally assumed that the observational errors and model errors are unbiased over some suitable period of time, e.g. the length of the forecast between times of data insertion. It is also assumed that the errors are in a Gaussian distribution. The majority of assimilation theory is developed based on these assumptions, which are, in fact, not realized. In particular, when the observations are biased, there would the expectation that the actual balance of geophysical terms is different from the balance determined by the model in the assimilation process. Furthermore, since the biases will have spatial and temporal variability, the balances determined by the assimilation are quite complex. Aside from biases between the observations and the model prediction, there are biases between different observation systems of the same parameters. These biases are potentially correctible if there is a known standard of accuracy defined by a particular observing system. However, the problem of bias is a difficult one to address and perhaps the greatest challenge facing assimilation (see Dee 2005). Bias is discussed in chapter *Bias Estimation* (Ménard).

Figure 2 above shows a set of products which comes from the assimilation system. These are (see chapters *Mathematical Concepts of Data Assimilation*, Nichols; *Evaluation of Assimilation Algorithms*, Talagrand):

- *Analysis:* The analysis is the merged combination of *model information* and *observational information*. The analysis is the estimate of the state of the system (in this case the atmosphere) based on the optimization criteria and error estimates;
- *Forecast/simulation:* The forecast/simulation is a model run that starts from an initial condition defined by the analysis. For some amount of time this model run is expected to represent the state of the system with some deterministic accuracy. For this case the model run is a forecast. After a certain amount of time the model run is no longer expected to represent the particular state of the system; though, it might represent the average state and the variance (i.e., the climate). In this case the model run is simply a simulation that has been initialized with a realistic state estimate at some particular time;
- *Observation minus forecast increment:* The observation minus forecast (O-F) increment gives a raw estimate of the agreement of the forecast information (i.e*.,* the first guess) with the observation information prior to assimilation. Usually, a small O-F increment indicates a high quality forecast, and O-F increments are used as a primary measure of the quality of the assimilation. O-F increments are exquisitely sensitive to changes in the system and are the primary quantity used for monitoring the stability and quality of the input data streams. Study of the O-F increment is useful for determining the spatial and temporal characteristics of some model errors;
- *Observation minus analysis increment:* The observation minus analysis (O-A) increment represents the actual changes to the model forecast that are derived from the statistical analysis algorithm. Therefore, they represent in some bulk sense the error weighted impact of the O-F increments. If the assimilation system

weighs the observations heavily relative to the forecast, then the O-A increments will have significant differences relative to the O-F increments. The opposite is also true; if the model information is weighed more heavily than the observational information then there will be little change represented by the O-F increments. If either of these extremes are realized the basic assumptions of the assimilation problem need to be reconsidered.

Assimilated data products are often said to be "value-added" (see also chapter *Data Assimilation and Information*, Lahoz et al.) The extra value comes from combining two sources of information under the premise that if the error sources are well represented and if the combination process is robust, then there is more information than in either individual source. The two basic sources of information are observed information and model information. Hence, if there is value added to the observed information, then that value comes from the model. Both the prognostic and the diagnostic attributes of the model contribute to the added value.

There are a number of types of information expected to come from the model. The observations are distributed in both space and time. The observations have different attributes; for instance, some observations are point values, while others represent deep layer means. The observations are not continuous; there are spatial and temporal gaps. The model represents the flow of the atmosphere. The model, therefore, takes the information from the observations and propagates that information. This fills in the gaps. Hence, at its most basic level the model is a physically based mapping routine.

From the point of view of information, the model propagates information from observed regions to unobserved regions. If the assimilation is robust, then this greatly improves knowledge of the state in unobserved regions. Further, if at one time a region is not observed and if at a future time the region is observed then, if the model has provided an improved state estimate, then the new observation can better refine the state estimate. That is, there is better use of the observational information. From a different perspective, this comparison of model prediction and observation provides a measure of quality of the assimilation system.

Another function of the model is to transfer information from one observed parameter to other observed parameters. For example, temperature and wind are related to each other. For many years, because temperature observations were by far the most prevalent observation type, temperatures were used to estimate the wind. Elson (1986) compared geostrophic estimates to a number of methods presumed to be more accurate. Such estimates of the ageostrophic wind are crucial, for instance, to weather forecasting and mass transport (see Holton 2004). One place that assimilation has had tremendous impact is in the estimate of mid latitude wind fields, where the geostrophic balance is strong and the wind is strongly influenced by the structure of the temperature field.

Perhaps best viewed as an extension of one observation type influencing another observation type, assimilation also provides estimates of unobserved quantities (see also chapter *Constituent Assimilation*, Lahoz and Errera). One quantity of specific interest is the vertical component of the wind. Because of the strong hydrostatic

stratification of the atmosphere, the vertical component of the wind is three orders of magnitude less than the horizontal components. It is difficult to measure; it remains mostly unmeasured. The vertical wind is, however, critically important to atmospheric physics, linking not only the conservation of thermodynamic energy and momentum, but it also is directly correlated with precipitation and release of latent heat through the condensation of water. Hence a goal of assimilation is to provide meaningful estimates of the vertical component of the wind through the correlated information provided by the temperature and horizontal velocity measurements. There are large errors associated with the estimates of the vertical wind.

The estimate of vertical wind follows from the divergence of the horizontal velocity field. The horizontal velocity is usually a resolved variable, by the nomenclature of Fig. 1, a primary product. Estimates of unobserved quantities also come from the parametrizations used to represent subscale processes. These quantities might include precipitation, clouds, turbulent kinetic energy, or in the case of chemistry-transport models, unobserved chemically reactive constituents or surface emissions. In general, the success of using the assimilation system to estimate unobserved quantities varies widely from one geophysical quantity to another.

Similar in spirit to estimating unobserved quantities, assimilation has the prospect of estimating incompletely observed quantities. An archetypical example is tropospheric ozone. There are many measures of the total amount of ozone in a column above a point on the surface of the Earth. There are also many measures of ozone column above the troposphere. Given the sensitivity of the ozone field to dynamical features in the atmosphere, especially synoptic-scale and planetary-scale waves, the dynamical mapping aspects of assimilation are reasonably expected to offer significant advantage in residual-based estimates of tropospheric ozone (see, for example, Štajner et al. 2008).

As will be discussed more fully below, the products from assimilated data sets may not be physically consistent. There are a number of ways to examine the issue of consistency. As mentioned in the discussion of Fig. 1, the equations of motion tell us that there are expected balances between variables. These balances suggest correlative behaviour between variables that reflect the physical connectivity. There is no reason that independent observations and their errors rigorously represent these balances. Similarly, the observations are not required to sample the mass field such that mass is conserved. We look to the model to develop these balances. How well the model does depends on the time-scales that connect the variables and the strength of the expected correlation and the quality of the observations and the model.

Perhaps the best way to look at the consistency problem is whether or not the conservation equation balances. In a well formulated model the conservation equation is solved; there is precise balance. The insertion of data acts like an additional forcing in the conservation equations. In general, this additional forcing will not average to zero over, say, the time-scale between data insertions. Conservation is not obtained. This is an important point to remember as many users of assimilated data sets assume that because they are essentially continuous in space and time, that the variables balance the conservation equation.

The consequences of this violation of conservation propagate through the model. There are fast modes in the model which will balance quickly and accurately. There are slow modes, for instance those balances revealed in the long-term space and time averages suitable for studying the general circulation, which will be influenced by the forcing that comes from the insertion of data. Hence, the assimilated data products might have better estimates than a free running model of primary products like temperature and wind, but the estimates of the derived products such as precipitation and the Eulerian-mean residual circulation (see Holton 2004) may be worse. That is, the analysis increments (i.e., data insertion) are a major part of the forcing. Molod et al. (1996) was one of the first to document the representation of the moisture and energy budgets in side-by-side free-running climate simulations and assimilated data using the same predictive model as used in the climate simulation.

## 4 Component Structure of an Atmospheric Model

This section lays out the component structure of an atmospheric model. The equations of motion for the atmosphere in tangential coordinates using altitude for the vertical coordinate ($x$, $y$, $z$) are given below (see Holton 2004). The first three equations represent the conservation of momentum components. The fourth equation is the mass continuity equation, and the fifth equation is the thermodynamic energy equation. The last equation is the equation of state (see chapter *General Concepts in Meteorology and Dynamics*, Charlton-Perez et al.).

$$\left.\begin{aligned}
&\frac{Du}{Dt} - \frac{u\mathrm{v}\tan(\phi)}{a} + \frac{uw}{a} = -\frac{1}{\rho}\frac{\partial p}{\partial x} + 2\Omega\mathrm{v}\sin(\phi) - 2\Omega w\cos(\phi) + \nu\nabla^2(u)\\
&\frac{D\mathrm{v}}{Dt} + \frac{u^2\tan(\phi)}{a} + \frac{\mathrm{v}w}{a} = -\frac{1}{\rho}\frac{\partial p}{\partial y} - 2\Omega u\sin(\phi) + \nu\nabla^2(\mathrm{v})\\
&\frac{Dw}{Dt} - \frac{u^2+\mathrm{v}^2}{a} = -\frac{1}{\rho}\frac{\partial p}{\partial z} - g + 2\Omega u\cos(\phi) + \nu\nabla^2(w)\\
&\frac{D\rho}{Dt} = -\rho\nabla\bullet\mathbf{u}\\
&c_v\frac{DT}{Dt} + p\frac{D\alpha}{Dt} = J \text{ or } \frac{c_p}{T}\frac{DT}{Dt} - \frac{R}{P}\frac{Dp}{Dt} = \frac{J}{T}\\
&p = \rho RT \text{ and } \alpha = \frac{1}{\rho}\\
&\frac{D}{Dt} = \frac{\partial}{\partial t} + \mathbf{u}\bullet\nabla
\end{aligned}\right\} \quad (1)$$

In Eq. (1), $t$ is time; $\phi$ is latitude; $a$ is radius of Earth, and $\Omega$ is the angular velocity of the Earth; $g$ is gravity; $\nu$ is a coefficient of viscosity; $c_v$ is specific heat at constant volume and $c_p$ is specific heat at constant pressure; $R$ is the gas constant for air; $\rho$ is density; $T$ is temperature; and $p$ is pressure. ($u$, $v$, $w$) = ($x$ (zonal), $y$ (meridional), $z$ (vertical)) velocity; $J$ is heating.

In addition, equations are needed which describe the conservation of trace constituents (see chapters in Part III, *Atmospheric Chemistry*). The generic form of these continuity equations are:

$$\frac{DQ_i}{Dt} + Q_i \nabla \bullet \mathbf{u} = P_{Q_i} - L_{Q_i} \tag{2}$$

Where $Q_i$ is the density of a constituent identified by the subscript $i$; $P$ and $L$ represent the production and loss from phase changes and photochemistry. An equation for water in the atmosphere, $Q_i = Q_{\mathrm{H2O}}$, is required for a comprehensive model. For water vapour, the production and loss terms are represented by evaporation and condensation. These are associated with significant consumption and release of heat, which must be accounted for in, $J$, the heating, de facto production and loss term of the thermodynamic energy equation. In general, in the atmosphere below the stratopause, heating due to the chemical reactions of trace constituents is assumed not to impact the heat budget of the atmosphere. It is possible for the spatial distribution of trace constituents, for example, ozone, to impact the absorption and emission of radiative energy; hence, there is feedback between the constituent distributions and diabatic processes in the atmosphere.

Water not only affects the atmosphere through the consumption or release of energy due to phase changes, but also affects the radiative balance of the atmosphere through both the distribution of vapour and through the distribution of clouds. Therefore, it is common in modern models to not only represent water vapour, but also to include an equation for cloud water, $Q_i = Q_{\mathrm{cloud}}$, which is partitioned between cloud liquid and cloud ice. The episodic and local scales of the phase changes of water and clouds offer one of the most difficult challenges of atmospheric modelling. This is important for modelling weather, climate, and chemistry.

Due to their impact on both the radiative budget of the atmosphere and formation of cloud water and ice, a set of constituent conservation equations for aerosols is required in a comprehensive atmospheric model. Like water vapour, the spatial and temporal scales of aerosols are often small, below the resolved scales of the model. Again, aerosol modelling provides significant challenges, and they are important for modelling weather and, especially, climate and chemistry.

The equations of motion and a suitable set of constituent continuity equations are the representative equations of the model (see Fig. 1). The equations of motion support many types of dynamical features, for example, waves, such as, Rossby waves, synoptic- or baroclinic-scale waves, gravity waves, Kelvin waves, etc. and vortices, such as hurricanes, tornadoes, etc. There is spatial and temporal heterogeneity in the forcing terms. Hence, the atmosphere is characterized by complexity, and this complexity is confronted when trying to build a predictive model out of the above equations. Further, the complexity is increased by the fact that discrete numerical representations of the equations of motion support a whole variety of behaviour unique to numerical approximation.

Atmospheric models are usually built from components. There are several useful paradigms for organizing their construction. In the first, the model can be divided

into processes, and the solution as a whole is the accumulation of these processes. This is called "process splitting" and has been discussed in, for instance, Strang (1968), Yanenko (1971) and McCrea et al. (1982). Another useful way to look at models is from the perspective of systems engineering, where the whole model system is built from systems of subsystems. This systems approach is useful when formulating strategies for model evaluation and validation; the interacting subsystems determine the performance of the model as a whole. It is, therefore, often difficult to relate model performance to the design characteristics of a particular component.

Recent efforts to organize the modelling community at different laboratories and in different countries have led to the formalization of a component architecture approach to organize the structure. In this approach there are base level components and composited components which rely on the results of the base level components. The interface between the components is formalized by two-way couplers, which transfer the needed information. The model as a whole is a composite of composited, coupled components. Figure 3 shows the Goddard Earth Observing System, version 5 (GEOS-5) component architecture as expressed in the Earth System Modeling Framework (Hill et al. 2004; http://www.esmf.ucar.edu/).

Referring to Fig. 3, the box labelled "agcm" represents the atmospheric general circulation model. The components represented here are appropriate for climate and weather. Additional components would be required to resolve the processes above the mesosphere; for example, to support space weather (Toth et al. 2005; see also chapter *Assimilation of GPS Soundings in Ionospheric Models*, Khattatov). Below "agcm" are two components which represent the fluid "dynamics" and the "physics." The fluid dynamical part of the model represents both the resolved flow and the drag associated with small (subgrid) scale gravity waves. The dynamics will

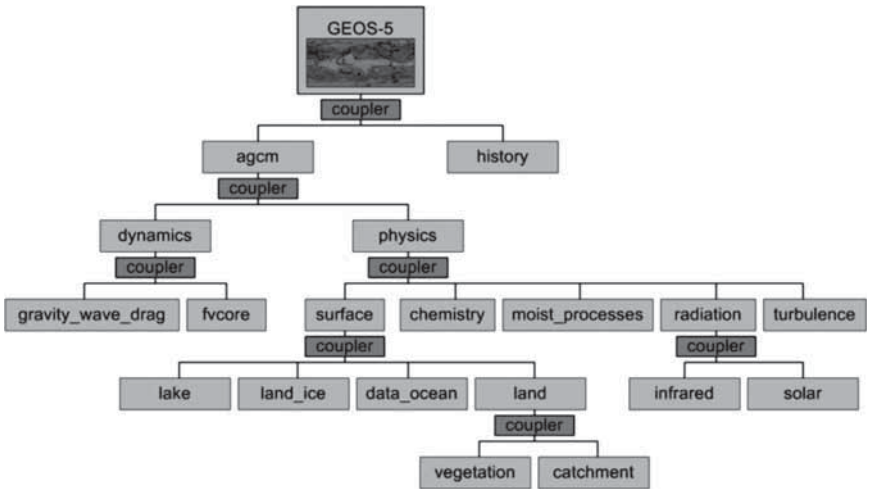



**Fig. 3** Earth System Modeling Framework (ESMF) component architecture of the Goddard Earth Observing System, version 5 (GEOS-5) atmospheric model (http://www.esmf.ucar.edu/about_us/). See text for detailed discussion

be discussed more fully below. The terms that form the components of the physics generally represent processes that occur on a scale smaller than resolved; again, they are subgrid. These are often called "parametrizations" (see Fig. 1). A useful, approximate concept is that those components collected under the term physics are treated as occurring only in a vertical column; hence, they can be extracted and tested in one-dimensional column models.[1] Those terms in the "dynamics" are fully three-dimensional; they connect the columns.

From left to right those components which composite as "physics" are as follows. The "surface" box represents, for an atmospheric model, the boundary conditions. Different variables characterize the transfer of momentum, mass, and energy from lakes, ice, ocean, and land (chapters *Ocean Data Assimilation*, Haines; *Land Surface Data Assimilation*, Houser et al., discuss models of the ocean and land, respectively). In this particular model the "land" model is a composite of a vegetation model and a catchment basin hydrology model. The next box to the right, "chemistry," is the interface to chemical production and loss terms which take place point-by-point in both the horizontal and the vertical. This is followed by the accumulation of the processes associated with water and its phase changes, "moist process": clouds, water vapour, liquid water, ice, convection, etc. Next are those processes needed to represent the absorption and reflection of both solar and terrestrial (infrared) "radiation." On the far right is a box labelled as "turbulence". Usually, in atmospheric models there is a separate parametrization which represents the turbulent mixing associated with the planetary boundary layer. More than the other processes in the composite of "physics," the processes in the planetary boundary layer may be connected in the horizontal; that is, they might not fit appropriately into the concept of column physics. As described here, these parametrizations connect momentum, mass, and energy in the vertical; there is transfer between model levels.

Figure 3 is a specific description of an atmospheric model, which relies on the representative equations listed above. In Fig. 1, a conceptual description for building a model was proposed. There are some obvious links. The boundary conditions appear explicitly, and Fig. 3 provides a framework for splitting up the processes of the representative equations. It remains to develop a discrete representation of equations and the identification of the primary and derived products from the model.

As stated at the beginning of the chapter the technical aspects of numerical modelling are left to comprehensive texts such as Jacobson (2005). Some numerical concepts will be developed here to demonstrate the art of model building. The focus will be on the "dynamics" part of the model (see Fig. 3). To demonstrate the concepts consider the thermodynamic energy equation and only the advection of temperature by the horizontal winds

$$\frac{\partial T}{\partial t} + \mathbf{u} \bullet \nabla T = \frac{\partial T}{\partial t} + u\frac{\partial T}{\partial x} + v\frac{\partial T}{\partial y} \tag{3}$$

[1] See information on column models at the National Center for Atmospheric Research – http://www.ccsm.ucar.edu/models/atm-cam/docs/scam/

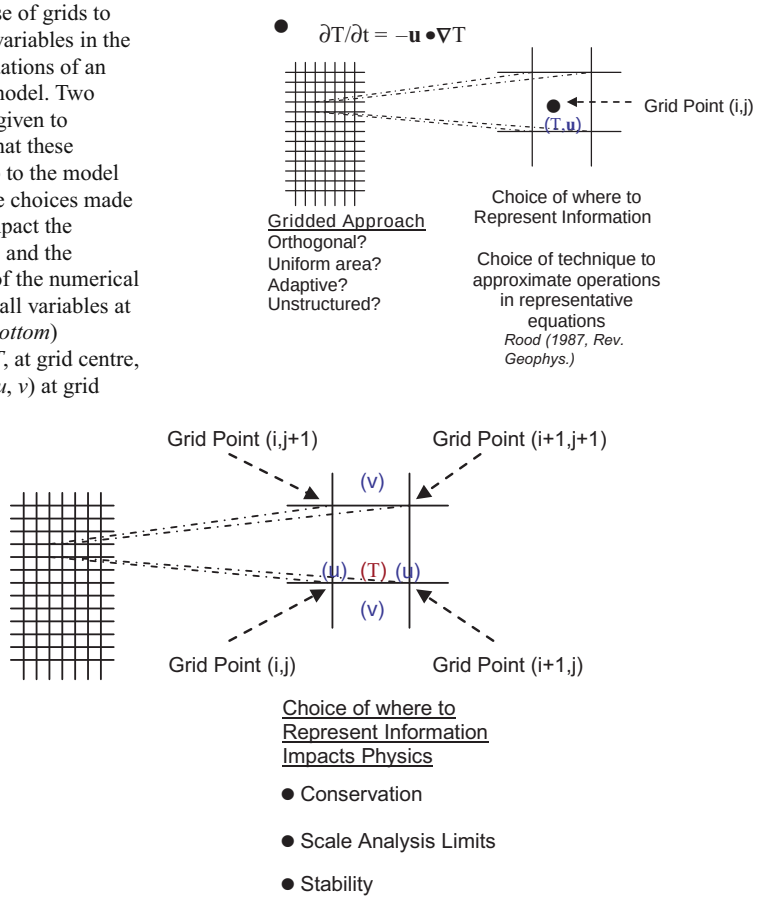



**Fig 4.** The use of grids to represent the variables in the governing equations of an atmospheric model. Two examples are given to demonstrate that these choices are up to the model developer. The choices made profoundly impact the characteristics and the performance of the numerical solution (*top*) all variables at grid centre, (*bottom*) temperature, *T*, at grid centre, and velocity (*u*, *v*) at grid edges

Attention will be focused on strategies to discretize the advective transport. Figure 4 illustrates the basic concepts. On the left of the figure a mesh has been laid down to cover the spatial domain of interest. In this case it is a rectangular mesh. The mesh does not have to be rectangular, uniform, or orthogonal. In fact the mesh can be unstructured or can be built to adapt to the features that are being modelled. The choice of the mesh is determined by the modeller and depends upon the diagnostic and prognostic applications of the model (see Randall 2000). The choice of mesh can also be determined by the computational advantages that might be realized.[2]

[2]Typical mesh sizes at the time of this chapter are 200 km for climate models down to 20 km for global weather models. Experiments are being run at resolutions as small as ~1 km. Computational resources limit resolution, but also as the resolution becomes finer the foundational assumptions of physical parametrizations must be reconsidered.

Using the mesh, index points are prescribed to determine location. In Fig. 4 (top) both the advective velocity and the temperature are prescribed at the centre of the cell. In Fig. 4 (bottom), the velocities are prescribed at the middle of the cell edges, and the temperature is prescribed in the centre of the cell. There are no hard and fast rules about where the parameters are prescribed, but small differences in their prescription can have large impact on the quality of the estimated solution to the equation, i.e., the simulation. The prescription directly impacts the ability of the model to represent conservation properties and to provide the link between the analytic equations and the theoretical constraints (see Fig. 1; see Rood 1987; Lin and Rood 1996, 1997; Lin 2004). In addition, the prescription is strongly related to the stability of the numerical method; that is, the ability to represent any credible estimate at all.

A traditional and intuitive approach to discretization is to use differences calculated across the expanse of the grid cell to estimate partial derivatives. This is the foundation of the *finite-difference method*, and finite-differences appear in one form or another in various components of most models. Differences can be calculated from a stencil that covers a single cell or weighted values from neighbouring cells can be used. From a numerical point of view, the larger the stencil, the more cells that are used, the more accurate the approximation of the derivative. *Spectral methods*, which use orthogonal expansion functions to estimate the derivatives, essentially use information from the entire domain. While the use of a large stencil increases the accuracy of the estimate of the partial derivatives, it also increases the computational cost and means that discretization errors are correlated across large portions of the domain.

One approach to solving the model equations is to take the continuous representative equations and make term-by-term numerical approximations to variables and their derivatives. There are many approaches to discretization of the dynamical equations that govern geophysical processes (Randall 2000; Jacobson 2005). Given that these equations are, in essence, shared by many scientific disciplines, there are sophisticated and sometimes similar developments in many different fields. One approach that has been recently adopted by several modelling centres is described in Lin (2004). In this approach the cells are treated as *finite volumes* and piecewise continuous functions are fit locally to the cells. These piecewise continuous functions are then integrated around the volume to yield the forces acting on the volume. This method, which was derived with physical consistency as a requirement for the scheme, has proven to have numerous scientific advantages. The scheme uses the philosophy that if the correlated physics are represented, then the accuracy of the scheme can be robustly built on a physical foundation. In addition, the scheme, which is built around local stencils, has numerous computational advantages.

The variables, *u, v, T,* and $Q_{H2O}$ are often termed the resolved or prognostic variables. Models are often cast into the form that surface pressure, $p_{sfc}$, is the prognostic equation for conservation of mass. These variables and their gradients are explicitly represented on the grid. The hydrostatic balance is a strong balance in the atmosphere. Most current global models are hydrostatic and do not include a prognostic equation for the vertical velocity, *w*; it is a diagnostic quantity. Cloud resolving

models and non-hydrostatic models do resolve the vertical velocity. Non-hydrostatic effects need to be considered if the horizontal resolution is finer than, approximately, 10 km. The importance of the consistent representation of the vertical velocity will be discussed more fully later in the chapter. Reactive constituents and aerosols can add to the list of resolved or prognostic variables. Generally, when the term prognostic is used to describe a variable, it means that a conservation equation has been written for that variable.

In contrast to the resolved or prognostic variables, there are a set of variables which are diagnosed at the grid scale. An example of this is the cloud mass flux between vertical layers of the atmosphere associated with the updrafts and downdrafts of cumulus clouds. There are some variables such as cloud liquid water and cloud ice which may be either explicitly resolved or diagnosed. This is dependent on the spatial resolution of the model. If the model has a resolution that is much larger that the scale of clouds, then cloud water and cloud ice have lifetimes too short to be advected from one grid box to another. In this case, these quantities are diagnosed in the column physics. The terminology prognostic and diagnostic are not precise; they are jargon. Many of the diagnostic variables are, in fact, predicted; therefore, they have the time-change attribute associated with the term "prognostic."

There is also a set of derived products associated with the model (see Fig. 1). For example, it is often productive to interpret atmospheric motions in terms of *vorticity* ($\nabla \times \mathbf{u}$) and *divergence* ($\nabla \bullet \mathbf{u}$). For large-scale, middle latitude dynamics, using pressure as the vertical coordinate, the relative vorticity, $\zeta$, is related to the geopotential, $\Phi$, by the following relationship

$$\zeta = \frac{1}{f}\nabla^2\Phi \tag{4}$$

$f$ is the Coriolis parameter. Geopotential, $\Phi$, is defined as $\Phi(z) = \int_0^z g dz'$,and is the variable which represents the height of a pressure surface when pressure, instead of height, is used as the vertical coordinate. (Geopotential can be related to a parameter with height dimension by dividing it by $g$; this is termed geopotential height.) The ability of the discretization method and the numerical technique to represent relationships such as the one described above is an important and underappreciated aspect of model construction. Lin and Rood (1997) show explicitly both a configuration of variables on the grid and a specification of averaging techniques that assures that the relationship between geopotential and vorticity is maintained in the discrete equations.[3]

Returning to the grids of Fig. 4, the spatial scale of the grid is related to the smallest scales which can be resolved in the model. As guidance, it takes a minimum of 8–10 grid boxes to resolve a wave meaningfully. There is *transport* and *mixing* which occurs at smaller spatial scales. Therefore, for both physical and numerical reasons there is the need to specify a subgrid mixing algorithm. In addition, explicit

[3]This constraint, therefore, implicitly links the numerical scheme to large-scale, rotationally dominated flows. As resolution is increased, the divergent component of the flow becomes larger. Therefore, different numerical considerations are expected to be required.

filters are used to counter errors that arise because of the discrete representation of continuous fields. Subgrid mixing and filters often take on the characteristics of *diffusion*. Their role in atmospheric models is complex and not well quantified. For instance, filters have been used to remove gravity waves in weather forecasting models (see chapter *Initialization*, Lynch and Huang). Given the important role of gravity wave dissipation in climate models, such a filter minimally complicates the quantitative representation of mixing in the atmosphere.[4] There are similar complications associated with the boundary layer turbulence parametrization. It is important to recognize that the dynamical core of the atmospheric model includes not only an approximation to the resolved advection, but also an algorithm for subgrid mixing, and filters to remedy numerical errors. All these pieces are tightly related to each other; they are often conflated.

As is apparent from the discussion above, there is not a unique or defined way to build an atmospheric model. With the benefit of many years of experience, there are a set of practices which are often followed. These practices evolve as experience is gained. There are decisions in model building which balance known sources of errors. There are decisions simply to give the model viable computational attributes. In many modelling environments there are parts of the code, components, which have not been altered in many years. There remain many open problems which need to be addressed and many paths proposed to address these problems. There are decisions in design and engineering, which contain more than a small element of art.

## 5 Consideration of the Observation-Model Interface

The interaction between the model and the observations takes place, ultimately, through the *statistical analysis* algorithm (see Fig. 2). There are many aspects of this interface which are described elsewhere in this book (e.g. see chapters in Part I, *Theory*; chapter *Constituent Assimilation*, Lahoz and Errera; and chapter *Land Surface Data Assimilation*, Houser et al.). The model and the observations are formally connected through the observation operator which can be as straightforward as interpolation routines or as complex as radiative transfer models which convert between the model variables and, for instance, the radiances that are observed by satellite instruments (see chapter *Assimilation of Operational Data*, Andersson and Thépaut). The model provides information directly to the quality control algorithm. Information from the analysis may be returned to the model through initialization algorithms which strive to filter out dynamical scales which are not important to the short-term forecast. The model and analysis interface can be a one-time passing of

[4]The initialization routine removes scales, for example, gravity waves that are detrimental to the short-term forecast. This is, in part, due to the fact that these scales are not well represented in either the model or the observations. Plus there are spurious sources of small scales related to imbalances due to scale errors and random error. It is incorrect to state that waves at these scales are not important to the atmosphere. They are important to both weather and climate. The behaviour of motions at these scales changes as the resolution of the model changes.

information, or there are numerous strategies for cycling the information across the interface to improve the balance in the model. Four-dimensional variational techniques and the incremental analysis update (Bloom et al. 1996) are examples of such cycling.

This section focuses on those geophysical variables that serve as the interface variables and where these variables are updated in the component architecture of the model (Fig. 3).

In order for assimilation to be a robust approach to analysing observations, there are a number of attributes that need to be considered. For example, are there enough data to specify the state variables of the atmosphere? If there are not enough observations to specify the state, then the model predictions are not likely to be a real source of information. Alternatively, if there are so many observations that the model is essentially specified by the observations, then a model is not needed for the analysis. The model must be of sufficient quality that it can propagate information from one observing time to the next. The observed variable must have a time-scale, a lifetime, such that information lasts from one observing time to the next; that is, there is the possibility of deterministic prediction. For the assimilation to be robust both the model and observations must contribute to the analysis; the error characteristics from one source or another should not always dominate the other.

For the atmosphere the geophysical parameter with, by far, the most complete coverage is temperature (see chapter *The Global Observing System*, Thépaut and Andersson). Since World War II there has been adequate coverage from surface measurements and balloons to support forecast-assimilation systems. With temperature observations it is possible to make credible estimates of winds by both the transference of information through the equations of motion and the propagation of information to chronically under-observed or unobserved regions. There is, also, a substantial body of horizontal wind observations and water vapour observations in the troposphere. Wind observations are especially important to the definition of the atmospheric state.

The temperature and wind observations are both primary products of the model; they are prognostic variables (see Fig. 1). Their spatial and temporal scales are such that their information is utilized and propagated by the model, especially in middle latitudes. From Fig. 3, these variables are directly provided by the dynamical core. The physics of the atmosphere are such that temperature and wind reflect integrated information. Temperatures and winds from assimilated data systems are often excellent estimates of the true state in the free troposphere and lower stratosphere.

Though there is a conservation equation for water vapour and water is a primary, prognostic variable, the water vapour distribution is largely determined in the "moist processes" component of the "physics" (Fig. 3). Because of phase changes, the spatial and temporal time-scales are both small. The observations of water which come from weather balloons reflect the small spatial scales of water in the atmosphere. These scales are far smaller than those represented by the model grid. The sampling network is not always representative of the state as a whole. The error characteristics are also a strong function of environment, i.e., temperature. Therefore, the representation of water from assimilated data sets is often not of quality for geophysical use (see chapter *Constituent Assimilation*, Lahoz and Errera).

One challenge that must be faced when using the observations from, for instance, balloons, is mapping the observational characteristics to the model. The model might be assumed to represent, for instance, the average temperature in a grid box. The balloon measurement might be appropriately considered a point measurement, at least relative to the model grid. Again, there is a more straightforward relation between modelled and observed temperatures and winds than modelled and observed water vapour.

Since 1979 satellite observations have come to dominate the absolute number of observations used in assimilation systems (see chapter *The Global Observing System*, Thépaut and Andersson). The first variables observed by satellites that were useful for assimilation are temperature and ozone observations. As compared with balloon measurements, satellite information is, often, smeared out over several model grid cells. It took many years to learn how to use satellite temperature observations effectively, and it was determined that mapping of the model information to the radiance space observed by the satellite often facilitated the use of observed information.

Ozone is an interesting variable to examine the model-observation-analysis interface. In some parts of the atmosphere the ozone distribution is determined by advection. Hence, the primary model information would come from the "dynamics" component (Fig. 3) in these regions. Given quality representation of the winds, ozone assimilation works well in these regions (see also chapter *Constituent Assimilation*, Lahoz and Errera). Other parts of the ozone distribution are primarily determined by processes contained in the "chemistry" component (Fig. 3). There are strong spatial gradients in the chemistry; in some places the time-scales are very short. Further, there are strong interdependencies with other gases, aerosols, and temperature (see chapter *Introduction to Atmospheric Chemistry and Constituent Transport*, Yudin and Khattatov). In these regions the assimilation is dominated by the chemical sources and sinks and the advection of ozone from other places and other times has little impact.

The examples outlined above highlight both the ability of the observing system to define the atmospheric state and the ability of the model to use the information. Experience to date shows that if the model information comes from the "dynamics" (Fig. 3) and the spatial gradients are resolved with some accuracy, then the assimilation can be robust. Alternatively, if the model information comes from the "physics" (Fig. 3) and the spatial and temporal scales are small, then the assimilation is likely to have large errors and be of little geophysical value.

Since 1979, and especially since the early 1990s, the amount, the quality, and the span of satellite observations have all grown tremendously. There are many geophysical parameters being measured in the atmosphere, and on the land, ocean, and ice surface. Some of the data, for instance, ocean surface winds have proven to have large impact in assimilation systems. Other observations have proven more difficult to use. The same ideas as described for atmospheric assimilation hold; the observing system must be able to define the state, and the model able to use the observations. Many of these new observations would have their interface with the model through the "physics" component. The spatial and temporal scales of the observations as

compared to their representation within the model are often not compatible. The composites that make up the variables in the model are often not what the satellite is observing. The greatest challenges in modelling lie in the representation of "physics," and one of the primary development paths for model-data assimilation should be the development of the model variable–observed variable interface.

## 6 Physical Consistency and Data Assimilation

Data assimilation has had dramatic impacts on the improvement of weather forecasts (see chapter *Assimilation of Operational Data*, Andersson and Thépaut). There has been wide-scale use of assimilated data sets in climate applications with some great success, as well as identification of a set of problems for which the assimilation analyses are not up to the application. Problems that rely on correlated behaviour of geophysical parameters that are not directly measured, i.e., those estimated by the model parametrizations, are difficult to address. Two examples of such problems, *hydrometeorology* and *constituent transport*, are discussed in chapter *Reanalysis: Data Assimilation for Scientific Investigation of Climate* (Rood and Bosilovich). The rest of this chapter will explore the attributes that distinguish data assimilation for weather from data assimilation for climate.

Weather forecasting is first and foremost concerned with providing quantitative estimates of a set of key variables which define local atmospheric conditions. Successful forecasts benefit from the atmosphere being organized into dynamical structures, e.g. waves and vortices, whose propagation is well represented by both the equations of the motion and the numerical methods used to approximate their solution. The observation system can resolve the actual configuration of the dynamical structures at a particular time. In the midst of the forecast-assimilation cycle, the model also has a configuration of the dynamical structures. Through well-defined interfaces, the assimilation routine hands a scale-dependent, balanced initial state to the predictive model, which, in principle, corrects the observed scales in the model state. The model propagates these features forward in time. As is well established, in mid latitudes, during winter, predictions of temperature and wind are useful for several days. When precipitation is associated with large scale dynamics, the prediction of precipitation is useful. Whether that precipitation will be rain or snow is a more difficult prediction. In the tropics and in the summer, when the dynamical features are of smaller scale and the localized moist processes are important to organization of the features, the length of the useful forecast is shorter. Van den Dool et al. (1990) discuss measures of forecast skill for high, medium and low temporal variations in the atmosphere, including the tropics and the extra-tropics; Waliser et al. (1999) discuss the predictability of tropical phenomena and the relationship to their time-scale.

"Weather" is a subset of the dynamical features that make up the Earth's climate. The role of weather is to transport energy, and consequently, water and other constituents. A good weather forecast is characterized by a good forecast of wind velocity, which is essentially the flux of momentum. Since the ultimate impact of

weather on climate is a transport process, climate is directly related to the divergence of fluxes.

The atmosphere, especially at mid latitudes, is dominated by rotational flows close to geostrophic and hydrostatic balance (see Holton 2004). The divergent component of the flow, that responsible for irreversible transport, is an order of magnitude smaller that the rotational part of the flow. Alternatively, the part of the flow most important to the quality of a weather forecast, the rotational part, is an order of magnitude larger than the part of the flow important for a physically consistent representation for climate, the divergent part.[5] While the fluxes are directly related to resolved variables such as the horizontal wind, the divergence of the fluxes are related to transience, non-linearity, and, ultimately, the dissipation of dynamical systems (see Andrews and McIntyre 1978).

A metric of physical consistency is whether or not the conservation equation balances. That is, when all advection, production and loss terms are accounted for, is an accurate estimate of the time rate of change of a quantity realized? If the numerical methods of a model are formulated from a foundation of physical consistency and numerical accuracy, then for a free-running model the budgets should balance. This is not achieved without attention to the details. The effects of corrective filters must be accounted for, and if the filters are a significant part of the conservation equation, then the numerical scheme must be reconsidered.

There is no reason to expect that the disparate observations of the Earth system will satisfy a conservation equation. Therefore, when the observation-based corrections are added to the model equations, imbalance is added. The observations act like a complicated source-sink term. Whether the model and observations in some time averaged sense satisfy a geophysical conservation equation depends upon many things. If there is bias between the model and the observations then the data insertion, the analysis increments, will be a forcing term of constant sign. If there is bias between the models and the observations, then that suggests that the model predicts a different mean state than is observed. If the biases in the resolved variables are "corrected" in the assimilation process, then there is an inconsistency between the values of those "corrected" resolved variables and the values that the physical parametrizations of the model generate. This inconsistency might have little or no effect on the short-term forecast; however, the data insertion is constantly forcing this imbalance, and it will impact those circulations induced by dissipating waves that are important to the climate (see Hoskins et al. 1985; Holton et al. 1995; Holton 2004). Data insertion will impact where and how resolved scales are formed and dissipated.

In order to demonstrate the impact of data insertion more clearly and more quantitatively, two examples will be developed. The first is based on the assimilation of

[5] To be clear, it is the estimate of the divergent part of the wind by data assimilation that is responsible for much of the improvement of weather prediction. However, it is true that in many instances that a reasonable short-term forecast at middle latitudes can be realized by the barotropic vorticity equation; hence, the non-divergent geostrophic wind. A good weather forecast might be viewed as, "how fast is the wind blowing in my face?" This is the flux of momentum.

temperature corrections into the thermodynamic equation. The second is based on the analysis of the vertical velocity in the transformed-Eulerian mean formulation of the equations of motion (see Holton et al. 1995; Holton 2004).

## Example 1: Observational Correction to the Thermodynamic Equation

To demonstrate the problem of physical consistency more quantitatively, consider the thermodynamic equation written with a simple heating rate and a loss rate proportional to temperature.

$$\frac{DT_f}{Dt} = \frac{\partial T_f}{\partial t} + \mathbf{u} \bullet \nabla T_f = H - \lambda T_f \tag{5}$$

$T_f$ is written to explicitly represent that this is the model forecast temperature. Two cases will be examined.

*Example 1, Case 1*: In Case 1 assume that the assimilation acts as a forcing term which relaxes the modelled temperature to an analysed temperature determined from the observations. This analysed temperature, for example, might be a gridded, least squares estimate from a variety of observations. The subscript *a* represents the analysis of the observational information.

$$\frac{DT_f}{Dt} = H - \lambda T_f - \lambda_a (T_f - T_a) \tag{6}$$

Note that the time-scale, $1/\lambda$ associated with the original equation follows from physical principles. The parameter $1/\lambda_a$ represents the relaxation time-scale associated with the analysis. The time-scale from the analysis follows from the design and assumptions of the data assimilation system (see Swinbank and O'Neill 1994). This appears as an additional forcing term; in the construct of Fig. 3, a "physics" term that is not in the model equations. Therefore, the estimated solution of the equation for $T_f$ evolves in the presence of this additional forcing term.

The equation can be rearranged as

$$\frac{DT_f}{Dt} = H + \lambda_a T_a - (\lambda + \lambda_a) T_f \tag{7}$$

The analysis can be viewed as a change to the loss rate. If the observations are biased relative to the forecast, then the observations are in the time average, a heating term. If the observations are unbiased in the time average, then this heating term averages away; still however, the loss rate is changed. The conservation equation is altered.

*Example 1, Case 2:* Case 2 implements the data-driven correction to the model equation by correction of the prognostic variable. That is, $T_f$ is replaced with a corrected temperature which is $T_f + \delta T_a$. On substitution into Eq. (5) and re-arranging the terms:

$$\frac{\partial T_f}{\partial t} + \mathbf{u} \bullet \nabla T_f - H + \lambda T_f = -(\frac{\partial(\delta T_a)}{\partial t} + \mathbf{u} \bullet \nabla(\delta T_a) + \lambda(\delta T_a)) \tag{8}$$

The terms on the left side are the model equations, which balance to zero in a free-running simulation. The terms on the right side represent an additional forcing to the model associated with the data insertion.

Under the condition that the model equation is satisfied, the left side of Eq. (8) is zero, then the following equation is obtained.

$$\frac{D(\delta T_a)}{Dt} = -\lambda(\delta T_a) \tag{9}$$

In this case, the increment from each data insertion relaxes to zero with time.

There are two intuitive time-scales to compare with the cooling time-scale $1/\lambda$. The first is the advective time-scale, which is a comparison of physically derived time-scales. This would divide the atmosphere into regions that are dominated by the "physics" and the "dynamics" as well as transition regions where both are important. The second time-scale is the time-scale associated with the frequency of insertion of observational information. From the point of view of "correcting" the model, the balance of these time-scales provides a mechanism to understand the system performance. In this case, following a parcel, the data insertion is a succession of corrections. These corrections relax to zero; hence, the state desired by the model. However, as the system is relaxing, the slow time-scales in the model are being constantly impacted by the observations. This impact lasts far longer than the direct impact of a single observation on the analysis increment. If there is bias between the model and the observations, then this represents a forcing to a new equilibrium state. The data insertion is, also, a source of variability at the time-scales of the observing system.

Another way to think about the role of the model in the assimilation system is to imagine the model as an instrument "observing" a suite of observations that describe the atmosphere, or more generally, the Earth. Some parts of the model directly observe this suite of observations and are well specified. Other parts of the observation suite are indirectly observed, and there are some unobserved variables which are diagnosed as part of the model "circuitry." The influence that the observations have on these indirectly determined and unobserved variables is strongly dependent on the design of the model and how information flows through the model. It is strongly dependent on the logic and construction of the numerical parametrizations.

To illustrate this, consider the component architecture of Fig. 3. Since this figure represents a solution to a set of conservation equations, then the model can be viewed as residing within a closed system. The correction of the model by observed information makes this an open system; the tenets of the conservation principle are no longer true. For the climate, which is the accumulation of the processes represented in the system, this insertion of information (forcing) from outside the system must be assumed to have consequences on the basic physical processes.

Figure 5 is a reduced version of Fig. 3; the box around the components shows that the model is a closed system. The coupler determines the transfer of information

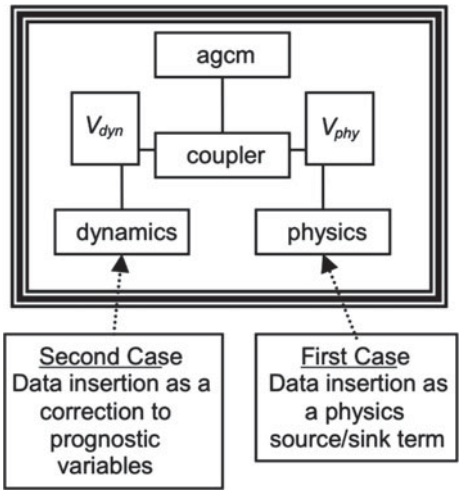



**Fig. 5** Schematic of model as a closed system, which accepts forcing from outside the system. Balance here is represented symbolically as if the model was an electrical circuit with a voltage difference across the "coupler." agcm stands for "atmospheric general circulation model"

between the "physics" and the "dynamics." There are numerous time and space scales present in the coupler, as well as those intrinsic to the "dynamics" and the "physics." If the model is viewed as an electronic instrument, as posed above, then there is a network of resistors which determine flow of signal through the system. There are capacitors that represent the time-scales of processes. The balance across the coupler is represented as a voltage difference, $V_{dyn}-V_{phy}$. The two data insertion scenarios described above are illustrated in Fig. 5. They both explicitly bring in information from outside of the system and to different sides of the coupler. They both would change some aspect of the system, represented symbolically by a change in the voltage difference across the coupler.

## Example 2: Horizontal Divergence and the Vertical Wind

The two cases in Example 1, above, used a simple form of the thermodynamic equation. The thermodynamic variables have the property of being in local equilibrium. However their relationship to the wind fields is not always local; the winds are related to the spatially integrated thermodynamic state. There is the possibility of action at a distance as momentum that is dissipated in one region can induce circulations which impact distant regions (see Hoskins et al. 1985; Holton et al. 1995; Holton 2004). Therefore, errors in the wind field are expected to impact the analysis in a different way than errors in the temperature field.

As pointed out above, for many years atmospheric observations were dominated by temperature measurements. Winds were estimated from the temperature observations. Assume that the horizontal winds are corrected by the temperature observations by transfer of information through the assimilation system.

$$u_c = u_f + \delta(u(\delta T_a)) \text{ and } v_c = v_f + \delta(v(\delta T_a)) \tag{10}$$

The subscript $c$ is the corrected forecast; subscript $f$ is the forecast. $u(\delta T_a)$ *and* $v(\delta T_a)$ are the corrections to the velocity field related to the correction in the temperature field that comes from the analysed observations. Through the mass continuity equation, the divergence of the horizontal wind is related to the vertical gradient of the vertical velocity. The divergence of the horizontal wind field, therefore, is the primary quantity that connects the thermodynamic equation and the momentum equations in the atmosphere (see Eq. 1 and Holton 2004). Schematically, the vertical velocity is a key variable connecting the "dynamics" and "physics" components of the model (see Fig. 3). Hence the vertical velocity is a key variable in the coupling of the "dynamics" and the "physics."

As an example, consider large-scale, mid latitude dynamical features. Scale analyses in the atmosphere shows that for these dynamical systems the divergence of the horizontal wind is an order of magnitude smaller than either of the individual terms that make up the divergence. That is, for a representative velocity scale $U$ and length scale $L$

$$\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} \text{ scales as } 0.1\frac{U}{L} \tag{11}$$

The divergence of the assimilation-corrected horizontal wind is

$$\frac{\partial u_c}{\partial x} + \frac{\partial v_c}{\partial y} = \frac{\partial u_f}{\partial x} + \frac{\partial v_f}{\partial y} + \frac{\partial(\delta(u(\delta T)))}{\partial x} + \frac{\partial(\delta(v(\delta T)))}{\partial y} \tag{12}$$

A 10% "correction" in the wind is, potentially, a 100% error in the divergence. It follows that there are similarly large errors in the vertical velocity.

As stated in the previous section, the vertical velocity is usually diagnosed in global models. The vertical velocity can, in general, be diagnosed in two ways. Following Holton (2004), in pressure coordinates, where $\omega \equiv \frac{Dp}{Dt}$ is the vertical velocity,

$$\omega_k(p) = \omega_k(p_{sfc}) - \int_{p_{sfc}}^{p} (\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y})_p dp \tag{13}$$

The subscript $k$ indicates that this velocity is diagnosed from the kinematics of the flow field. $p_{\text{sfc}}$ is the surface pressure.

The vertical velocity can also be diagnosed from the thermodynamic equation. Again, in pressure coordinates and following Holton (2004), assuming the diabatic terms, $J$, can be ignored,

$$\omega_T(p) = S_p^{-1}\left(\frac{\partial T}{\partial t} + u\frac{\partial T}{\partial x} + v\frac{\partial T}{\partial y}\right) \tag{14}$$

Where $S_p$ is the static stability parameter in pressure coordinates. The subscript $T$ indicates this estimate of the vertical velocity is from the thermodynamic equation. In this simplified case the consistency question would ask whether or not these two estimates of the vertical velocity are equal. Experience shows that this is not the case in assimilated data sets, and the errors in the divergence dominate the calculation in $\omega_k$.

To illustrate this problem further, and to make the connection to the climate problem clearer, it is useful to consider the transformed Eulerian-mean formulation of the equations of motion (see Holton et al. 1995; Holton 2004). This formulation has proven a powerful paradigm for understanding the general circulation and constituent transport and is an important example of the physical constraints discussed in Fig. 1. The transformed Eulerian-mean approximates the compensating transport of the waves (represented by a prime) and the Eulerian-mean meridional circulation (represented by an over bar). In this case the diabatic terms cannot be ignored, and one estimate of the residual mean vertical velocity,$\overline{w}^*$, is called the diabatic (subscript $d$) vertical velocity and should equal

$$\overline{w}^*_d(z) = \frac{R\bar{J}}{HN^2 c_p} \tag{15}$$

For convenience, the vertical coordinate, $z$, is log pressure-height. $N^2$ is the square of the buoyancy frequency, and $H$ is a constant scale height ~7 km.

By definition the corresponding analogue to the kinematic estimate is

$$\overline{w}^*_k(z) = \overline{w} + \frac{R}{H}\frac{\partial(v'T'/N^2)}{\partial y} \tag{16}$$

In this case the question of consistency comes to whether or not $\overline{w}^*_k = \overline{w}^*_d$ is true.

In general this equality is not realized from assimilated data sets, even in the relatively simple case of the stratosphere (see Schoeberl et al. 2003).

Finally, this form of the exposition of the concepts of physical consistency is illustrated in Fig. 6. The value of the vertical velocity presented to the coupler should be the same from the diagnostics via the "physics" and "dynamics." If this is not the case, then the assimilation is physically inconsistent. This particular exposition through the vertical velocity is perhaps the most relevant and important in data assimilation. It is relevant not only to climate and chemistry transport, but to weather. It poses a physical constraint for assimilation – can physically consistent thermodynamic and kinematic vertical velocities from the model be maintained in the assimilation? Or more generally – can the physical balance of the model be maintained in the presence of the assimilation of observations? This is a formidable task.

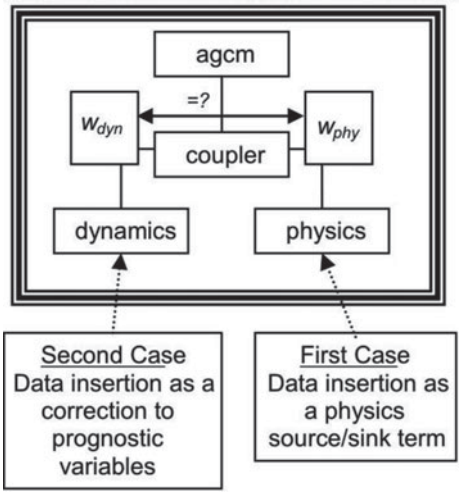



**Fig. 6** Schematic of model as a closed system, which accepts forcing from outside the system. Balance here is represented as consistency between vertical velocity estimates coming from the "physics" or "dynamics" components. agcm stands for "atmospheric general circulation model"

## 7 Summary

This chapter introduced the fundamental ideas that a scientist needs to understand when building or using models in Earth system science research. Rather than focusing on technical aspects of modelling and data assimilation, the chapter focused on a number of underlying principles. These principles, if adhered to, will allow the models and model products to be used in quantitative, data-driven research.

With regards to stand-alone models in the absence of data assimilation, it was emphasized that the underlying physics should be well represented. Specifically, the need to represent correlated behaviour between geophysical parameters was emphasized. A strategy for meeting such a design criteria is to assure that the discrete, numerical approximation to the continuous equations honours the balance conditions that are used in the development of theoretical constructs. This emphasizes "consistency," perhaps at the expense of formal numerical accuracy, as accurate numerical techniques do not guarantee physical consistency. Data assimilation was introduced as the addition of a forcing term to the model that is a correction based on observations. This additional forcing term changes the balance of forces. Therefore, budgets calculated from assimilated data are not expected, a priori, to be robust for geophysical applications.

The role of the model in data assimilation was discussed. It is the assimilation of observational information into the predictive-diagnostic model that sits at the foundation of the value and the potential value of the information produced by data assimilation. In applications ranging from mapping, to improved predictions, to generation of unobserved geophysical variables, data assimilation stands as an essential ingredient of modern Earth system science. The future development of data assimilation includes both the improvement of the models and the better use of information

provided by the model. Model improvements include a more robust link in the models between resolved scales and subgrid physical parametrizations. Specifically, with regard to the link to data assimilation, the interface between the subgrid information and the observations needs more attention (see Zhang and Lin 1997). Better use of model information includes using the information developed by the model that connects the correlative physics important to the climate – how is this, first, preserved, then improved, when data is inserted into the model?

Several frames of reference were offered for thinking about models, model construction, and physical consistency. A summary version of these concepts follows. There are many time-scales represented by the representative equations of the model. Some of these time-scales represent balances that are achieved almost instantly between different variables. Other time scales are long, important to, for instance, the general circulation which will determine the distribution of long-lived trace constituents. It is possible in assimilation to produce a very accurate representation of the observed state variables and those variables which are balanced on fast time scales. On the other hand, improved estimates in the state variables are found, at least sometimes, to be associated with degraded estimates of those features determined by long time-scales. Conceptually, this can be thought of as the impact of bias propagating through the physical model (see Dee 2005). With the assumption that the observations are fundamentally accurate, this indicates errors in the specification of the physics that demand further research. The identification, the management, the correction, and the elimination of sources of bias are crucial for improving the physical robustness and self-consistency of assimilated data sets.

**Acknowledgments** I thank Minghua Zhang and Ivanka Štajner for reviewing this chapter.

## References

Andrews, D.G. and M.E. McIntyre, 1978. Generalized Eliassen-Palm and Charney-Drazin theorems for waves on axisymmetric mean flows in compressible atmospheres. *J. Atmos. Sci.*, **35**, 175–185.

Bloom, S.C., L.L. Takacs, A.M. da Silva and D. Ledvina, 1996. Data assimilation using incremental analysis updates. *Mon. Weather Rev.*, **124**, 1256–1271.

Daley, R., 1991. *Atmospheric Data Analysis*, Cambridge University Press, New York, 457pp.

Dee, D.P., 2005. Bias and data assimilation. *Q. J. R. Meteorol. Soc.*, **131**, 3323–3342.

Dee, D.P., L. Rukhovets, R. Todling, A.M. da Silva and J.W. Larson, 2001. An adaptive buddy check for observational quality control. *Q. J. R. Meteorol. Soc.*, **127**, 2451–2471.

Elson, L.S., 1986. Ageostrophic motions in the stratosphere from satellite observations. *J. Atmos. Sci.*, **43**, 409–418.

Hill, C., C. DeLuca, V. Balaji, et al., 2004. Architecture of the Earth System Modeling Framework. *Comp. Sci. Eng.*, **6**, 18–28.

Holton, J.R., 2004. *An Introduction to Dynamic Meteorology*, Elsevier Academic Press, San Diego, 535pp.

Holton, J.R., P.H. Haynes, M.E. McIntyre, et al., 1995. Stratosphere-troposphere exchange. *Rev. Geophys.*, **33**, 403–439.

Hoskins, B.J., M.E. McIntyre and A.W. Robertson, 1985. On the use and significance of isentropic potential vorticity maps. *Q. J. R. Meteorol. Soc.*, **111**, 877–946 (Correction, *Q. J. R. Meteorol. Soc.*, **113**, 402–404, 1987).

Jacobson, M.Z., 2005. *Fundamentals of Atmospheric Modeling*, 2nd edition, Cambridge University Press, New York, 813pp.

Johnson, S.D., D.S. Battisti and E.S. Sarachik, 2000. Empirically derived Markov models and prediction of tropical Pacific sea surface temperature anomalies. *J. Climate*, **13**, 3–17.

Kalnay, E., 2003. *Atmospheric Modeling, Data Assimilation, and Predictability*, Cambridge University Press, Cambridge, 364pp.

Lin, S.-J., 2004. A "vertically Lagrangian" finite-volume dynamical core for global models. *Mon. Weather Rev.*, **132**, 2293–2307.

Lin, S.-J. and R.B. Rood, 1996. Multidimensional flux-form semi-Lagrangian transport schemes. *Mon. Weather Rev.*, **124**, 2046–2070.

Lin, S.-J. and R.B. Rood, 1997. An explicit flux-form semi-Lagrangian shallow-water model on the sphere. *Q. J. R. Meteorol. Soc.*, **123**, 2477–2498.

Lynch, P., 2003. Introduction to initialization. In *Data Assimilation for the Earth System*, NATO Science Series: IV. Earth and Environmental Sciences 26, Swinbank, R., V. Shutyaev and W. Lahoz (eds.), Kluwer Academic Publishers, Dordrecht, The Netherlands, pp 97–111, 378pp.

McCrea, G.J., W.R. Gooden and J.H. Seinfeld, 1982. Numerical solution of the atmospheric diffusion equation for chemically reacting flows. *J. Comput. Phys.*, **45**, 1–42.

Molod, A., H.M. Helfand and L.L. Takacs, 1996. The climatology of parameterized physical processes in the GEOS-1 GCM and their impact on the GEOS-1 data assimilation system. *J. Climate*, **9**, 764–785.

Plumb, R.A. and M.K.W. Ko, 1992. Interrelationships between mixing ratios of long lived stratospheric constituents. *J. Geophys. Res.*, **97**, 10145–10156.

Randall, D.A. (ed.), 2000. *General Circulation Model Development: Past, Present, and Future*, Academic Press, San Diego CA, 807pp.

Rood, R.B., 1987. Numerical advection algorithms and their role in atmospheric transport and chemistry models. *Rev. Geophys.*, **25**, 71–100.

Schoeberl, M.R., A.R. Douglass, Z. Zhu and S. Pawson, 2003. A comparison of the lower stratospheric age-spectra derived from a general circulation model and two data assimilation systems. *J. Geophys. Res.*, **108**, 4113.

Štajner, I., K. Wargan, S. Pawson, et al. 2008. Assimilated ozone from EOS-Aura: Evaluation of the tropopause region and tropospheric columns. *J. Geophys. Res.*, **113**, D16S32, doi: 10.1029/2007JD008863.

Strang, G., 1968. On the construction and comparison of difference schemes. *SIAM J. Numer. Anal.*, **5**, 506–517.

Swinbank, R. and A. O'Neill, 1994. A stratosphere troposphere data assimilation system. *Mon. Weather Rev.* **122**, 686–702.

Swinbank, R., V. Shutyaev and W.A. Lahoz (eds.), 2003. *Data Assimilation for the Earth System*. NATO Science Series: IV. Earth and Environmental Sciences 26, Kluwer Academic Publishers, Dordrecht, The Netherlands, 378pp.

Toth, G., I.V. Sokolov, T.I. Gombosi, et al., 2005. Space weather modeling framework: A new tool for the space science community. *J. Geophys. Res.*, **110**, A12226, doi:10.1029/2005JA011126.

Trenberth, K.E. (ed.), 1992. *Climate System Modeling*, Cambridge University Press, Cambridge, 788pp.

Van den Dool, H.M. and S. Saha, 1990. Frequency dependence in forecast skill. *Mon. Weather Rev.*, **118**, 128–137.

Waliser, D.E., C. Jones, J.-K.E. Schemm and N.E. Graham, 1999. A statistical extended-range tropical forecast model based on the slow evolution of the Madden-Julian Oscillation. *J. Climate*, **12**, 1918–1939.

Yanenko, N.N., 1971. *The Method of Fractional Steps*, Spinger-Verlag, New York, 160pp.

Zhang, M.H. and J.L. Lin, 1997. Constrained variational analysis of sounding data based on column-integrated conservations of mass, heat, moisture, and momentum: Approach and application to ARM measurements. *J. Atmos. Sci.*, **54**, 1503–1524.

# Numerical Weather Prediction

**Richard Swinbank**

## 1 Introduction

Numerical weather prediction (NWP) entails the use of computer models of the atmosphere to simulate how the state of the atmosphere is likely to evolve over a period of several hours up to 1 or 2 weeks ahead. This approach is central to modern operational weather forecasting: it is the improvements in NWP systems that have led to continual improvements in the skill of weather forecasts over recent decades.

The essential steps in NWP are:

- Making observations of the current weather;
- Assimilating the observations into a numerical model, to represent the current atmospheric state;
- Integrating the model to simulate the future evolution of the atmosphere;
- Generating products to inform users about the forecast weather.

Many detailed aspects of this process are described elsewhere in this book (Part I, *Theory*; Part II, *Observations*; and companion chapters in Part III, *Meteorology and Atmospheric Dynamics*). The aim of this chapter is to describe how the processes fit together to produce operational weather forecasts.

In this chapter we will review the development of NWP techniques that have contributed to this major improvement in forecast skill, and mention some techniques that should lead to future improvements. We will describe the main stages in the NWP process, described in more detail elsewhere in the book, and see how they are applied in the context of operational weather forecasting. This, and other, chapters in the book outline the development of data assimilation techniques. Coupled with the increasingly sophisticated use of satellite observations, this has led to major improvements in the quality of initial data for weather forecasting. At the same time, increases in computer power have permitted major advances in numerical

R. Swinbank (✉)
Met Office, Exeter, UK
e-mail: richard.swinbank@metoffice.gov.uk

W. Lahoz et al. (eds.), *Data Assimilation*, DOI 10.1007/978-3-540-74703-1_15,

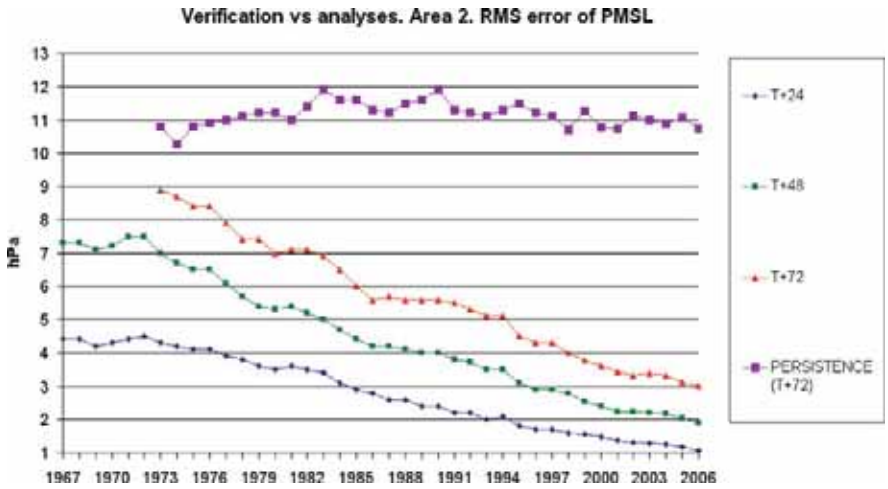



**Fig. 1** Root mean square (RMS) error in mean sea level pressure (hPa) over the North East Atlantic, illustrating how forecast errors have decreased over the past 40 years. Forecast errors are shown for 3 different forecast ranges: 24 (*blue*), 48 (*green*) and 72 (*blue*) h; for reference the T+72 persistence error is shown in magenta

weather prediction models, including more sophisticated physical parametrizations, numerical techniques and improved resolution. These factors have led to major improvements in forecast skill. Figure 1 illustrates how forecasts of mean sea level pressure have improved over the last 40 years (see also Fig. 4 in chapter *Assimilation of Operational Data*, Andersson and Thépaut). By this measure, the current skill of the 3-day forecast is the same as the skill of the 1-day forecast less than 25 years ago.

This chapter covers the use of NWP for a variety of applications, at different spatial scales and at different time ranges. To meet this range of requirements, national meteorological services often run a range on numerical models, from high resolution local models for very short-range prediction through to global models used for medium range forecasting (up to around 2 weeks) and beyond. To illustrate this chapter, we draw upon examples from the UK Met Office and other meteorological services.

## 2 Observations

### *2.1 Operational Observing System*

The primary input to NWP systems comes from regular measurements made by many meteorological instruments which are deployed worldwide to form a *Global Observing System*, GOS. Since the middle of the nineteenth century, a network of meteorological stations has been established to take regular measurements of atmospheric pressure, temperature, humidity, wind and other weather elements for use by

weather forecasters. Later, techniques were developed for making measurements of the upper levels of the atmosphere, using balloon-borne instruments. Radiosondes came to be a crucial part of the observing system following the International Geophysical Year (IGY) in 1957.

A further impetus to the development of the observing network came around the time of the First GARP (Global Atmospheric Research Programme) Global Experiment (1978/1979), when satellite observations started to become regularly used for operational weather forecasting. These observations included both cloud track winds and satellite temperature soundings. In addition to these observation types, the observing network also includes observations from aircraft, ships and drifting buoys. Satellite measurements include scatterometer measurements of surface winds, sea surface temperatures, and soundings of temperature and humidity derived from GPS (Global Positioning System) signals. Many of these observation types are described in more detail in chapter *The Global Observing System* (Thépaut and Andersson).

The dissemination of these observations is coordinated by the WMO (World Meteorological Organisation) WWW (World Weather Watch) programme. The data are exchanged between meteorological centres using the GTS (Global Telecommunication System). This ensures that meteorological data from all over the globe are exchanged in a timely basis, for use in operational weather forecasts.

In order to get the most up to date forecasts possible, operational weather forecasts need to be run very close to real time. So, each forecast is necessarily based on only those observations that are received within a few hours of when they were taken. This can be a particular constraint for satellite data, since the measurements may be relayed to ground stations only once an orbit. The ground systems to support operational weather satellites are designed to support this near real time requirement. It is also often helpful to process data from research satellites in order to make use of the novel types of measurements, and assess their benefit in an operational NWP framework (see chapter *Research Satellites*, Lahoz). For that reason, it is beneficial to arrange for a fast delivery stream for research satellite data.

Once the observations are received at operational NWP centres, they are collected in observation databases. At the start of an operational data assimilation run, all the observations from the relevant time window are extracted from the database and prepared for the data assimilation system.

## *2.2 Quality Control*

Each of the observations is subject to a variety of possible errors; there may be random measurement errors or systematic biases resulting from calibration errors. One also needs to account for the fact that observations at particular location may not be entirely representative of the surrounding area; these types of errors are usually referred to as *errors of representativeness*. Additionally, there may be serious errors resulting from instrument malfunction or transmission errors (for example); these are often referred to as *gross errors*. It is important that any observations suffering from gross errors are screened out and not used by the assimilation, otherwise they

could seriously degrade the quality of the weather forecast. This is the purpose of *quality control*, the subject of this subsection. In addition, any systematic biases should be removed; the bias correction of data is treated in detail in chapter *Bias Estimation* (Ménard).

Perhaps the most important part of pre-processing observations ready for assimilation is the quality control step. The aim is to detect all observations that have a high probability of suffering from gross errors. There are a number of checks that are part of this process, for example:

- Is the observation self consistent? For example, in a radiosonde sounding, is the vertical temperature structure physically plausible?
- Is the observation consistent with earlier measurements from the same observing station?
- Is an observation consistent with its neighbours? This "buddy check" can often highlight gross errors such as incorrect locations in ship reports;
- Is the observation consistent with a short-range forecast, often referred to as the background state? The background state reflects our a priori knowledge of the current state of the atmosphere.

One possible approach, as used at the Met Office, is to assess the "probability of gross error" (PGE) for each observation (Dharssi et al. 1992). Each observation is initially assigned a PGE value, dependant on the observation type. For example we expect about 1.5% of SYNOP pressure observations to be "bad" and therefore assign them an initial PGE of 0.015. The estimated PGE is then updated as each of the relevant checks is carried out. At the end of the process, an observation is rejected if the PGE is greater than 0.5.

In the Met Office system, the buddy check compares each observation against up to about 12 neighbouring observations, again updating the PGE. Surface observations are compared with other surface data. Radiosondes, aircraft and cloud-track winds are each compared with other observations of the same type, and radiosondes are compared with aircraft data. In general, observations with the same callsign are not allowed to buddy check each other, since they are likely measured by the same instrument, although this criterion is not applied to cloud-track winds. The buddy check compares differences from the background, rather than observed values themselves, e.g. two ships both 5 hPa lower than background will buddy check well, even if one is in the middle of a depression and the other has pressure 10 hPa higher 100 km away.

For the background check, the observations, which can be represented by the vector $\mathbf{y}$, are compared with the equivalent values derived from the background state, $\mathcal{H}(\mathbf{x}^b)$, where $\mathbf{x}^b$ denotes the background state and $\mathcal{H}$ the (non-linear) observation operator, which maps a model state to observation space. The innovation vector $\mathbf{y}-\mathcal{H}(\mathbf{x}^b)$ is therefore a measure of the departure of each observation from a common atmospheric background state.

One approach is to consider the expected probability distribution function (PDF) of the observation minus background (or short-term forecast) differences. Generally,

one would expect this PDF to be close to a Gaussian distribution with a width reflecting both the instrument errors and errors of representativeness. However, for observations affected by a gross error, the observation minus background PDF would be expected to be very broad and shallow. So, if the observation minus background difference is small, one can infer that the observation is likely to belong to the set of observations without gross errors, and where the difference is large the observation is likely to be affected by a gross error. For any particular observation minus background differences, Bayesian statistics can be used to update the estimated PGE. For more details of the practical application of these concepts, see Lorenc and Hammon (1988).

These concepts are taken further in the variational quality control approach, also known as VarQC (see chapter *Assimilation of Operational Data*, Andersson and Thépaut). A modified observation error probability density function (the standard PDF plus a gross error distribution) is used in the calculation of the observation term in the cost function ($J_o$, as discussed below). Because the gradient of $J_o$ is very small, a very low weight is given to an observation that is a long way from the background. One common approach is to turn VarQC off for the first few iterations of the variational assimilation (described in Sect. 3). When VarQC is turned on for later iterations, the bad data are essentially ignored.

## 3 Data Assimilation

### *3.1 Introduction*

The aim of this section is to give an overview of the various data assimilation methods used for operational NWP. While this section describes the application of different assimilation methods to NWP, the reader is referred to other chapters of this book for mathematical details (see chapters in Part I, *Theory*). We will give an overview of the various approaches used for the objective analysis of meteorological data, culminating with an account of the state-of-the-art in meteorological data assimilation. For the interested reader, Kalnay (2003) gives a good overview of the historical development of data assimilation methods.

In the very early days of objective analysis of meteorological data, the approach used was to fit polynomial functions to the observation values (Panovsky 1949). Gilchrist and Cressman (1954) developed the method further by introducing a region of influence for each observation, and they suggested the use of a background field (from a previous forecast). In the approach of Bergthorsson and Döös (1955) the background field plays a more central role – their technique was based on an analysis of observation minus background differences, rather than the observation values themselves. They attempted to optimize the weights given to each observation based on the accuracy of different observation types, as compiled on a database. Later variations on the technique involved multiple iterations of the analysis – the Successive Correction Method (e.g. Cressman 1959).

Perhaps the most important breakthrough was the adoption of statistical interpolation techniques, which came to prominence in meteorology with the book by Gandin (1963). This put the hitherto pragmatic approach to data analysis onto a proper statistical basis, usually referred to as *Optimal Interpolation* (or OI). The weights given to the observations were properly related to the observation errors. At the same time, the background field was no longer just the starting point for the calculation of the analysis, but instead was recognized as being another useful source of information, with its own error characteristics. Arguably, this is the key idea in the development of data assimilation.

Once computer power made them feasible, data assimilation schemes based on OI were implemented in several operational centres in the late 1970s, e.g., at the European Centre for Medium-Range Weather Forecasts, ECMWF (Lorenc 1981) and the Met Office (Lyne et al. 1982). However, the first implementations had to make major approximations to make the calculations feasible. The ECMWF scheme was based on small analysis volumes, while in the Met Office scheme a maximum of 8 observations influenced each grid-point. These drastic simplifications meant that initial implementations of OI were, in fact, rather suboptimal. In later years this gave the term "Optimal Interpolation" a rather bad name within the meteorological community.

Over about the next 20 years, data assimilation schemes continued to develop, using various approximations to solve the basic statistical equations (see Lorenc 1986). Later analysis schemes provided improved approximations. For example, the Analysis Correction scheme of Lorenc et al. (1991) is, in a sense, a hybrid between OI and the successive correction method. However the key breakthrough has likely been the adoption of variational methods to determine the solution to the OI equations (e.g. Le Dimet and Talagrand 1986).

## *3.2 Variational Methods*

The principle underlying variational data assimilation schemes is that one can construct a global *cost function J* to quantify the mismatch between a model state vector $\mathbf{x}$, and the available information, comprising the background state $\mathbf{x}^b$ (i.e., the best prior estimate of the model state) and (new) observations $\mathbf{y}$ ($T$ denotes transpose):

$$J = \frac{1}{2}[\mathbf{x} - \mathbf{x}^b]^T \mathbf{B}^{-1}[\mathbf{x} - \mathbf{x}^b] + \frac{1}{2}[\mathbf{y} - \mathcal{H}(\mathbf{x})]^T \mathbf{R}^{-1}[\mathbf{y} - \mathcal{H}(\mathbf{x})] \tag{1}$$

The analysis $\mathbf{x}^a$ is defined as the vector $\mathbf{x}$ that minimizes the cost function, i.e., the model state that best fits all the available information. In this equation, $\mathbf{R}$ denotes the error covariance of the observations (taking into account errors of representativeness) and $\mathbf{B}$ the error covariance of the background state. The two terms in Eq. (1) are sometimes denoted $J_b$ and $J_o$, respectively the background and observation components of $J$. This form of variational data assimilation is referred to as 3D-Var (three-dimensional variational), since the equations are solved in three spatial dimensions at a single time. While based on the same statistical considerations as

OI, variational data assimilation does not require the major simplifications employed by the early OI schemes.

The first operational implementation of a three-dimensional variational data assimilation scheme was the Spectral Statistical Interpolation (SSI) system of the US National Centers for Environmental Prediction (NCEP) (Parrish and Derber 1992). This was followed by implementation at ECMWF a few years later (Courtier et al. 1998) and many other operational NWP centres since then.

Because of the large number of variables, variational data assimilation schemes do not perform the minimization of $J$ in the model space but, instead, use a transformed or *control space*. The elements of this control space are the *control variables*; these control variables are chosen in such a way that errors in each control variable can be assumed to be uncorrelated with one another. For example, in the SSI system, the leading control variable is stream function representing non-divergent flow. Additional control variables represent the divergent flow, unbalanced pressure (the component of the mass field not in balance with the wind field) and humidity. The background error covariance $\mathbf{B}$ is defined using a series of control variable and spatial transforms (Parrish and Derber 1992; Lorenc et al. 2000).

It is generally not straightforward to estimate the error covariance values that populate the $\mathbf{B}$ matrix. In the first implementation of SSI, Parrish and Derber (1992) suggested that the forecast errors could be estimated from the difference between pairs of forecasts that verify at the same time. Although this can only give the covariance of the forecast difference, it was found these estimated covariances gave better results than previous estimates computed from forecast minus observation differences. In view of its success, this so-called "NMC method" was widely adopted by operational NWP centres. However, some more satisfactory ensemble-based techniques are now coming into use, as described later.

The NASA Data Assimilation Office (now Global Modeling and Assimilation Office, GMAO) developed an alternative approach, known as PSAS (Physical-space Statistical Analysis System; Cohn et al. 1998). PSAS can be considered to be the dual of 3D-Var, solving the same statistical problem as 3D-Var in observation space, then mapping the solution to model space. PSAS calculates the analysis using the following equation (where $\mathbf{H}$ is the linearization of the observation operator $\mathcal{H}$ about the background trajectory – see chapter *Mathematical Concepts of Data Assimilation*, Nichols):

$$\mathbf{x}^a = \mathbf{x}^b + \mathbf{B}\mathbf{H}^T[\mathbf{H}\mathbf{B}\mathbf{H}^T + \mathbf{R}]^{-1}[\mathbf{y} - \mathcal{H}(\mathbf{x}^b)] \tag{2}$$

The observation-space approach of PSAS is cheaper than the conventional model-space approach if the number of observation values (the dimension of observation space) $p$ is much smaller than the dimension of the model state space $n$. However, the relatively large value of $p$ in operational systems (resulting from the vast quantities of satellite data) means that the PSAS approach is now significantly less competitive than 3D-Var. Reflecting this consideration, the GMAO has

now adopted the NCEP Gridpoint Statistical Interpolation (GSI) scheme (Wu et al. 2002), which was developed from the SSI scheme.

The 3D-Var approach assumes that all observations are valid at the same time, even though they are generally taken over a time window, of typically 6 h. In 3D-FGAT (First Guess at the Appropriate Time), a variant of 3D-Var, the $J_o$ term is calculated by comparing observation values with the background at the relevant observation times. 3D-FGAT goes some way to improving the use of observations when the weather is changing quickly, in a relatively inexpensive manner.

A full treatment of the time variation of both the observations and model state is afforded by four dimensional variational data assimilation (4D-Var). In this algorithm the cost function minimization is carried out over a time window that is typically 6 or 12 h for NWP. The $J_o$ terms takes account of the misfit between the observations $\mathbf{y}_i$ and the model state $\mathbf{x}_i$ at each time-step $i$ in the assimilation window:

$$J = \frac{1}{2}[\mathbf{x}_o - \mathbf{x}_o^b]^T \mathbf{B}_o^{-1}[\mathbf{x}_o - \mathbf{x}_o^b] + \frac{1}{2}\sum_{i=0}^{N}[\mathbf{y}_i - \mathcal{H}(\mathbf{x}_i)]^T \mathbf{R}^{-1}[\mathbf{y}_i - \mathcal{H}(\mathbf{x}_i)] \qquad (3)$$

In principle, the model state $\mathbf{x}_i$ is defined by integrating the full non-linear NWP model from the initial state at the beginning of the time window; the integration of the non-linear model $\mathcal{M}$ over one time-step is shown by $\mathbf{x}_i = \mathcal{M}_i(\mathbf{x}_{i-1})$. However, for a practical solution of the 4D-Var problem, we need to adopt an incremental approach and work in terms of perturbations $\delta\mathbf{x}$ to the first-guess non-linear model trajectory (Courtier et al. 1994). The time evolution of the perturbations can be estimated using a linear model, $\mathbf{M}$, to approximate $\mathcal{M}$: $\boldsymbol{\delta}\mathbf{x}_i = \mathbf{M}_i\boldsymbol{\delta}\mathbf{x}_{i-1}$. The adjustment required to the initial conditions is estimated using the adjoint of the linear model (see chapter *Variational Assimilation*, Talagrand). This cycle of a forward integration of the linear model and the backward integration of the adjoint model is repeated many times to minimize the cost function. The integration of the non-linear model may be repeated to get an updated version of the full model trajectory, followed by further iterations of the linear and adjoint models. This is referred to as a multiple outer loop approach.

The 4D-Var method is the current state-of-the-art in operational data assimilation, and is used by many of the leading operational NWP centres. The version of 4D-Var that we have just described assumes that the model is perfect, i.e., it is a strong constraint. However, model errors can be significant, particularly over longer assimilation time windows. To address this issue, weak constraint versions of the 4D-Var algorithm are being developed, including simplified representations of the effect of model errors (see Trémolet 2007).

As part of the data assimilation process, it is generally beneficial to control spurious high-frequency oscillations in numerical forecasts. A popular method of controlling this noise is normal mode initialization (Machenhauer 1977). More recently, Lynch and Huang (1992) introduced an alternative method of initialization known as digital filter initialization (see chapter *Initialization*, Lynch and Huang).

In the context of 4D-Var, it is reasonably straightforward to apply a digital filter as a weak constraint. An additional term (usually denoted $J_c$) is added to the standard cost function equation, Eq. (1), to penalize departures of the model state $\mathbf{x}$ from the filtered values of that state. Gauthier and Thépaut (2001) applied this to the 4D-Var system at Météo-France, and it has also been applied at other centres including the Met Office.

An alternative approach to data assimilation is the Kalman filter. This is formally similar to the statistical assimilation methods previously discussed, but the major difference is that the forecast error covariance $\mathbf{P}^f$is evolved using the linear model $\mathbf{M}$ itself, rather than approximating it as a constant covariance matrix $\mathbf{B}$. The Kalman filter equations are:

$$\mathbf{x}_i^f = \mathbf{M}_{i-1}\mathbf{x}_{i-1}^a, \tag{4a}$$

$$\mathbf{P}_i^f = \mathbf{M}_{i-1}\mathbf{P}_{i-1}^a\mathbf{M}_{i-1}^T + \mathbf{Q}_{i-1}, \tag{4b}$$

$$\mathbf{x}_i^a = \mathbf{x}_i^f + \mathbf{K}_i\left[\mathbf{y}_i - \mathbf{H}\mathbf{x}_i^f\right], \tag{4c}$$

$$\mathbf{K}_i = \mathbf{P}_i^f\mathbf{H}_i^T\left[\mathbf{R}_i + \mathbf{H}_i\mathbf{P}_i^f\mathbf{H}_i^T\right]^{-1}, \tag{4d}$$

$$\mathbf{P}_i^a = \left[\mathbf{I} - \mathbf{K}_i\mathbf{H}_i\right]\mathbf{P}_i^f. \tag{4e}$$

Starting with an analysis, $\mathbf{M}$ is used to produce a forecast at the next time-step, Eq. (4a). In parallel with this, Eq. (4b) is used to derive the forecast error covariance $\mathbf{P}^f$ from the previous analysis error covariance and the model error covariance $\mathbf{Q}$. Equation (4c) is the analysis step, using the weight matrix (Kalman gain) defined in Eq. (4d); this step is exactly the same as in PSAS, see Eq. (2). Equation (4e) updates the error covariance to take account of the assimilated data. The basic Kalman filter uses a linear model $\mathbf{M}$, and can be shown to give the same analysis $\mathbf{x}^a$ as the corresponding 4D-Var scheme. The Extended Kalman filter is a development of this scheme that uses a non-linear model$\mathcal{M}$, and might be viewed as the "gold standard" of data assimilation.

For the large problems that need to be tackled for operational NWP, the cost of the covariance evolution Eqs. (4b) and (4c) is prohibitive. However, the Kalman filter is applied in various approximate forms for example in some constituent assimilation work (see chapter *Constituent Assimilation*, Lahoz and Errera) and in the Ensemble Kalman filter discussed below.

## *3.3 Assimilation of Satellite Soundings*

Satellite observations are, generally speaking, measurements of radiation at a range of different wavelengths. Since these measurements are related to atmospheric model variables in rather complex ways, the treatment of satellite soundings is often closely tied to the data assimilation system. Chapter *Assimilation of*

*Operational Data* (Andersson and Thépaut) gives a much more detailed account of the assimilation of satellite data but, for completeness, we summarize some of the key issues here.

The most basic approach to the assimilation of satellite data is to assimilate retrievals, e.g., temperature profiles. These retrieved data are often supplied by the agencies that provide the satellite instruments. When satellite data first started to be assimilated for NWP, it was common to use retrievals, and assimilate them in much the same way as profiles measured by, for example, radiosondes. However, there are a number of disadvantages to this approach. First, the error characteristics of retrievals are generally poorly known; it is hard to track the errors (and particularly their correlation) through all the steps of the calculation. Second, it is likely that the retrieval product will suffer from a poorer prior estimate of the atmosphere than the background data from an NWP model; assimilating the retrievals could even have a negative impact. Rodgers (2000) gives a thorough account of the application of inverse methods to atmospheric soundings.

An improved approach is the use of locally produced retrievals. A 1-D variational approach can be used to derive retrievals using background information from profiles extracted from a recent short-range forecast. While the prior information in this case would generally be very accurate, the errors in the retrieval may still be hard to characterize. In particular, it may be hard to avoid the assimilation using the forecast background twice: once directly and once indirectly via the retrieval.

Variational techniques allow the direct assimilation of radiance observations, and therefore avoid the need for an explicit retrieval step. For radiance assimilation, the observation operator $\mathcal{H}$ – see Eq. (1), incorporates a (simplified) radiative transfer model that maps the atmospheric profile to radiance space. The forecast background provides the prior information to supplement the radiances. Furthermore, the inversion is further constrained by the assimilation of other observations. This procedure also has the advantage that radiance errors are much easier to characterize than the retrieval errors.

For NWP, a pragmatic approach to the assimilation of satellite data is adopted. For vertical temperature soundings, it is much easier to characterize the radiance errors, and much effort has been spent to develop suitable radiative transfer models, so radiance assimilation is the best approach. However, there are still some situations where the assimilation of retrievals may be the most practical approach; for example, it is much more straightforward to assimilate winds derived from tracking clouds between satellite images than to assimilate the image data themselves.

## *3.4 Ensemble Assimilation Methods*

One promising assimilation method that is not yet in widespread use is the *Ensemble Kalman filter* (EnKF) – see chapter *Ensemble Kalman Filter: Current status and potential* (Kalnay). The principle is that an ensemble of $m$ data assimilation cycles

are carried out simultaneously, where *m* is chosen to be sufficiently large to enable the ensemble to give a good representation of the probability distribution of possible atmospheric states. All the ensemble members assimilate the same set of observations, but in order to keep them independent a different set of random perturbations is added for each ensemble member. Rather than explicitly modelling the forecast error covariance evolution (as in an Extended Kalman filter), the evolution of the forecast error covariance is calculated from the ensemble spread.

A particularly promising approach is the use of a localized Ensemble Kalman filter (e.g. Ott et al. 2004). With a limited number of ensemble members *m*, the error covariance estimated from the ensemble is very likely to indicate spurious long-distance correlations. As a result, assimilating an observation at a particular location will lead to spurious analysis increments at some distance. To avoid this problem, an Ensemble Kalman filter may be run for smaller (perhaps overlapping) analysis volumes. An alternative approach is to use localization functions to multiply the correlations derived from the ensemble to ensure that they do not spread more than a few 1,000 km.

An Ensemble Kalman filter is very much cheaper than an Extended Kalman filter. At the same time, it does not require the development of a linear and adjoint model. Research at the Canadian Met Service (e.g. Houtekamer and Mitchell 2005) has shown better performance than 3D-Var, and approaching the performance of 4D-Var.

Rather than replacing 4D-Var with ensemble methods, some operational NWP centres are considering a *hybrid assimilation method* whereby an ensemble approach is used to estimate the background error covariances for use in a 4D-Var system. Currently, ECMWF uses an ensemble data assimilation system to estimate the climatological background error covariances used in their 4D-Var operational system (Fisher 2003).

The next logical step is to derive the background error covariances from a current real time ensemble, allowing representation of the actual flow-dependent "errors of the day". However it is not possible to make an accurate estimate of the error covariances without using a rather large ensemble. To circumvent this problem, Hamill and Synder (2000) ran experiments with a simple 3D-Var data assimilation system, using a judicious blend of climatological error covariances with those estimated from short-range ensemble forecasts. They showed that this hybrid system gave improved performance compared with either using just the climatological errors or just the ensemble-based errors. More recently, Wang et al. (2008a, b) have developed a hybrid Ensemble Transform Kalman filter (ETKF) – 3D-Var data assimilation for the WRF (Weather Research and Forecast) model, and confirmed the benefit of the hybrid approach, particularly in data sparse regions. The hybrid technique could also improve on 4D-Var assimilation systems; although 4D-Var systems implicitly evolve the background error covariances during the assimilation time window, they are still limited by the static specification of error covariances at the start of the window. Ensemble methods are discussed further in the next section.

# 4 Numerical Modelling

## *4.1 Development of Numerical Models*

Numerical models of the atmosphere are described in some detail in chapter *The Role of the Model in the Data Assimilation System* (Rood). The aim of this section is to describe how they are applied in an operational NWP context.

The interest in using physically based models to forecast the weather goes back to Richardson (1922), who attempted to forecast the change in surface pressure over Europe from observed data, albeit rather unsuccessfully. When electronic computers were first developed, numerical weather forecasting started to become a realistic proposition. One of the pioneers was Jules Charney, who realised that the most practical approach was to use "filtered" equations of motion, based on quasi-geostrophic balance. Subsequently, Charney et al. (1950) computed a 1-day forecast using a barotropic (one-level) filtered model, using the ENIAC computer. The first real-time, operational NWP forecast was run in Sweden in September 1954. The USA followed the next spring, but it was not until 1965 that the Met Office started running numerical weather forecasts.

In a visionary article, Charney (1951) saw that, while very useful for understanding dynamical processes, quasi-geostrophic equations would not be sufficiently accurate to allow continued progress in NWP. Primitive equation models would be required, incorporating conservation laws for mass, energy and momentum. In addition the major effects of adiabatic physical processes (rainfall, radiation and boundary layer processes) would need to be represented. These developments have all subsequently been implemented in NWP models.

It has been found that the accuracy of NWP models is strongly influenced by their resolution: the higher the resolution, the more accurate the model. However, halving the resolution in each direction not only means that 8 times as many grid points are required, but also that the number of time-steps need to be doubled to keep the computations stable. Alongside the escalating costs of solving the dynamical equations, the physical parametrization schemes are becoming increasingly sophisticated. As a result, running atmospheric models has always required some of the fastest supercomputers available.

However, the potentially escalating cost of numerical modelling has been kept in check by the development of modern discretization techniques, including the use of semi-implicit and semi-Lagrangian schemes, which have less stringent stability conditions on the time-step, and more accurate space discretizations. For example, Davies et al. (2005) describe the numerical integration methods used for the dynamical core of the Met Office Unified Model.

In many models, the equations are primarily solved on a set of grid points. These grid points are typically arranged in a rectangular fashion on a latitude-longitude grid. The main disadvantage is that near the poles the east-west grid lengths become very small, which means that a shorter integration time-step is required (unless semi-Lagrangian techniques are used), or spatial filtering is required. For limited area models, this problem can be avoided by using a rotated grid (see Fig. 2). For

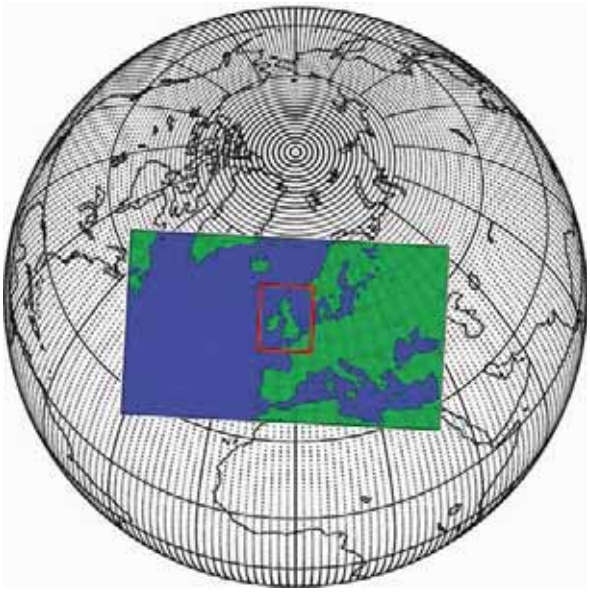


**Fig. 2** Domains covered by the Met Office Global, North Atlantic and Europe (NAE) and 4 km United Kingdom (UK4) models

global models, several different model grid geometries have been proposed, including icosahedral grids, the "cubed sphere" or the "Fibonacci grid", see Purser (1999) and Swinbank and Purser (2006). A popular alternative approach is to use spherical harmonics to solve the dynamical equations on a sphere. In practice, many global numerical models employ a combination of spectral and grid-point techniques.

## *4.2 Model Configurations*

Early experiments with numerical prediction models used models that covered regional or continental scales. The main focus was on forecast ranges of around 1–2 days. As the skill of NWP models improved, it became feasible to extend the forecast range. This entailed the use of large model domains, since information affecting a particular location would propagate from further afield. In the 1970s, several operational NWP centres were using models that covered most of a hemisphere, for example the Met Office "octagon model" that covered the Northern Hemisphere extra-tropics. 1978/1979 saw the First GARP Global Experiment (FGGE), which coincided with a major improvement in the availability of operational satellite soundings. This gave an impetus to the development of global NWP models, and the start of operational medium-range weather forecasts (notably at the then recently-founded ECMWF).

The use of global models opened up the prospect of seamless forecasting, from short-range potentially to seasonal or even climate forecasts. Indeed, modern numerical models share a strong common heritage with general circulation models used for global climate simulations. For example, at the Met Office the Unified Model

(Cullen 1993) is run in different configurations for both NWP and climate simulations. For seasonal and longer ranges, the atmospheric model is often run with a coupled ocean in order to simulate changes to the ocean and consequent changes to atmospheric forcing. Models of the vegetation and land surface may also be coupled to the atmosphere and ocean; such models are often referred to as Earth System Models.

While global models are necessary for extended forecast ranges, regional models are the best way to produce short-range forecasts. Limited area models can be run at very high resolutions to allow much more detailed forecasts than would otherwise be possible. They can also generally be run closer to real time, since they can be initialized using local observations, rather than needing to wait for observations to be collected from around the globe. The mix of observations assimilated into a regional model may include data derived from radar measurements of rainfall and wind, for example, but may not include the range of satellite observations used in global models.

Boundary conditions for the limited area models are generally provided by nesting inside models covering larger domains. For example, at the Met Office three main NWP model configurations are currently in use: the global model; a North Atlantic European (NAE) model; and mesoscale model covering the UK. Figure 2 illustrates the model domains. The model resolutions are being improved every few years; in the current (2009) configuration the global model has a grid-length of around 40 km, the NAE model 12 km and the UK mesoscale model 4 km.

A similar set of model configurations is run by other national meteorological services. However, many centres do not run their own global model, but use boundary conditions provided by other centres; for example, some European regional models use the ECMWF global model to provide boundary conditions. Although not always possible, it is preferable to nest different configurations of the same model together, so that the boundary conditions are as consistent as possible with the formulation of the limited area model.

## 5 Ensemble Forecasting

### *5.1 Benefits of Ensemble Forecasts*

So far, we have considered deterministic forecasts, in which a single set of outcomes is predicted. For example, it may be predicted that 5 mm of rainfall will occur in Exeter between 9 and 12 GMT next Tuesday. But, it may well be that the actual rainfall is 8 mm, or it may occur in the afternoon instead, or, if the weather is showery, the rain may miss Exeter completely. For many applications, it is very helpful to be able to issue *probabilistic forecasts*, indicating the range of likely outcomes, rather than forecasting whether a single event will occur.

Uncertainties will exist in the initial conditions of each weather forecast, and these will grow during the forecast period. Lorenz (1963) discovered the fact that the atmosphere, in common with many other dynamical systems, has a finite limit

of predictability. By performing two runs made with the same model, with initial conditions that differed only with round-off errors, he found that the two solutions completely diverged. For the large-scale atmospheric circulation, Lorenz estimated a limit of deterministic predictability of around 2 weeks. But individual small-scale weather features, such as showers, are much less predictable: a more realistic time-scale may be just an hour or two.

Leith (1974) first proposed the idea of performing ensemble forecasting with a number of ensemble members rather than the conventional single deterministic forecast. An ensemble should be designed to reflect uncertainties in the forecast at any given time. So, the $m$ members of an ensemble forecast should represent, in some sense, the probability distribution of possible model states. Forecast ensembles are often defined such that one of ensemble members reflects the best available estimate of the state of the atmosphere, and how it evolves – usually referred to as the *control forecast*. Typically, the initial conditions for the control forecast are produced using a state-of-the-art data assimilation system, as used for a deterministic forecast. Other ensemble members are derived by adding perturbations to those initial conditions, as discussed below.

An interesting property of ensemble forecasts is that the error of the ensemble mean of many forecasts should be less than statistical errors in a single forecast. Consider the deviation $\mathbf{a}$ of forecast model variables with respect to climatology. The true state of the atmosphere is denoted $\mathbf{a}_0$. The value $\tilde{\mathbf{a}}$ denotes an unbiased estimate of $\mathbf{a}_0$, whose expected value at long lead times (averaged over many forecasts) is zero: $< \tilde{\mathbf{a}} >= 0$. If we were to use climatology to estimate $\mathbf{a}_0$, the expected error covariance would be $<(\mathbf{0} - \mathbf{a}_0)(\mathbf{0} - \mathbf{a}_0)^T>=\mathbf{A}$. A single deterministic forecast $\tilde{\mathbf{a}}$ would have an error covariance $<(\tilde{\mathbf{a}} - \mathbf{a}_0)(\tilde{\mathbf{a}} - \mathbf{a}_0)^T> = <\tilde{\mathbf{a}}\tilde{\mathbf{a}}^T+\mathbf{a}_0\mathbf{a}_0{}^T - \tilde{\mathbf{a}}\mathbf{a}_0{}^T - \mathbf{a}_0\tilde{\mathbf{a}}^T>$, which would tend to a limit of $2\mathbf{A}$, since the last two terms would be zero at long lead times. However, if $\overline{\mathbf{a}}$ is the average of the ensemble of $m$ forecasts, then its error covariance $<(\overline{\mathbf{a}} - \mathbf{a}_0)(\overline{\mathbf{a}} - \mathbf{a}_0)^T>$ tends to a limit of $(1 + 1/m)\mathbf{A}$. In other words, the root mean square error of a deterministic forecast will saturate at around $\sqrt{2}$ times the error of a forecast based on climatology, while the mean of an ensemble should converge to the climatological average with a spread equivalent to the climatological error.

## *5.2 Initial Condition Perturbations*

Early ensemble prediction experiments used lagged averaged forecasting (e.g. Hoffman and Kalnay 1983), in which forecasts initialized at one time are combined with forecasts initialized at $m$–1 previous times, i.e., the initial perturbations simply reflect the differences between the initial analysis fields and a set of forecasts valid at that time.

Toth and Kalnay (1993) developed an *error breeding* approach in order to ensure that the perturbations better reflected the uncertainties in the initial conditions. Perturbed forecasts are run in parallel to the unperturbed control forecast.

On a regular basis (e.g. every 6 h) the perturbations are scaled back to a standard size (defined using the same norm), consistent with typical uncertainties in the analysis. It was found that the perturbations generated acquired a fast growth rate. This error breeding method favours the fastest growing modes, referred to as the leading *Lyapunov vectors*. While this method generates a good estimate of the modes that lead to uncertainties in the initial conditions, the bred vectors derived from a set of perturbation runs can be strongly correlated with one another. To ensure that the ensemble better samples all the uncertainties, a variation has recently been introduced into the NCEP system in which the vectors are transformed, using the "Ensemble Transform", to ensure they are mutually orthogonal. In a further level of sophistication, the Met Office uses the Ensemble Transform Kalman filter (ETKF; Wang and Bishop 2003), so that the perturbations are adjusted to take account of the information introduced by the assimilation of observations. Further details about the Met Office ensemble forecast systems are given by Bowler et al. (2008).

An alternative approach is to base the ensemble perturbation on singular vectors, i.e., the fastest growing modes, determined over a specified period. In the ECMWF system the initial condition perturbations are based on the singular vectors that grow fastest over a 48-h optimization time using a total energy norm.

Ideally, the initial condition perturbations should reflect uncertainties at the initial time, i.e., the structure of the initial perturbations should be closely related to the analysis area covariance. The error-breeding and related methods, including the ETKF, are consistent with this approach, since they are based on perturbations that have grown in the period leading up to the analysis time. On the other hand, the singular vectors highlight modes that grow in the initial forecast period. The use of evolved singular vectors, based on the maximum growth over an extended period rather than just the initial time-step, brings the singular vector method more into line with the other approaches, while retaining the link with the fastest growing modes.

## *5.3 Accounting for Model Errors*

During the forecast, the ensemble spread should ideally reflect the increasing (root mean square, RMS) error as the forecast proceeds. In practice, it is usually found that the ensemble spread grows more slowly during the forecast than the RMS error. The main reason for this is that much of the increasing forecast error reflects shortcomings in the numerical model's representation of the real atmosphere. One approach to this problem is to introduce some stochastic fluctuations in the tendencies calculated by the model's physical parametrization schemes (Buizza et al. 1999). These fluctuations should be applied to the tendencies in a manner that reflects uncertainties in the model physics. In a similar way, the "random parameters" component of the Met Office stochastic physics varies several of the parameters used by the physical parametrizations. The parameters are (slowly) varied, in a random manner, within each parameter's range of uncertainty.

A complementary approach is to represent stochastic kinetic energy backscatter. In this type of scheme, some of the kinetic energy that is lost (unrealistically) in numerical diffusion is reinstated by introducing extra random fluctuations in the wind fields. Shutts (2005) first applied this type of scheme in the ECMWF model, using cellular automata to generate pseudo-random patterns for a stream function forcing field. These additional perturbations increased the ensemble spread and had a beneficial impact on probabilistic measures of forecast skill. The Met Office stochastic kinetic energy backscatter scheme was designed to have a similar effect, though it uses a difference technique to generate "random" wind fluctuations. A new version of the stochastic kinetic energy backscatter scheme will also take into account energy dissipation associated with convection as well as with diffusion.

Another way of accounting for model errors in ensemble forecasts is to use more than one model. NWP models from different operational centres will have different strengths and weaknesses. By building an ensemble from forecasts of different models, one can construct a *multi-model ensemble* whose spread reflects the uncertainties in model parametrizations. In the context of seasonal forecasting, it has been shown that the combination of ensemble forecasts from different models results in more skill than the single model ensembles considered separately (e.g. the DEMETER project; see Palmer et al. 2004). This improvement is not just a result of the increased ensemble size, but is also due to complementary information provided by the different climate forecast systems (i.e., the combinations of data assimilation and numerical model).

As part of the international THORPEX (THe Observing system Research and Predictability EXperiment) programme, research is being done on the benefit of building a grand ensemble, combining ensemble forecasts from different centres. In the THORPEX Interactive Grand Global Ensemble (TIGGE) project (see http://tigge.ecmwf.int/), several global NWP centres are running regular medium-range ensembles and making the output available for research. Figure 3 (from Johnson and Swinbank 2009) compares *Brier skill* scores from three single-model ensembles with different versions of multi-model ensembles combining the three models. This study, and other studies based on TIGGE data (Park et al. 2008; Matsueda and Tanaka 2008), demonstrated only limited benefit of multi-model ensembles for forecasts of 500 hPa height and sea level pressure. On the other hand, multi-model techniques give better benefit for 2 m temperature and, to a lesser extent, 850 hPa temperature. It was also found that a simple multi-model ensemble (giving each ensemble the same weight) performed almost as well as more complex weighting schemes that took into account differing model errors.

Another method of correcting for model errors is to estimate them from a set of retrospective forecasts. Hamill et al. (2006) built up a large set of *reforecasts* using a fixed (T62 resolution) version of the NCEP GFS (Global Forecast System) model, initialized from the NCEP/NCAR reanalyses (Kalnay et al. 1996). They refer to the data as "reforecasts" by analogy with the fixed data assimilation system used to construct reanalysis datasets. By comparing the reforecasts with suitable verifying analyses or observations, it is possible to characterize model errors and apply them to current forecasts. Recent results from ECMWF (R. Hagedorn, personal communication) indicate that the benefit from applying corrections based

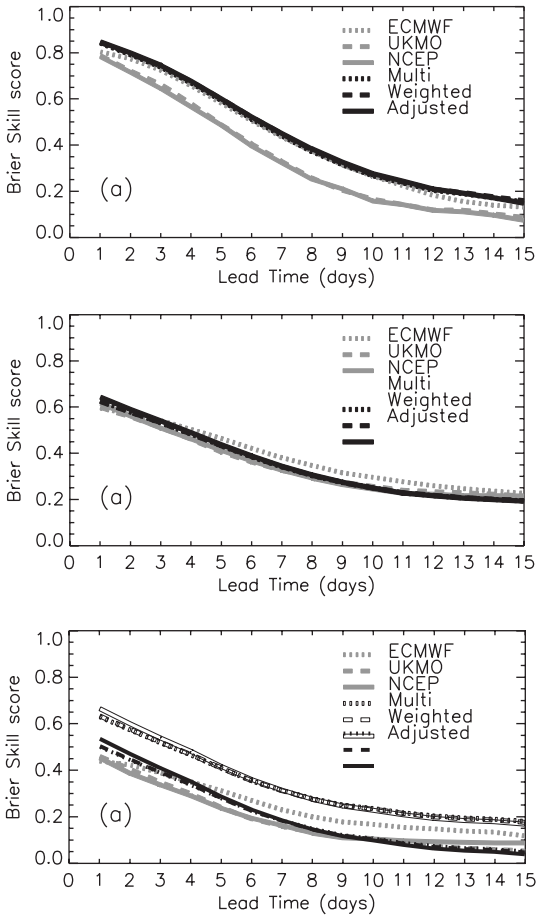



**Fig. 3** Brier skill scores for: mean sea level pressure greater than the climatological mean (*top plot*); 2 m temperature greater than the climatological mean (*middle plot*); 2 m temperature greater than 90th percentile (*bottom plot*). The *grey lines* show the bias-corrected single-model ensembles (ECMWF, Met Office and NCEP) and the *black lines* show three difference multi-model ensembles: simple combination (*dotted*), weighted (*dashed*), weighted and variance adjusted (*solid*). The data are globally averaged over 120 days ending 29 April 2008

on reforecasts is similar in magnitude to that obtainable from multi-model ensembles. Results from Hagedorn et al. (2008) indicate that calibrated multi-model forecasts are better than calibrated single-model forecasts, so it is likely to be worth employing both techniques in tandem.

## 6 Forecast Products

### *6.1 Weather Forecasts*

The output from NWP models needs a certain amount of processing before the information is presented to the public as a weather forecast. Some operational NWP centres also produce a range of more specialized services and products for particular

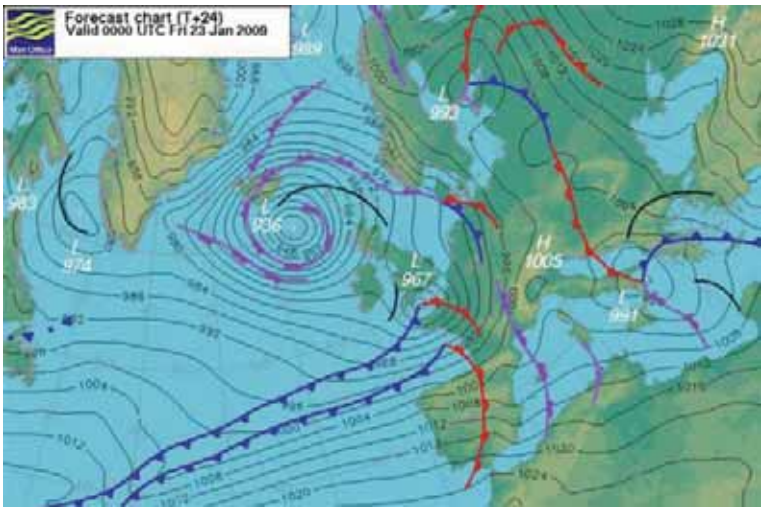



**Fig. 4** Example of forecast mean sea level pressure chart for the North Atlantic and Europe. The contours show pressure in hPa and the coloured lines with symbols show fronts (*blue*: cold front; *red*: warm front; *purple*: occluded front)

groups of customers in both the public and private sectors. There are also a range of commercial companies that take raw model output produced by the public weather services and process them to generate added value products. The aim of this section is to give a brief overview of how NWP output is processed to meet the needs of a wide variety of customers.

Perhaps the most basic type of forecast information is a synoptic map of mean sea level pressure, as shown in Fig. 4. To a meteorologist, or other well-informed user, this gives a readily interpreted overview of the expected weather. Surface winds roughly follow the isobars, but with some frictional flow towards low pressure centres. Fronts mark the areas where large-scale rainfall is expected. The flow patterns can also give an indication of expected temperatures or shower activity. In order to make the weather forecasts more generally understandable to the general public, maps are often produced using a set of symbols to show the expected weather. Figure 5a shows an example of this kind of simplified map, for the same forecast as shown in Fig. 4. Weather forecasts for the general public may also include more specific information, such as surface air temperature and surface wind. The forecast should also include at least some indication of cloud amount, amount of precipitation (highlighting if snow is expected) and visibility (especially if it will be poor).

## *6.2 Site-Specific Information*

In many cases, weather forecasts need to be made for specific locations. A simple approach would be to interpolate the NWP model output directly to each site where the forecast is required. However, the raw model output is very likely to

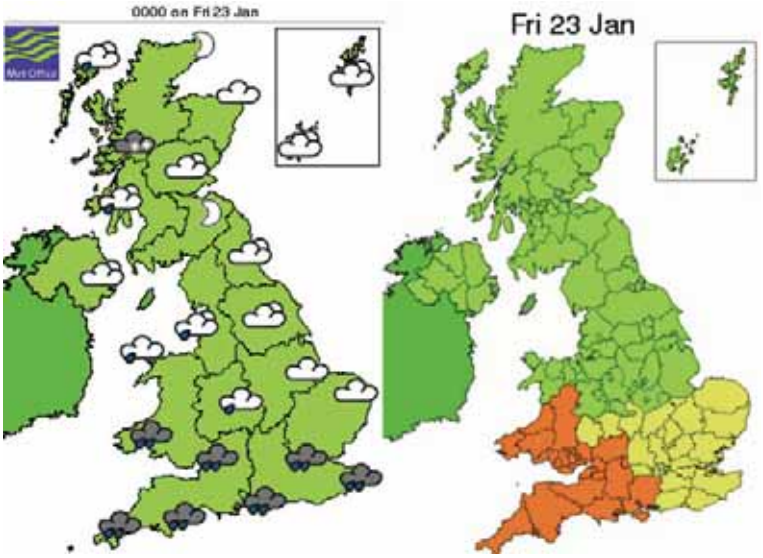



**Fig. 5** Examples showing simpler presentations of the forecast from Fig. 4: (**a**) forecast weather-symbol map, *left-hand plot*; (**b**) map indicating areas where severe weather has been forecast (*green indicates* no warning, *yellow* "be aware" and *orange* "be prepared"), *right-hand plot*. The weather warning map covers the whole day, while the weather symbol map is for just one time; in this case the warning map highlights the risk of overnight heavy rain, shown in the 0000 UTC forecast maps

need adjustment: the model output likely contains biases; the model topography is unlikely to represent the local topography at a particular location in question.

To correct for these effects, a method known as *Model Output Statistics* (MOS) is often applied. This is essentially a multiple linear regression method, where the predictors are model forecast variables, and may include other parameters such as the time of day or time of year. The predictand is a set of observations at the site under consideration. Based on a sequence of training data, a set of regression coefficients are derived, to quantify how the observed values are related to the model forecast values. These regression coefficients can then applied to subsequent NWP model output values in order to calculate the corrected forecast at that site. As in any statistical regression, the quality of the results is dependent on the quality and length of the training data.

A variation on the MOS method is based on the Kalman filter equations, Eq. (4). In the standard MOS approach, the regression coefficients are calculated only once, but by using a Kalman filter the regression coefficients are updated. This allows the system to track changes to the NWP model. This may be a more satisfactory method of taking into account seasonal variations than using a multi-annual training period.

## *6.3 Probabilistic Forecasts*

With some more post-processing the model output can be used to produce warnings of severe weather events. While experienced forecasts can interpret the model

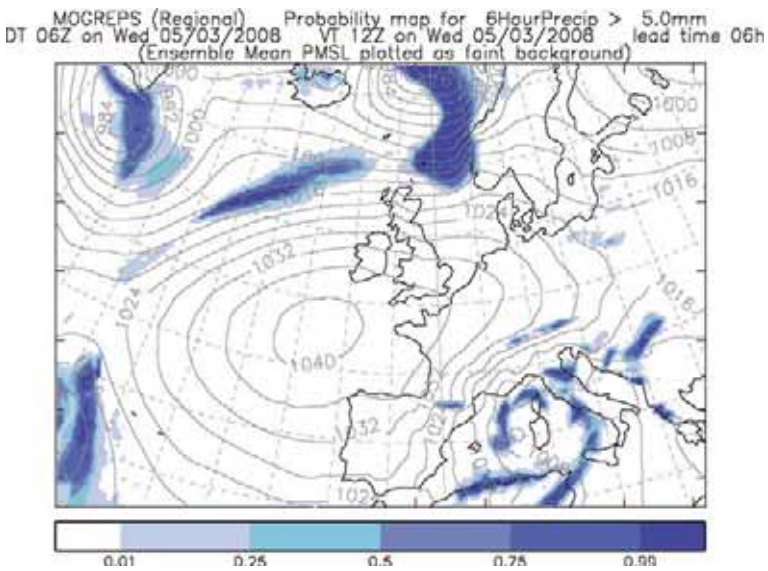



**Fig. 6** A chart showing the spatial variation in the probability of the 6-h rainfall exceeding 5 mm

output and give end users an indication of forecast uncertainties, more objective probabilistic forecasts can be based on the output from ensemble forecasts.

An ensemble prediction system is usually designed so that, as far as possible, each ensemble member is as likely as any other. So, the probability of a particular event occurring can be estimated from the proportion of ensemble members in which that particular events is forecast. For example the event may be that at least 5 mm of rain occurs in a specified 6-h period. By scanning the forecast precipitation field from all the ensemble members, it is straightforward to plot a map showing the estimated probability of rainfall exceeding the specified threshold. Figure 6 gives an example of a probabilistic rainfall forecast map.

Another popular way of presenting forecasts for a particular location is to produce a set of time series plots showing forecasts for different weather elements, such as temperature, rainfall or wind speed. These are often referred to as *meteograms*. Figure 7 shows an example of site-specific temperature forecast derived from an ensemble prediction. This shows the expected temperature, based on the ensemble mean, while the red and orange shading indicates the likely range of values. The previous day's observed temperature is also shown, for reference.

## *6.4 Warnings of High-Impact Weather*

Forecasting severe weather is a particularly important aspect of weather forecasting. Severe weather events, such as floods or winds, have a very high socio-economic impact, disrupting a wide range of everyday activities. It is vitally important that the public and the emergency services are well-prepared when severe weather events

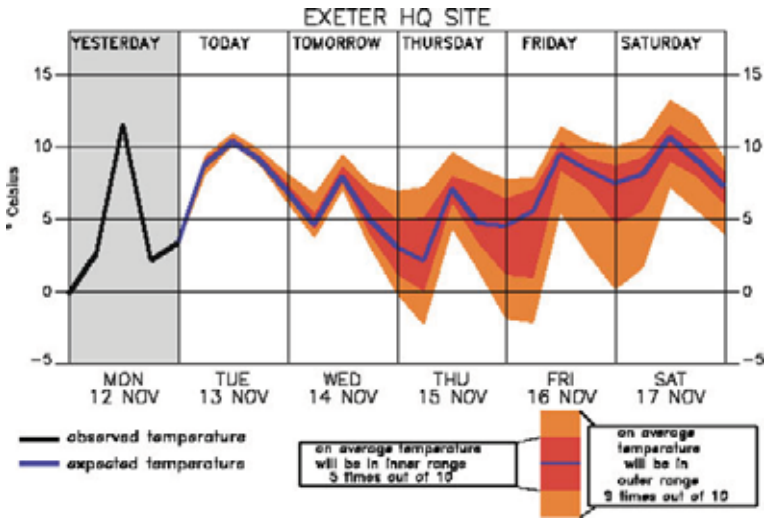



**Fig. 7** Example of a site-specific temperature forecast, showing associated levels of confidence

are expected. So, at the Met Office, and other weather services, warnings are issued and widely publicised when severe weather is expected.

Figure 5b is an example of how severe weather events are advertised on the Met Office website. Different colours are used to indicate the severity of the expected weather for a particular day: yellow indicates that the public should be aware of the likelihood of bad weather; orange indicates the need to be prepared for severe weather and, in the most severe cases, red indicates action should be taken. Alongside the map, the website gives more details about the nature of severe weather, and the time period when it is expected. For the example shown in Fig. 5b, very heavy rainfall was forecast for the indicated areas (see Fig. 5a).

Heavy rainfall can lead to localized flooding, where the intensity of rainfall overwhelms local drainage systems; these are sometimes referred to as pluvial floods. Very often, prolonged periods of heavy rainfall can lead to rivers bursting their banks and fluvial flooding. Prediction of fluvial flooding depends not only on forecasting rainfall but also on understanding the hydrology of relevant river catchments. Another type of flooding that we need to be able to forecast is coastal flooding that can result from storm surges. Sea level is affected by storms as well as astronomical tides; where adverse weather coincides with the highest (spring) tides, the sea may overtop coastal defences.

While some severe weather events, such as strong winds or heavy rainfall, are linked to particular storms, there are other types of weather that can have high socio-economic impacts. For this reason, we often refer to “high-impact weather”. While a brief dry spell is of little consequence, a prolonged drought can have a devastating effect on agriculture. A heat wave comprising several unusually hot days (and nights) can have a major impact on public health – for example the heat wave that

affected western Europe in summer 2003. Prolonged hot, dry weather can also lead to devastating wildfires.

## *6.5 Improving the Prediction of High-Impact Weather*

To improve the prediction of high-impact weather, the WMO is co-ordinating a decade-long research programme, known as THORPEX (see above). The aim of THORPEX is to co-ordinate research into a range of topics that potentially improve the skill of weather forecasts. Once the benefit of novel techniques have been demonstrated, it is planned that prototype products would be delivered to prospective users via a Global Interactive Forecast System (GIFS).

We have already touched on some of the work covered by the THORPEX programme, including research into improvements to data assimilation and observing systems and the TIGGE project on ensemble forecasting. Another THORPEX research topic has been the use of targeted observations to reduce uncertainties in the prediction of high-impact weather events. The principle is that errors in the outcome of a weather forecast can generally be traced back through the earlier stages of the integration of the NWP model. For example, an ensemble forecast might show the risk of development of a severe storm at a 3-day lead time. Tracing back the evolution of the ensemble might indicate a sensitive area in the 1-day forecast where additional observations would reduce the uncertainty in the forecast evolution. So, by making additional observations in that area, perhaps by deploying dropsondes from an aircraft, one should be able to improve the skill of the storm forecast. Several different techniques have been used to predict these sensitive areas: NCEP and the Met Office use a technique based on the ETKF, while ECMWF use a singular vector method.

In the USA, NOAA (National Oceanic and Atmospheric Administration) runs a regular Winter Storm Reconnaisance (WSR) programme, in which aircraft are deployed in the Pacific Ocean to make additional targeted observations using dropsondes. The first WSR campaign was carried out in early 1999 and was found to improve the forecast in 18 out of 25 cases (Szunyogh et al. 2000). Following this success, the programme has been repeated every winter. Other observation targeting experiments have been run in the North Atlantic, including FASTEX (Fronts and Atlantic Storm Track Experiment) in 1997 and ATReC (Atlantic THORPEX Regional Campaign) in 2003. Generally speaking, the Atlantic observation targeting campaigns have been less successful than the Pacific WSR campaigns. One factor is that the North Atlantic is not so big, and not so data sparse as the North Pacific. Also, more recent Atlantic campaigns have tended to use 4D-Var systems to evaluate the impact of observation targeting. 4D-Var systems tend to be better than 3D-Var at exploiting sparse data, so benefit less from additional targeted data. A study by ECMWF explored the value of observations in targeted sensitive areas over an extended period (Buizza et al. 2007). They confirmed that targeted observations are on average more valuable than observations in randomly selected areas, but their overall impact on forecast skill was rather marginal.

Where new techniques have been shown to be cost effective, it is planned to use them as the basis for new products to be delivered to users, particularly to alert them to forecast high-impact weather events. It is planned that GIFS will build on the WMO Information System (WIS) that has been designed for the international distribution of weather-related information. The WMO Severe Weather Forecast Demonstration Project (SWFDP) demonstrated how information about high-impact weather for southern Africa could successfully be relayed from global NWP centres to the South African Weather Service (SAWS). In turn, SAWS interpreted the data and distributed the information to the national weather services in several of the less-developed countries in that region, helping them improve warning services to their local communities. The success of the SWFDP in southern Africa has led to interest in setting up SWFDP sub-projects in other regions, as well as extending the project in southern Africa to a wider region. It provides a good example for how new high-impact weather products developed under THORPEX could be used to improve regional weather warnings.

## 7 Conclusions

This chapter gives a general overview of the weather forecasting process, starting with making observations, assimilating the data into an NWP model, running a numerical forecast and producing a range of forecast products. Our aim has been to give readers a general appreciation of the techniques that have led to the rapid advances in weather forecast skill that we have seen over recent decades. In this single chapter, there has been little room to go into detail; other aspects of numerical weather prediction, and specifically data assimilation, are covered elsewhere in this book in more detail. We have included discussion of some novel techniques which should enable further advances in the skill of operational weather forecasts over the coming years.

## References

Bergthorsson, P. and B. Döös, 1955. Numerical weather map analysis. *Tellus*, **7**, 329–340.

Bowler, N.E., A. Arribas, K.R. Mylne, K.B. Robertson and S.E. Beare, 2008. The MOGREPS short-range ensemble prediction system. *Q. J. R. Meteorol. Soc.*, **134**, 703–722.

Buizza R., C. Cardinali, G. Kelly and J.N. Thépaut, 2007. The value of observations. II: The value of observations located in singular-vector-based target areas. *Q. J. R. Meteorol. Soc.*, **133**, 1817–1832.

Buizza, R., M. Miller and T.N. Palmer, 1999. Stochastic representation of model uncertainties in the ECMWF ensemble prediction system. *Q. J. R. Meteorol. Soc.*, **125**, 2887–2908.

Charney, J., 1951. *Dynamical Forecasting by Numerical Process*. *Compendium of meteorology*. American Meteorological Society, Boston.

Charney, J., R. Fjørtoft and J. von Neumann, 1950. Numerical integration of the barotropic vorticity model. *Tellus*, **2**, 237–254.

Cohn, S.E., A. da Silva, J. Guo, M. Sienkiewicz and D. Lamich, 1998. Assessing the effects of data selection with the DAO Physical-space Statistical Analysis System. *Mon. Weather Rev*., **126**, 2912–2926.

Courtier P, E. Andersson, W. Heckley, et al., 1998. The ECMWF implementation of three-dimensional variational assimilation (3D-Var). I: Formulation. *Q. J. R. Meteorol. Soc*., **124,** 1783–1807.

Courtier, P., J.N. Thépaut and A. Hollingsworth, 1994. A strategy for operational implementation of 4D-Var, using an incremental approach. *Q. J. R. Meteorol. Soc.*, **120**, 1367–1387.

Cressman, G., 1959. An operational objective analysis system. *Mon Weather Rev.*, **87**, 367–374.

Cullen, M.J.P., 1993.The unified forecast/climate model. *Meteorol. Mag.*, **122**, 81–94.

Davies, T.D., M.J.P. Cullen, A.J. Malcolm, et al., 2005. A new dynamical core for the Met Office's global and regional modelling of the atmosphere. *Q. J. R. Meteorol. Soc*., **131**, 1759–1782.

Dharssi, I., A.C. Lorenc and N.B. Ingleby, 1992. Treatment of gross errors using maximum probability theory. *Q. J. R. Meteorol. Soc.*, **118**, 1017–1036.

Fisher, M., 2003. Background error covariance modelling. *Recent Developments in Data Assimilation for Atmosphere and Ocean*, ECMWF Seminar proceedings, Reading, UK, pp 45–63.

Gandin, L., 1963. Objective analysis of meteorological fields. *Gidromet.*, Leningrad (English translation Israel Program for Scientific Translation, Jerusalem, 1965).

Gauthier, P. and J.-N. Thépaut, 2001. Impact of the digital filter as a weak constraint in the pre-operational 4D-Var assimilation system of Météo-France. *Mon. Weather Rev.,* **129**, 2089–2102.

Gilchrist, B. and G. Cressman, 1954. An experiment in objective analysis. *Tellus*, **6**, 309–318.

Hagedorn, R., T.M. Hamill and J.S. Whitaker, 2008. Probabilistic forecast calibration using ECMWF and GFS ensemble reforecasts. Part 1: Two-meter temperatures. *Mon. Weather Rev*., **136**, 2608–2619.

Hamill, T.M. and C. Snyder, 2000. A hybrid ensemble Kalman filter – 3D variational analysis scheme. *Mon. Weather Rev*., **128**, 2905–2919.

Hamill, T.M., J.S. Whitaker and S.L. Mullen, 2006. Reforecasts, an important dataset for improving weather predictions. *Bull. Amer. Meteorol. Soc.*, **87**, 33–46.

Hoffman, R.N. and E. Kalnay, 1983. Lagged average forecasting, an alternative to Monte Carlo forecasting. *Tellus*, **35A**, 100–118.

Houtekamer, P.L. and H.L. Mitchell, 2005. Ensemble Kalman Filtering. *Q. J. R. Meteorol. Soc.*, **131**, 3269–3289.

Johnson, C. and R. Swinbank, 2009. Medium-range multi-model ensemble combination and calibration. *Q. J. R. Meteorol. Soc.*, **135**, 777–794.

Kalnay, E., 2003. *Atmospheric Modeling, Data Assimilation and Predictability*, Cambridge University Press, Cambridge, 364pp.

Kalnay E., M. Kanamitsu, R. Kistler, et al., 1996. The NCEP/NCAR 40-year reanalysis project. *Bull. Amer. Meteorol. Soc.*, **77**, 437–471.

Le Dimet, F-X. and O. Talagrand, 1986. Variational algorithms for analysis and assimilation of meteorological observations: Theoretical Aspects. *Tellus*, **38A**, 97–110.

Leith, C.E., 1974. Theoretical skill of Monte Carlo forecasts. *Mon. Weather Rev.*, **102**, 409–418.

Lorenz, E.N., 1963. Deterministic non-periodic flow. *J. Atmos. Sci.*, **20**, 130–141.

Lorenc, A.C., 1981. A global three-dimensional multivariate statistical analysis scheme. *Mon. Weather Rev.*, **109**, 701–721.

Lorenc, A.C., 1986. Analysis methods for numerical weather prediction. *Q. J. R. Meteorol. Soc.,***112**, 1177–1194.

Lorenc, A.C., S.P. Ballard, R.S. Bell, N.B. Ingleby, P.L.F. Andrews, D.M. Barker, J.R. Bray, A.M. Clayton, T. Dalby, D. Li, T.J. Payne and F.W. Saunders, 2000. The Met. Office global 3-dimensional variational data assimilation scheme. *Q. J. R. Meteorol. Soc.,* **126**, 2991–3012.

Lorenc, A.C., R.S. Bell and B. Macpherson, 1991. The meteorological office analysis correction data assimilation scheme. *Q. J. R. Meteorol. Soc.,***117**, 59–89.

Lorenc, A.C. and O. Hammon, 1988. Objective quality control of observations using Bayesian methods – Theory, and a practical implementation. *Q. J. R. Meteorol. Soc.*, **114**, 515–543.

Lynch, P. and X.-Y. Huang, 1992. Initialization of the HIRLAM model using a digital filter. *Mon. Weather Rev.,* **120**, 1019–1034.

Lyne, W.H., R. Swinbank and N.T. Birch, 1982. A data assimilation experiment, with results showing the atmospheric circulation during the FGGE special observing periods. *Q. J. R. Meteorol. Soc.,* **108**, 575–594.

Machenhauer, B., 1977. On the dynamics of gravity oscillations in a shallow water model with applications to normal mode initialization. *Contrib. Atmos. Phys.*, **50**, 253–271.

Matsueda, M. and H.L Tanaka, 2008. Can MCGE outperform the ECMWF ensemble? *SOLA*, **4**, 77–80. doi:10.2151/sola.2008-020.

Ott, E., B.R. Hunt, I. Szunyogh, A.V. Zimin, E.J. Kostelich, M. Corazza, E. Kalnay, D.J. Patil and J.A. Yorke, 2004. A local ensemble Kalman filter for atmospheric data assimilation. *Tellus*, **56A**, 415–428.

Panovsky, H., 1949. Objective weather-map analysis. *J. Appl. Meteor*, **6**, 386–392.

Park, Y.-Y., R. Buizza and M. Leutbecher, 2008. TIGGE: Preliminary results on comparing and combining ensembles. *Q. J. R. Meteorol. Soc*., **134**, 2051–2066. Also published as ECMWF Technical Memorandum No. 548.

Parrish, D.F. and J.C. Derber, 1992. The National Meteorological Center's spectral statistical interpolation analysis scheme. *Mon. Weather Rev.*, **120**, 1747–1763.

Purser, R.J., 1999. Non-standard grids. *Recent Developments in Numerical Methods for Atmospheric Modelling*, ECMWF seminar 7–11 September 1998, pp 44–72.

Richardson, L., 1922. *Weather Prediction by Numerical Process*. Cambridge University Press, Cambridge.

Rodgers, C.D., 2000. *Inverse Methods in Atmospheric Sounding: Theory and Practice*. World Scientific, Singapore, 238pp.

Shutts, G., 2005. A kinetic backscatter algorithm for use in ensemble prediction systems. *Q. J. R. Meteorol. Soc.*, **131**, 3079–3102.

Swinbank, R. and R.J. Purser, 2006. Fibonacci grids: A novel approach to global modelling. *Q. J. R. Meteorol. Soc.*, **132**, 1769–1793.

Szunyogh, I., Z. Toth, R.E. Morss, S.J. Majumdar, B.J. Etherton and C.H. Bishop, 2000. The effect of targeted dropsonde observations during the 1999 winter storm reconnaissance program. *Mon. Weather Rev.*, **128**, 3520–3537.

Toth, Z. and E. Kalnay, 1993. Ensemble forecasting at NMC – the generation of perturbations. *Bull. Amer. Meteorol. Soc.*, **74**, 2317–2330.

Trémolet, Y., 2007. Model-error estimation in 4D-Var. *Q. J. R. Meteorol. Soc.*, **133**, 1267–1280.

Wang, X., D.M. Barker, C. Snyder and T.M. Hamill, 2008a. A hybrid ETKF-3DVAR data assimilation for the WRF model. Part I: Observing system simulation experiment. *Mon. Weather Rev.*, **136**, 5116–5131.

Wang, X., D.M. Barker, C. Snyder and T.M. Hamill, 2008b. A hybrid ETKF-3DVAR data assimilation for the WRF model. Part II: Real observation experiments. *Mon. Weather Rev.*, **136**, 5132–5147.

Wang, X. and C.H. Bishop, 2003. A comparison of Breeding and Ensemble Transform Kalman Filter Ensemble Forecast Schemes. *J. Atmos. Sci.*, **60**, 1140–1158.

Wu, W-S., R.J. Purser and D.F. Parrish, 2002. Three-dimensional variational analysis with spatially inhomogeneous covariances. *Mon. Weather Rev.*, **130**, 2905–2916.

# Part IV
# Atmospheric Chemistry

# Introduction to Atmospheric Chemistry and Constituent Transport

**Valery Yudin and Boris Khatattov**

## 1 Importance of Chemistry

Atmospheric photochemical processes often occurring at altitudes of tens of kilometres above the surface can be of paramount importance to the existence of life on Earth. Formed by complex chemical and photodissociation processes, the ozone layer absorbs harmful ultraviolet radiation in the stratosphere before it reaches the Earth surface. The deoxyribose nucleic acid molecules (DNAs) of most organisms absorb very strongly at wavelengths around 300 nm. Had this radiation not been prevented from reaching the ground, it would have caused immediate and significant tissue damage and led to formation of cancer cells and genetic mutations.

Knowing what processes control formation and destruction of ozone molecules is important for monitoring and predicting changes in ozone abundances. The springtime "ozone hole" phenomenon over the Antarctic continent discovered in the 1980s clearly demonstrated how human (anthropogenic) activities can destroy the natural chemical balance in the atmosphere and potentially lead to disastrous consequences.

It took years of scientific studies and debates to discover the complete chain of related physical effects and chemical reactions and prove that the dramatic rapid destruction of ozone was originally caused by industrial emissions of chlorofluorocarbons. Chlorine and bromine radicals released in the process of photodissociation of chlorofluorocarbons act as catalysts in fast ozone loss cycles. This property combined with persistent patterns of atmospheric circulation, very cold temperatures in the Antarctic, and absence of solar radiation during the long Antarctic winter eventually causes almost complete destruction of October ozone in the lower stratosphere, where most of the ozone resides. Recently, wintertime mini ozone holes were observed in the Northern Hemisphere above western Europe and Russia.

V. Yudin (✉)
SAIC, Global Modeling Assimilation Office, Code 610.1, Goddard Space Flight Center, Greenbelt, MD 20771, USA; Atmospheric Chemistry Division, National Center for Atmospheric Research, Boulder, CO, USA
e-mail: vyudin@ucar.edu

W. Lahoz et al. (eds.), *Data Assimilation*, DOI 10.1007/978-3-540-74703-1_16,

The risk of global warming highlights another strong link between concentrations of atmospheric trace gases and global environmental conditions. Some atmospheric gases, for instance, $CO_2$ and $H_2O$, trap radiation emitted by the Earth's surface. Increased concentrations of these gases are likely to lead to temperature increases in the troposphere since normally this radiation would have escaped to space. While the direct proof of a relationship between recent increases in atmospheric carbon dioxide and observed global temperature trends is not easy due to the large natural variability of temperature records, there exist enough scientific evidence and modelling studies pointing to this connection beyond reasonable doubt.

The pollution of air near the surface and in the troposphere is another illustration of the importance of atmospheric chemistry to our well-being. Carbon monoxide, nitrogen species, complex organic compounds and small particulate matter contained in car exhaust and by-products of industrial production and incomplete combustion lead, through a complex chain of chemical transformations, to the formation of smog, acid rain, and increased amounts of tropospheric ozone. While stratospheric ozone shields us from harmful radiation, the increased levels of ozone in the troposphere can lead to complications in patients with cardio-vascular diseases and, in some cases, increased rates of hospital admissions and mortalities recorded the next day.

Chains of chemical transformations leading to formation of a particular constituent are often very long and complex. This is particularly true for the chemistry of tropospheric pollutants chemistry, where chemical interactions between hundreds and thousands of compounds should be monitored. Modelling and understanding these processes requires significant resources and high quality observations with global coverage. The framework of chemical data assimilation (see chapter *Constituent Assimilation*, Lahoz and Errera) can facilitate this task by uncovering and making use of relationships between observed and simulated quantities in a mathematically consistent and rigorous fashion. In the following sections we will give a brief overview of elementary photochemistry and transport of atmospheric constituents chemistry and related observations. A much more detailed presentation of chemistry and dynamics in the whole atmosphere can be found, for instance, in the textbooks of Brasseur et al. (1999) and Brasseur and Solomon (2000).

## 2 Atmospheric Processes Affecting the Composition

Interactions between thermodynamics, radiation and chemistry on various spatial and temporal scales define distributions of the radiatively and chemically active species and their variability. Figure 1 schematically illustrates processes that affect the atmospheric composition in the boundary layer, free troposphere, and stratosphere, from the Equator to the Pole. The separation boundaries marked by the height of the planetary boundary layer (PBL) and the location of the tropopause are relatively transparent and their locations vary on different time-scales, exhibiting diurnal, seasonal and sudden (event driven) variations. Penetration and mixing of stratospheric and tropospheric air masses, ventilation of boundary layer by vertical convection, and tropical and polar air intrusions to mid latitudes create

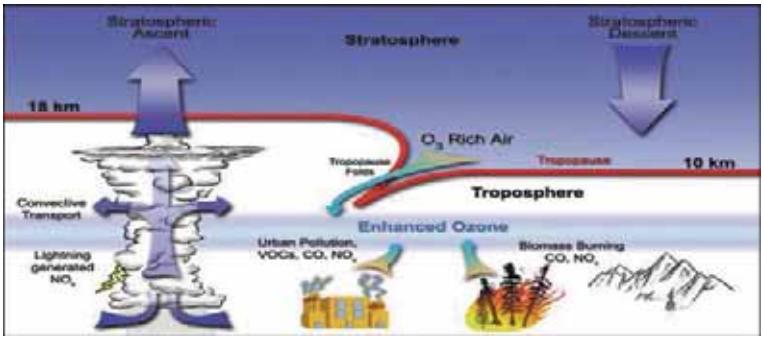



**Fig. 1** Schematic illustration of transport of constituents from the surface to the free atmosphere, and stratosphere-troposphere air mass exchanges. Source: EOS Aura website

variable chemical weather events on the global and regional scales from the surface to the stratopause. Although the description of interactions between chemistry, radiation and dynamics is vital for atmospheric composition studies, in this introduction we will discuss separately the chemical and transport processes near the surface, in the free troposphere, and in the stratosphere. Next, we illustrate the separation boundaries characterizing the chemistry and transport across the tropopause, highlighting the penetration of pollutants from natural and anthropogenic surface emissions.

## *2.1 Elementary Chemical Processes*

The Earth's atmosphere can be thought of as a combustion system where the energy of the Sun drives a variety of chemical transformations. The composition of the atmosphere is determined by complex chemical mechanisms. Each mechanism consists of a few to sometimes hundreds of elementary chemical reactions. For example, the photolysis of $O_2$ is responsible for initiating the chemistry involved in the production of ozone, $O_3$, in the stratosphere:

$$O_2 + h\nu \longrightarrow O + O$$
$$O_2 + O + M \longrightarrow O_3 + M$$

The process of absorption of a photon by a molecule results in a change in the energy level of the molecule. In this process the photon disappears. Photons come in different "colours" or frequencies corresponding to different energies. "Blue" photons have more energy than "yellow" and more energy than "red". High energy photons can break up molecules; this process is called photodissociation (photolysis). Examples of photolysis reactions are shown below:

$$O_2 + h\nu \longrightarrow O + O$$
$$O_3 + h\nu + \longrightarrow O_2 + O$$
$$NO_2 + h\nu \longrightarrow NO + O$$

Rates of the photodissociation reactions depend on the amount of sunlight (number of photons) and the absorption cross-section of the molecule that absorbs photons. These rates are often called photodissociation coefficients or photolysis rates, *J*. Details of the calculation of these rates can be found in Brasseur et al. (1999). For this brief introduction, it is important to note that the rate of change of a particular chemical due to photodissociation is directly proportional to the corresponding photolysis rate multiplied by the chemical's concentration:

$$\frac{d[\mathrm{O_3}]}{dt} = -J_{\mathrm{O_3}} \cdot [\mathrm{O_3}]$$

The most common reactions between atmospheric chemicals are bimolecular (by number of reagents) reactions of the type

$$\mathrm{AB} + \mathrm{C} \longrightarrow \mathrm{A} + \mathrm{BC}$$

An example of such reaction is

$$\mathrm{NO} + \mathrm{O_3} \longrightarrow \mathrm{NO_2} + \mathrm{O_2}$$

In such a reaction, the rate of disappearance of reagents, equal to the rate of appearance of the products, is

$$\frac{d[\mathrm{AB}]}{dt} = \frac{d[\mathrm{C}]}{dt} = -\frac{d[\mathrm{A}]}{dt} = -\frac{d[\mathrm{BC}]}{dt} = k \cdot [\mathrm{AB}] \cdot [\mathrm{C}]$$

$k$ in this equation is called the reaction rate. This coefficient is usually a strong function of temperature:

$$k = \mathrm{A} \cdot \exp(-\Delta \mathrm{E}/\mathrm{RT})$$

Another common type of atmospheric chemical reaction is the trimolecular reaction of the type

$$\mathrm{A} + \mathrm{B} + \mathrm{C} \longrightarrow \text{products}$$

Some examples are

$$\mathrm{NO} + \mathrm{NO} + \mathrm{O_2} \longrightarrow \mathrm{NO_2} + \mathrm{NO_2}$$
$$\mathrm{O_2} + \mathrm{O} + \mathrm{N_2} \longrightarrow \mathrm{O_3} + \mathrm{N_2}$$

Similarly to bimolecular reactions, the rate of disappearance of reagents (and appearance of products) is given by

$$\frac{d[\mathrm{A}]}{dt} = \frac{d[\mathrm{B}]}{dt} = \frac{d[\mathrm{C}]}{dt} = -\frac{d[\text{products}]}{dt} = k \cdot [\mathrm{A}] \cdot [\mathrm{B}] \cdot [\mathrm{C}]$$

Chemical transformations in the Earth atmosphere must be considered together with the influence of the movement of air masses in the vertical and horizontal directions, turbulent diffusion and molecular mixing of molecules, and the dependence of chemical reactions rates on temperature and pressure. The next sections review chemistry and transport in the stratosphere and troposphere, highlighting the current research tendency to understand and constrain uncertainties in atmospheric photochemistry by means of models and observations.

## *2.2 Stratospheric Chemistry*

The concentration of ozone in the stratosphere is determined by a balance between its production and losses. In a purely oxygen (simplified) atmosphere, processes controlling ozone concentration are:

$$O_2 + h\nu \longrightarrow O + O \qquad (J_1)$$

$$O_2 + O + M \longrightarrow O_3 + M \qquad (k_1)$$

$$O_3 + h\nu \longrightarrow O_2 + O \qquad (J_2)$$

$$O + O_3 \longrightarrow O_2 + O_2 \qquad (k_2)$$

In the three body reaction ($k_1$), M stands for any inert molecule such as $O_2$ or $N_2$. The coupled system of two differential equations that describes the time evolution of ozone and atomic oxygen can be written as:

$$\frac{d[O]}{dt} = 2 \cdot J_1 \cdot [O_2] - k_1 \cdot [O][O_2] + J_2[O_3] - k_2 \cdot [O][O_3]$$

$$\frac{d[O_3]}{dt} = k_1 \cdot [O][O_2] - J_2[O_3] - k_2 \cdot [O][O_3]$$

For given initial conditions (concentrations of ozone and atomic oxygen at time $t = t_0$), this system of first order differential equations can be solved to monitor $[O_3]$ and [O] as a function of time. In reality, the concentration of ozone is directly and indirectly affected by many other chemicals. Particularly important are the catalytic ozone destruction cycles of the following form:

$$X + O_3 \longrightarrow XO + O_2$$

$$\underline{O + XO \longrightarrow O_2 + X}$$

$$\text{Overall: } O + O_3 \longrightarrow O_2 + O_2$$

In this cycle, chemical X leads to destruction of one ozone molecule without being destroyed itself. As a result, a single molecule X can destroy a very large number of $O_3$ molecules. In the stratosphere some of the most important catalytic ozone loss cycles involve NO, Cl, and OH in place of X, for example:

$$NO + O_3 \longrightarrow NO_2 + O_2$$

$$\underline{O + NO_2 \longrightarrow O_2 + NO}$$

$$\text{Overall: } O + O_3 \longrightarrow O_2 + O_2$$

**Table 1** A typical set of stratospheric photochemical reactions

| | |
|---|---|
| k001: $N_2O_5 + H_2O(a) \rightarrow 2^*HNO_3$; | k042: $HCl + OH \rightarrow H_2O + Cl$; |
| k002: $O + O_3 \rightarrow 2^*O_2$; | k043: $HCl + O \rightarrow OH + Cl$; |
| k003: $O(^1D) + O_3 \rightarrow 2^*O_2$; | k044: $HOCl + OH \rightarrow H_2O + ClO$; |
| k004: $O(^1D) + N_2 \rightarrow O + N_2$; | k045: $ClONO_2 + O \rightarrow ClO + NO_3$; |
| k005: $O(^1D) + O_2 \rightarrow O + O_2$; | k046: $ClONO_2 + OH \rightarrow HOCl + NO_3$; |
| k006: $O(^1D) + H_2O \rightarrow 2^*OH$; | k047: $ClO + NO_2 + M \rightarrow ClONO_2 + M$; |
| k007: $O(^1D) + H_2 \rightarrow H + OH$; | k050: $NO_2 + O \rightarrow NO + O_2$; |
| k008: $O(^1D) + CH_4 \rightarrow OH + CH_3$; | k051: $NO + O_3 \rightarrow NO_2 + O_2$; |
| k009: $O + O_2 + M \rightarrow O_3 + M$; | k052: $NO + HO_2 \rightarrow NO_2 + OH$; |
| k016: $OH + CO \rightarrow CO_2 + H$; | k053: $NO_2 + O_3 \rightarrow NO_3 + O_2$; |
| k017: $CH_4 + OH \rightarrow CH_3 + H_2O$; | k054: $HNO_3 + OH \rightarrow NO_3 + H_2O$; |
| k019: $H_2 + OH \rightarrow H_2O + H$; | k055: $HNO_4 + OH \rightarrow H_2O + O_2 + NO_2$; |
| k020: $H + O_3 \rightarrow O_2 + OH$; | k057: $NO_2 + OH + M \rightarrow HNO_3 + M$; |
| k021: $H + HO_2 \rightarrow 2^*OH$; | k058: $NO_2 + HO_2 + M \rightarrow HNO_4 + M$; |
| k022: $OH + O \rightarrow O_2 + H$; | k059: $NO_3 + NO_2 + M \rightarrow N_2O_5 + M$; |
| k023: $OH + O_3 \rightarrow O_2 + HO_2$; | k060: $N_2O_5 + M \rightarrow NO_2 + NO_3 + M$; |
| k024: $OH + OH \rightarrow H_2O + O$; | k061: $HNO_4 + M \rightarrow HO_2 + NO_2 + M$; |
| k025: $OH + HO_2 \rightarrow H_2O + O_2$; | j001: $O_2 \rightarrow 2^*O$ |
| k026: $HO_2 + O_3 \rightarrow 2^*O_2 + OH$; | j002: $O_3 \rightarrow O_2 + O$ |
| k027: $HO_2 + O \rightarrow O_2 + OH$; | j003: $O_3 \rightarrow O_2 + O(^1D)$ |
| k028: $HO_2 + HO_2 \rightarrow H_2O_2 + O_2$; | j004: $HO_2 \rightarrow O + OH$ |
| k029: $H_2O_2 + OH \rightarrow H_2O + HO_2$; | j005: $H_2O_2 \rightarrow 2^*OH$ |
| k030: $H + O_2 + M \rightarrow HO_2 + M$; | j006: $NO_2 \rightarrow NO + O$ |
| k031: $Cl + O_3 \rightarrow ClO + O_2$; | j007: $NO_3 \rightarrow NO_2 + O$ |
| k032: $Cl + CH_4 \rightarrow HCl + CH_3$; | j008: $NO_3 \rightarrow NO + O_2$ |
| k033: $Cl + H_2 \rightarrow H + HCl$; | j009: $N_2O_5 \rightarrow NO_2 + NO_3$ |
| k034: $Cl + HO_2 \rightarrow O_2 + HCl$; | j010: $HNO_3 \rightarrow OH + NO_2$ |
| k035: $Cl + HO_2 \rightarrow OH + ClO$; | j011: $HNO_4 \rightarrow OH + NO_3$ |
| k036: $Cl + H_2O_2 \rightarrow HO_2 + HCl$; | j012: $HNO_4 \rightarrow HO_2 + NO_2$ |
| k038: $ClO + O \rightarrow Cl + O_2$; | j016: $HOCl \rightarrow OH + Cl$ |
| k039: $ClO + NO \rightarrow Cl + NO_2$; | j017: $ClONO_2 \rightarrow Cl + NO_3$ |
| k040: $ClO + OH \rightarrow HO_2 + Cl$; | j018: $ClONO_2 \rightarrow Cl + NO_2 + O$ |
| k041: $ClO + HO_2 \rightarrow HOCl + O_2$; | j026: $HCl \rightarrow H + Cl$ |

Table 1 presents an example of a set of photochemical reactions that one usually needs to take into account when modelling stratospheric chemistry with moderate accuracy. Several important cycles (sulphur, bromine, iodine and heterogeneous chemistry) are beyond these limited set of 87 reactions. Table 2 illustrates a set of non-linear chemical equations corresponding to the core set of the stratospheric reactions given in Table 1 above; for more details see Khattatov et al. (1999).

## *2.3 Tropospheric Chemistry*

Tropospheric chemistry is considered to be the next theoretical and experimental frontier in the understanding and prediction of Earth's atmospheric composition and climate. To advance this discipline over the foreseeable future will be a great

**Table 2** A system of coupled ODEs corresponding to reactions in Table 1

[1:] d[H ]/dt = j026*HCl + k007*O($^1$D)*$H_2$ + k016*OH*CO + k019*$H_2^*$OH − k020*H*$O_3$ − k021*H*$HO_2$ + k022*OH*O − k030*H*$O_2^*$M + k033*Cl*$H_2$

[2:] d[OH ]/dt = j004*$HO_2$ + 2*j005*$H_2O_2$ + j010*$HNO_3$ + j011*$HNO_4$ + j016*HOCl+ 2*k006*O($^1$D)*$H_2O$ + k007*O($^1$D)*$H_2$ + k008*O($^1$D)*$CH_4$ − k016*OH*CO − k017*$CH_4^*$OH − k019*$H_2^*$OH + k020*H*$O_3$ + 2*k021*H*$HO_2$ − k022*OH*O − k023*OH*$O_3$ − 2*k024*OH*OH − k025*OH*$HO_2$ + k026*$HO_2^*O_3$ + k027*$HO_2^*$O − k029*$H_2O_2^*$OH + k035*Cl*$HO_2$ − k040*ClO*OH − k042*HCl*OH + k043*HCl*O − k044*HOCl*OH -k046*$ClONO_2^*$OH + k052*NO*$HO_2$ − k054*$HNO_3^*$O − k055*$HNO_4^*$OH − k057*$NO_2$OH*M

. . ..
. . ..
. . ..

[15:] d[$N_2O_5$ ]/dt = −j009*$N_2O_5$ − k001*$N_2O_5^*H_2O$(a) + k059*$NO_3^*NO_2^*$M − k060*$N_2O_5^*$M

[16:] d[O ]/dt = 2*j001*$O_2$ + j002*$O_3$ + j004*$HO_2$ + j006*$NO_2$ + j007*$NO_3$+ j018*$ClONO_2$ − k002*O*$O_3$ + k004*O($^1$D)*$N_2$ + k005*O($^1$D)*$O_2$ − k009*O*$O_2^*$M − k022*OH*O + k024*OH*OH − k027*$HO_2^*$O − k038*ClO*O − k043*HCl*O − k045*$ClONO_2^*$O − k050*$NO_2^*$O

[17:] d[O($^1$D) ]/dt = j003*$O_3$ − k003*O($^1$D)$O_3$ − k004*O($^1$D)*$N_2$ − k005*O($^1$D)*$O_2$ − k006*O($^1$D)*$H_2O$ − k007*O($^1$D)*$H_2$ − k008*O($^1$D)*$CH_4$

[18:] d[$O_3$ ]/dt = −j002*$O_3$ − j003*$O_3$ − k002*O*$O_3$ − k003*O($^1$D)*$O_3$ + k009*O*$O_2^*$M − k020*H*$O_3$ − k023*OH*$O_3$ − k026*$HO_2^*O_3$ − k031*Cl*$O_3$ − k051*NO*$O_3$ − k053*$NO_2^*O_3$

challenge for atmospheric science research because global and regular observations of tropospheric species are needed to constrain their highly uncertain budgets, and investigate mechanisms of chemical transformations of inorganic and organic compounds. In the troposphere, key monitored trace gas species include $H_2O$, $O_3$, OH, $NO_x$ (= NO + $NO_2$), CO, some important hydrocarbons, and aerosols. The chemical transformations of these species are also affected by the presence of clouds, and rain and snow.

Tropospheric $O_3$ and $H_2O$ define the oxidation of many species through the hydroxyl radical (OH). Its primary source in the troposphere is the photodissociation of $O_3$ (hν <411 nm) followed with reaction with $H_2O$:

$$O_3 + h\nu \longrightarrow O_2 + O(^1D)[h\nu < 411\,\text{nm}]$$
$$H_2O + O(^1D) \longrightarrow 2OH$$

Only a small portion of the O($^1$D) reacts with $H_2O$. The quenching of O($^1$D) by inert air molecules returns it back to atomic oxygen O($^3$P), and after recombination of $O_2$ and O($^3$P), the restoration of $O_3$ is accomplished. There are two major chains of tropospheric $O_3$ production controlled by OH. In the presence of nitrogen oxides, the oxidation of hydrocarbons can lead to $O_3$ production through oxidation of CO:

$$\begin{aligned}
&CO + OH + O_2 \longrightarrow CO_2 + HO_2\\
&HO_2 + NO \longrightarrow OH + NO_2\\
&NO_2 + h\nu \longrightarrow NO + O\\
&\underline{O + O_2 + M \longrightarrow O_3 + M}\\
&\text{Overall}: \; CO + 2O_2 \longrightarrow CO_2 + O_3
\end{aligned}$$

As seen above, the concentrations of OH, $HO_2$ and NO + $NO_2$ are not affected by this cycle of CO oxidation. To simplify the study of stratospheric chemistry, the odd nitrogen and hydrogen families ($NO_x$ = NO + $NO_2$ and $HO_x$ = OH + $HO_2$, respectively) can be introduced. This helps to increase effective chemical lifetimes (e.g. $NO_x$ is longer lived than NO or $NO_2$) and relax the requirements for numerical schemes employed in the solution of chemical equations. OH is involved in the key reaction that describes loss of $NO_x$:

$$OH + NO_2 + M \longrightarrow HNO_3 + M$$

Nitric acid ($HNO_3$) in the troposphere is a soluble substance; it is only moderately reactive. This non-reactive reservoir ($HNO_3$) has a high likelihood of being removed from the atmosphere before it can be modified to release the $NO_x$ radicals, NO and $NO_2$.

Both $HO_x$ radicals (OH and $HO_2$) react directly with $O_3$:

$$\begin{aligned}
&HO_2 + O_3 \longrightarrow OH + 2O_2\\
&OH + O_3 \longrightarrow HO_2 + O_2
\end{aligned}$$

The net production of OH defined by the above chemical reactions depends on the availability of $NO_x$ in the troposphere. Production of $NO_x$ by industrial processes and lightning is an important factor in controlling the distributions of nitrogen oxides. Under high $NO_x$ loading, the smog cycle of ozone production (Crutzen 1974) is an effective link in the following chain of photochemical transformations in the troposphere:

$$\begin{aligned}
&CH_4 + OH \longrightarrow CH_3 + H_2O\\
&CH_3 + O_2 + M \longrightarrow CH_3O_2 + M\\
&CH_3O_2 + NO \longrightarrow CH_3O + NO_2\\
&CH_3O + O_2 \longrightarrow HCHO + HO_2\\
&HO_2 + NO \longrightarrow OH + NO_2\\
&NO_2 + h\nu \longrightarrow NO + O\\
&\underline{O + O_2 + M \longrightarrow O_3 + M}\\
&\text{Overall}: \; CH_4 + 4O_2 \longrightarrow HCHO + H_2O + 2O_3
\end{aligned}$$

The last net reaction is catalysed by both the $HO_x$ and $NO_x$ families. Other reactions, including those in the developing branch of inorganic chemistry, can

be also important to close the budgets of key species and pollutants in the troposphere (Brasseur et al. 1999). Constraining tropospheric ozone and related tracers using observations is a key task of current and future intensive observational campaigns and space environmental missions. The external sources of nitrogen compounds, hydrocarbons, CO, and $CO_2$, are still the most uncertain parameters in the chemistry-transport models (CTMs) that aim to forecast chemical weather. The combination of land data and atmospheric monitoring of species from space is considered to provide a promising database for the optimization of chemical weather models; such monitoring aims to include the strengths, locations and temporal variations of the constituent concentrations and their surface emissions.

## *2.4 Surface Emissions*

Recently, understanding of regional surface concentrations and sources of pollution has been greatly advanced by the global monitoring of tropospheric constituents from space. Multi-year space-borne CO, $CO_2$, $CH_4$ and $NO_2$ retrievals provide a powerful database for the inverse studies that adjust surface boundary conditions in the CTMs (see chapter *Inverse Modelling and Combined State-source Estimation for Chemical Weather*, Elbern et al.).

For optimization of emissions, statistical estimation schemes operate with misfits of observed minus simulated concentrations to invert the strength of surface sources required to simulate the observed concentrations in the atmosphere. The natural surface sources are largely related to sudden biomass burning and wildfire events. Their locations and timing can be identified by space-borne monitoring of burnt areas and fire counts (Giglio et al. 2006). Anthropogenic activities, including industrial, agricultural and biofuel emissions, can also provide a substantial impact on the total net surface fluxes of pollutants. The simultaneous optimization of multi-species surface fluxes is currently considered to be the way to provide a consistent adjustment of correlated emissions in chemical models of the troposphere (van der Werf et al. 2006).

The textbook of Enting (2002) provides a good introduction to the practical formulation and solution of the inverse problems associated with the estimation of regional and global sources of atmospheric pollution. Depending on data sources, techniques for estimation (i.e., modelling) of the surface emissions can be separated into two types: “top–down” and “bottom–up” algorithms. The “bottom–up” schemes use land surface data, and international reports of the environmental protection agencies and committees. The bottom–up schemes employ empirical relationships and models to estimate surface emissions by adding together and compiling various sources of information. The bottom–up studies provide the strength and geographical distribution of the surface characteristics of pollutants on monthly and annual scales. These data for major pollutants form the basis of emission inventories. These inventories represent the critical input for modelling and forecasting pollution budgets and the evolution of pollution plumes. They can be viewed as the background field for the top–down optimization methods that aim to adjust

averaged sources by inserting specific observations typical for a given year, month, or a shorter time interval.

Figure 2 shows an example of the biomass burning emissions of CO used in the MOZART (Model for OZone And Related Tracers) CTM (Horowitz et al. 2003) for April–May 2000–2003, adjusted using the monthly MODIS (MODerate resolution Imaging Spectroradiometer) fire counts (Giglio et al. 2006). An adjustment in the original climate inventory of CO sources has been made to introduce observed year-to-year variations in CO surface fluxes and simulate interannual CO variability in the free atmosphere observed by the MOPITT (Measurements Of Pollution In The Troposphere) instrument (Fig. 3). (Acronyms are given in the *Appendix*.)

To predict accurately the observed distribution of pollutants, the joint chemical assimilation of observed constituents in the free atmosphere and inverse optimization of corresponding surface sources are necessary, especially for urban and industrial areas (see chapter *Inverse Modelling and Combined State-source Estimation for Chemical Weather*, Elbern et al.). Assimilation of observations without correction of systematic model errors associated with surface emissions can violate the assumption of unbiased errors (see chapter *Mathematical Concepts of Data Assimilation*, Nichols). Misspecification of surface concentrations or fluxes of pollutants in models leads to systematic errors in their forecast. The capability to suppress forecast biases related to errors in surface emissions before (or in the course of) assimilation can greatly enhance the quality of combined model-data analysis, and provide consistent optimization of concentrations and emissions. The practical solution of the generalized inverse or combined source-state estimation problem depends on the quality of data, forecast uncertainty and model formulations. Using

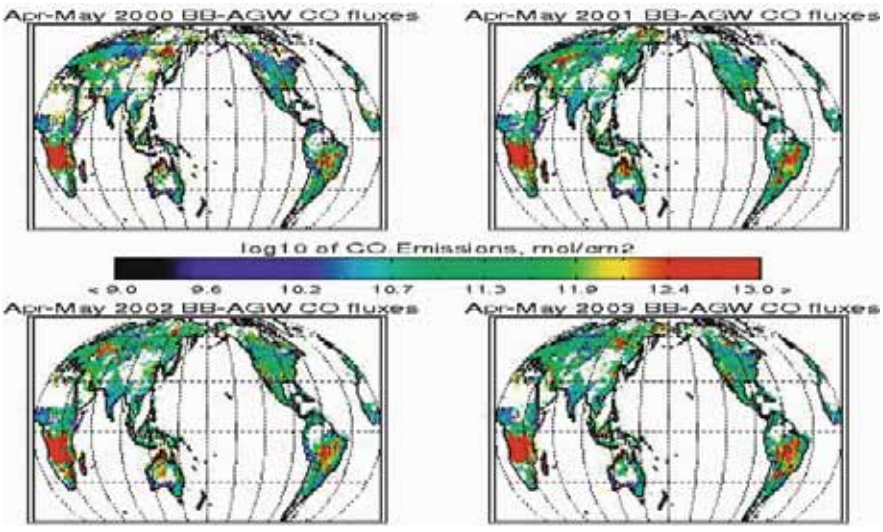



**Fig. 2** The biomass burning emissions of CO used in the chemistry transport model for April–May 2000–2003 adjusted using the monthly MODIS/Terra fire counts (Giglio et al. 2006). *Red* indicates relatively high values; *blue* indicates relatively low values

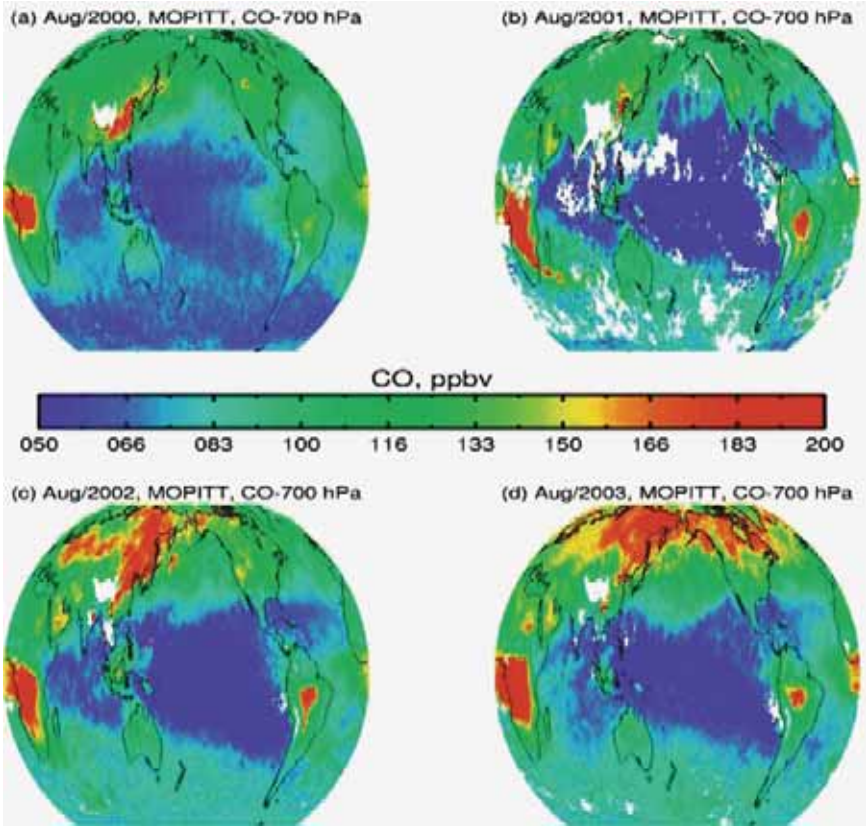



**Fig. 3** Interannual CO variability in the free atmosphere (700 hPa) observed by the MOPITT instrument during August for the years 2000 (*top left*), 2001 (*top right*), 2002 (*bottom left*) and 2003 (*bottom left*). *Red* indicates relatively high values; *blue* relatively low values

space-borne constituent observations from different instruments for inversion can result in data-source dependent surface fluxes for a given model. In turn, for a given dataset the different formulations of models (transport, chemistry and numerics) can also produce a spread in the solutions to the inverse problem. Characterization of uncertainties of satellite measurements, evaluation of systematic data discrepancies and model sensitivity to tunable parameters, can provide realistic error estimates of the joint state-source estimation results.

The pioneering space-borne measurements of $NO_2$, $CH_4$, and CO from the Envisat, Terra, Aqua and Aura missions (see chapter *Research Satellites*, Lahoz), initiated a number of model-data analysis studies that sought to optimize uncertainties in the emissions of these observed species. Figure 4 illustrates how the top–down optimization of surface emissions using MOPITT CO retrievals improved MOZART CTM simulations during the TRACE-P observational campaign in March 2001 over the coastline of South Asia (Petron et al. 2004). With these optimized emissions incorporated into the CTM, the data assimilation of MOPITT CO

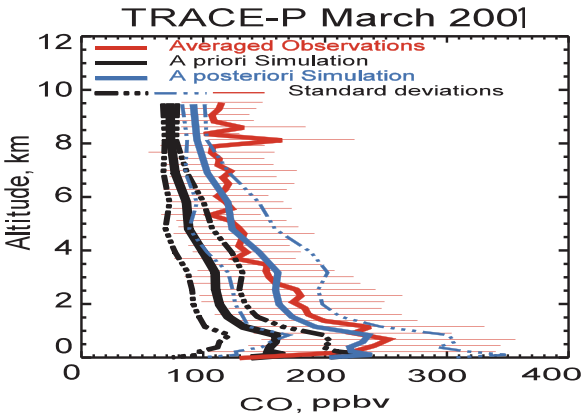



**Fig. 4** Example of observed (*red*) and simulated (*black* and *blue*) CO during the TRACE-P observational campaign in March 2001. The two simulations correspond to two types of CO surface emissions: the EDGAR-2 emission inventory (*black*); and surface CO fluxes optimized using the monthly averaged MOPITT/Terra CO retrievals (*blue*)

retrievals can be substantially improved. The study of Yudin et al. (2004) demonstrated the importance of constraining CO emissions before the assimilation of MOPITT data into the CTM.

The importance of year-dependent optimization of emissions in models is clearly demonstrated by year-to-year variations of CO constrained by MOPITT measured radiances for the period 2000–2008. In practice, the top–down estimates of surface emissions from systematic differences between simulated and observed concentrations in the mid troposphere will depend on the a priori distribution of the surface fluxes created by the land-surface products (fire counts, burned area products, etc.).

Several environmental models have been recently used to update the estimates of surface fluxes (Randerson et al. 2004) and evaluate uncertainties in bottom–up emission inventories. Recent studies show that these inventories still have a large level of uncertainty. To validate these emissions, model simulations of pollutants with proposed surface emissions are compared to surface station data and measurements collected during intensive observational campaigns.

Using inverse top–down methods, global multi-instrument space-borne constituent data from recent and planned satellite missions will help constrain further emissions of pollutants. It is worth noting that global top–down inverse modelling results depend strongly on the influence functions that describe the response of atmospheric pollutant concentrations to changes in surface emissions. The spatial and temporal structure of these functions is controlled by resolved and subgrid transport processes in the planetary boundary layer (PBL) and free atmosphere (layers above the PBL). These transport mechanisms, which deliver pollutants into the atmosphere from surface sources, will be discussed next.

## *2.5 Transport of Chemicals from Sources in the PBL and Convection*

The surface of the Earth is a rigid boundary that creates a frictional drag on air mass motion. The frictional effects are dominant in the PBL, where the horizontal

mean circulation is relatively weak. The theory of turbulent eddies describes the thermodynamics and tracer transport in the PBL; its vertical depth is controlled by vertical eddy fluxes of heat and water. The PBL contains about 10% of the total mass of the atmosphere. Depending on location, and the type of surface, the characteristic height of the PBL shows seasonal and diurnal changes varying from 10–100 m to 1–2 km. For instance, the nighttime PBL is much shallower than its daytime counterpart.

Surface-sourced pollutants begin to spread horizontally in the PBL; turbulent constituent fluxes then control this spread and the mixing of the pollutants with ambient air. Depending on the thermal stratification of the PBL, constituents can be trapped near the surface or can be ventilated out of the PBL by convective transport. Extreme concentrations of pollutants (smog) are observed for stable layers associated with temperature inversions. In these stable layers, constituents in warm air at the Earth's surface (both land and water) can rapidly be advected upward into the free atmosphere. Chapter *General Concepts in Meteorology and Dynamics* (Charlton-Perez et al.) discusses the general circulation in the free atmosphere.

The convective circulation in the clear-sky and cloudy tropical atmosphere is schematically shown in the left hand side of Fig. 1. In the deep clouds, the fast uplifting of moist air allows short-lived chemically active radicals to be advected into the upper troposphere and lower stratosphere, and affect the budget of ozone and related species in this region. The slow downdrafts of moist heavy air from convective clouds and their mixture with adjacent dry air masses close the budget associated with the convective transport of chemical species. In global models there are two types of convective schemes: deep and shallow convection schemes that parametrize the processes in deep tropical clouds and clouds near the top of the PBL, respectively. Triggering of convective processes depends on the stability of the PBL and eddy fluxes of heat and moisture. The convective transport of pollutants such as CO, $HNO_3$, $NO_x$ and hydrocarbons is parametrized in a manner analogous to water vapour convection. In global CTMs, both the subgrid vertical mass fluxes controlled by convection, and the vertical diffusion, are important mechanisms that describe the exchange of air masses across the boundaries marked by the height of PBL and the location of the tropopause (the boundary between the troposphere and stratosphere) – see Fig. 1.

## *2.6 Circulation and Transport*

The circulation of the troposphere and the stratosphere, and its variations associated with variable synoptic weather systems and planetary wave structures, determines the global and regional distribution of chemicals in the free atmosphere. Several conceptual transport models have been proposed to highlight the key features of tracer dynamics in the tropics, and mid and high latitudes. These concepts aim to interpret observed troposphere-stratosphere mass exchange, intrusions of tracers between the subtropical and polar transport barriers, and the summer-winter inter-hemispherical transport of air in the mid and upper stratosphere. Figure 1 illustrates schematically

the descent of air in the polar regions during stratospheric winter, and the development of the Hadley cell with vertical upwelling in the tropics and descent of air in adjacent subtropical regions. It also highlights the intrusions of extra-tropical stratospheric ozone into the troposphere that occur during tropopause folding events. The chapters on transport in Andrews et al. (1987), Brasseur et al. (1999) and Brasseur and Solomon (2002), present the fundamental concepts and provide details on the spread and mixing of passive and reactive species by the atmospheric circulation. The next subsections discuss briefly the global transport and mixing of tracers in the troposphere, stratosphere, and across the tropopause.

### 2.6.1 Tropospheric Circulation and Mixing

The remarkable differences between the tropospheric and stratospheric circulations can be explained by the roles of tropospheric water (in all its phases) and the Earth's surface. For example, the distribution of land and water can create significant zonal asymmetries in the radiation absorbed by the Earth's surface. This differential absorption of radiation controls the circulation systems that establish large-scale transport pathways of tracers in the troposphere. In comparison to the atmosphere, the Earth's surface is a more effective absorber of solar radiation. This feature explains the presence of a net radiative cooling in the mid and upper troposphere, with heating tending to be dominant near the surface. Furthermore, the latent heat release associated with water vapour condensation in the cloudy troposphere, acts to increase the instability of the vertical layers. Taken together, all these processes allow one to use the concept of radiative convective equilibrium to approximate globally the tropospheric thermal regime.

Convective transport of heat and moisture ensures neutral stability in the troposphere, and provides vertical stirring and dissipation of the vertical perturbations. Convective mixing of air parcels across isentropic surfaces is accompanied by diabatic heat release during the condensation of water vapour, thereby creating baroclinically unstable regions. In the troposphere, moist air parcels rise along vertically tilted isentropic surfaces and the transport of air masses is accompanied by cloud formation and precipitation.

The distribution of the land and the ocean, and variations in land topography, are important for the formation of tropospheric circulation systems. At large spatial scales, these mechanisms are responsible for the geographical variation of the storm tracks showing, for instance, the prevailing eastward synoptic storm tracks in mid latitudes, and the equatorial east–west circulation cells (Walker cells) in the tropics associated with the heat distribution over the oceans and continents. Land-sea and valley–mountain breezes are well-known examples of regional circulations induced by differential topography. Diagnostic studies of the global circulation indicate that the troposphere is well-mixed on the time-scales of months (i.e., a season), although strong and fast convective mixing can be observed at shorter time-scales and at regional spatial scales. These relatively fast dynamical time-scales ensure effective vertical and meridional mixing in the troposphere, and explain the weak

spatial gradients of long-lived tropospheric tracers, such as $CH_4$, $N_2O$, CFC-11, and CFC-12.

In order to quantify seasonal changes in tracer budgets, a number of authors have computed the effective meridional circulation and large-scale eddy mixing attributable to flows simulated by General Circulation Models (GCMs) and represented by meteorological analyses (e.g. Plumb and Mahlman 1987; Holton et al. 1995; Haynes and Shuckburgh 2000a, b; Lyjak and Yudin 2005). Figure 5 below from Yudin et al. (2000) shows examples of meridional mass stream functions for January and July derived from the NCAR (National Center for Atmospheric Research) climate middle atmosphere model. For the zonal mean diagnostics, the "residual" or *Transformed Eulerian-Mean* (*TEM*) circulation (middle column of Fig. 5) represents the effective transport that takes into account compensating effects between the Eulerian mean (left column of Fig. 5) and the eddy fluxes associated with deviations of the three-dimensional flow from the zonal mean. The TEM is based on the reformulation of zonal mean temperature balances and approximates the large-scale Lagrangian mean transport, i.e., the mean circulation that

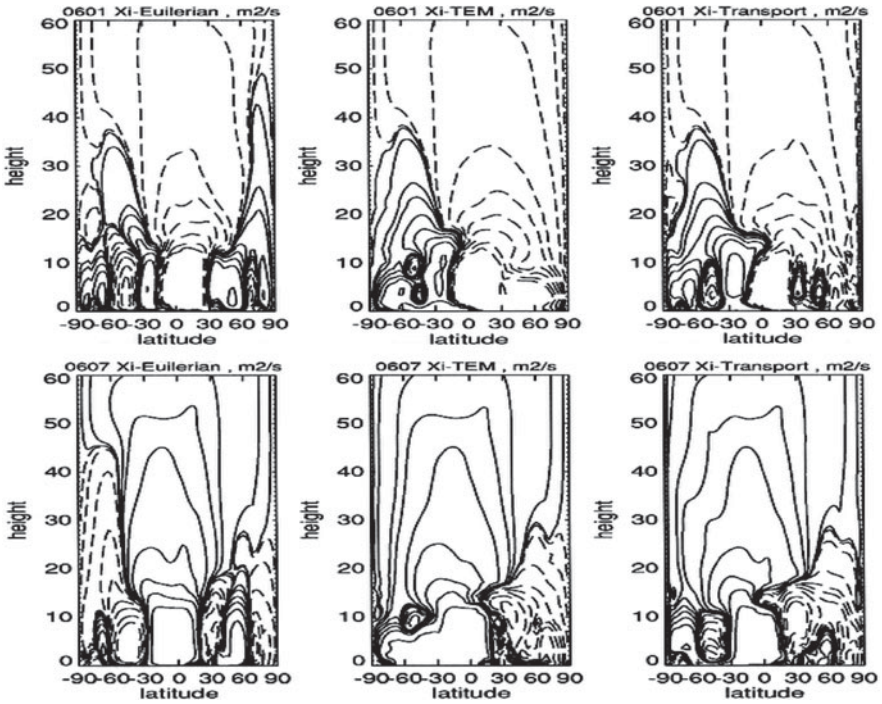



**Fig. 5** The Eulerian (*left column*); TEM (*middle column*); and transport (*right column*) meridional stream functions derived from the sixth year of a MACCM2/NCAR model simulation. The *top row* shows results for January; the *bottom row* presents results for July. The *dashed stream* function contours designate the clockwise mass transport (from south to north); *solid lines* show the courterclockwise mass transport

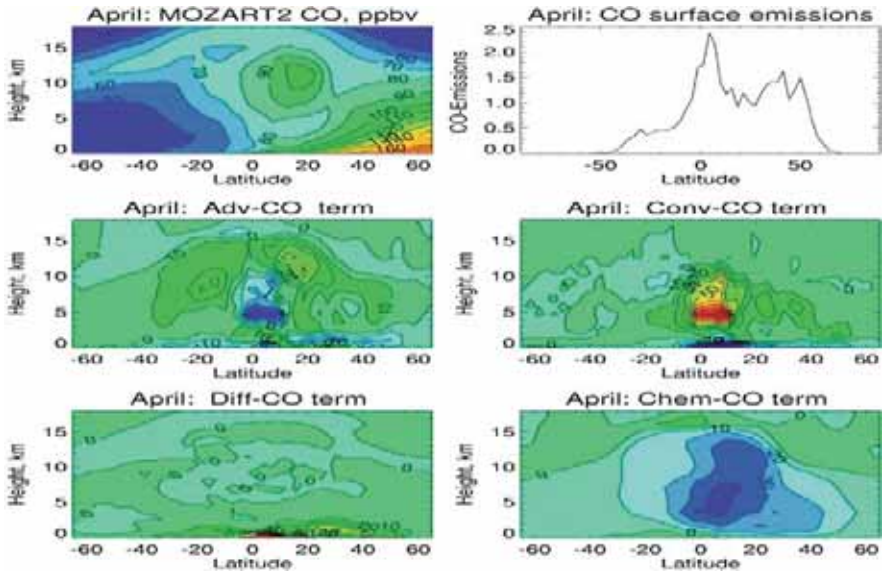



**Fig. 6** Zonal mean monthly (April) carbon monoxide (CO) in the troposphere simulated by the NCAR MOZART CTM (*top left panel*), and distribution of surface emissions (*top right panel*) and chemistry-transport terms: advection (*middle left panel*), convection (*middle right panel*), diffusion (*bottom left panel*), and chemical production-loss (*bottom right panel*)

follows a set of fluid particles (Andrews et al. 1987). The TEM framework successfully represents the tracer transport in atmospheric flows where large-scale non-zonal oscillations can be described by the superposition of stationary and weakly dissipative planetary waves.

For dissipative flows, one can introduce the concept of the transport meridional circulation for passive tracers (Plumb and Mahlman 1987). This treatment is analogous to the TEM approach described above. This concept aims to describe both the transport and mixing of tracers induced by transient and dissipative wave motions. The right column of Fig. 5 shows the stream function of the transport circulation. From Fig. 5 we can see that the major differences between the TEM and transport circulations occur in the well-mixed troposphere and lower stratosphere, while in the mid and upper stratosphere both circulations provide patterns similar to those of the meridional circulation.

Figure 6 summarizes the concepts discussed in this subsection. It shows the budget of the April zonal mean CO distribution predicted by the NCAR MOZART CTM. Diagnostic simulations of the CO budget show the interplay of the different mechanisms that control the distribution of CO between the surface and the tropopause.

### 2.6.2 Stratospheric Circulation and Mixing

As seen from the structure of the various meridional streamfunctions (Fig. 5), tropospheric air enters the stratosphere mainly in the tropics. In the stratosphere this

air moves polewards. By mass conservation this circulation must be closed by stratospheric air that returns to the upper troposphere/lower stratosphere (UTLS) at polar and mid latitudes. For wintertime stratospheric flows, the Eulerian circulation (left column of Fig. 5) deviates substantially from the TEM and transport circulations. In contrast to the situation for the summer months, the wintertime meridional mass circulation is affected by strong planetary wave activity. The wintertime polar night vortex in both hemispheres provides a transport barrier for air mass exchange between mid and high latitudes. In the winter stratosphere, breaking planetary waves form a surf zone around the edge of the polar vortex. Chapter *General Concepts in Meteorology and Dynamics* (Charlton-Perez et al.) provides further details.

Diagnostics of quasi-horizontal mixing associated with wave breaking are usually computed at isentropic surfaces; they help examine annual and interannual changes in the mixing properties of the stratospheric flow. However, note that differences between model (e.g. GCM) simulations and analysis schemes combining models and observations can affect the mixing properties of the flow. Figures 7 and 8 illustrate differences between wave amplitudes and mixing characteristics of stratospheric analyses produced by ECMWF (European Centre for Medium-Range Weather Forecasts), Met Office and GEOS-3/GMAO (Goddard Earth Observing

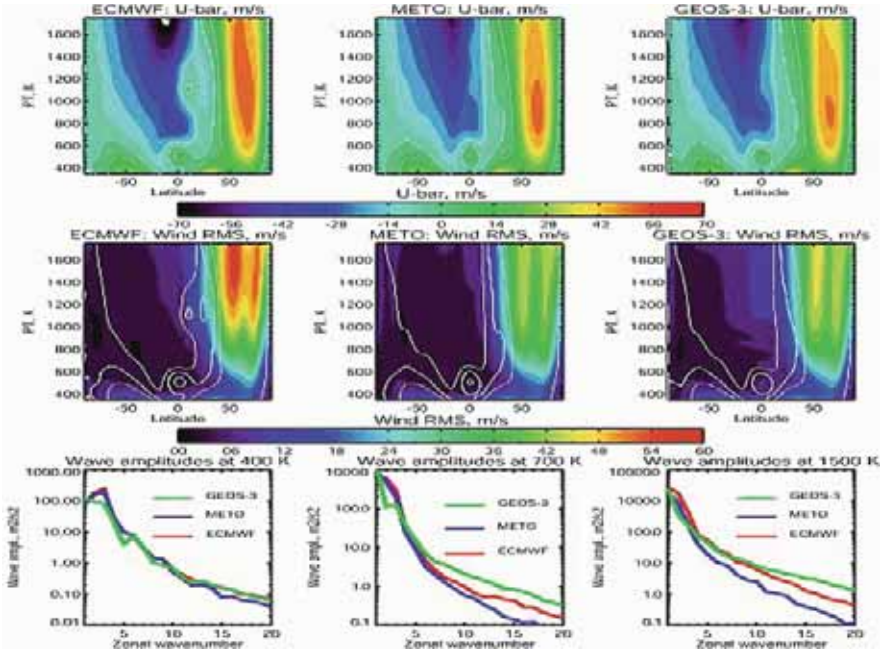



**Fig. 7** *Top* and *middle panels*: Analyses and diagnostics for ECMWF (*left panels*); Met Office (*middle panels*); and GEOS-3 (*right panels*). Zonal mean winds (*top panel*); kinetic eddy energy averaged for the period 15–30 January 2000 (*middle panel*). The *bottom panel* shows (*left to right*) kinetic energy distribution for the first 20 zonal modes at 400, 700, and 1,500 K for the three systems. The *x*-axis in the *top* and *middle panels* corresponds to equivalent latitude; the *x*-axis in the *bottom panel* corresponds to zonal wavenumber. See text for acronyms

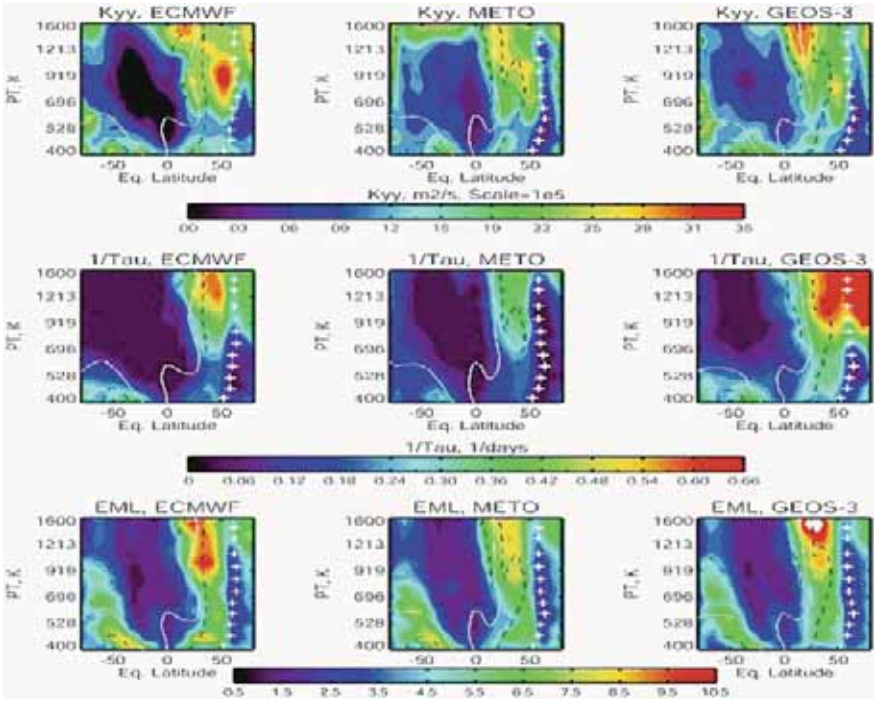



**Fig. 8** Averaged (15–30 January 2000) mixing diagnostic results for ECMWF (*left panels*), Met Office (*middle panels*), and GEOS-3 (*right panels*) stratospheric analyses. Mixing coefficient $K_{yy}$ (*top panels*); inverse e-folding time-scales ($\lambda$) obtained with the CAS (contour advection with surgery) technique (*middle panels*); distribution of the equivalent mixing lengths $L_n = L_e/(2\pi a\cos\phi)$ (*bottom panels*). Crosses show the position of the vortex boundaries in terms of the maximum of the zonal wind jet. The *white lines* depict the position of the zero wind line of the zonal winds. The adjacent dashed lines show the $-10$ and 10 $ms^{-1}$ wind contours. The *x*-axis in the plots corresponds to equivalent latitude. The *y*-axis in the plots corresponds to potential temperature (K)

System/Global Modeling Assimilation Office) (see also chapter *The Role of the Model in the Data Assimilation System*, Rood).

### 2.6.3 Transport and Chemistry Across the Tropopause

A fundamental aim of UTLS studies is to quantify the processes which control the atmospheric composition in terms of air transported from the stratosphere and air transported from the upper troposphere. For this quantification it is convenient to define conceptual boundaries, particularly the tropopause, which separates the regions of interest. This boundary is a notional concept and plays a crucial role in our understanding of Stratosphere-Troposphere Exchange (STE).

There are several definitions of the tropopause, for example those based on static stability, chemical composition and potential vorticity (PV) outside the tropics

(Shapiro 1981; Wirth 2000; Zahn et al. 2004). The tropopause is conventionally defined as the level above which the rate of temperature decrease with height (the "lapse rate") does not exceed the threshold value of –2 $Kkm^{-1}$, provided the average vertical temperature gradient above this level does not fall below this value again within a 2 km zone. The tropopause is also associated with a sharp transition in the values of concentrations of tracers and radiatively active gases such as ozone, methane, nitric acid and water vapour.

The climatological height of the tropopause as a function of latitude has traditionally been explained by radiative-convective adjustment models. However, these models do not adequately explain the position of the tropopause height in the extratropics. It has been suggested (Held 1982) that the tropopause height is controlled by the dynamical effects of synoptic-scale baroclinic eddies. Outside the tropics, the "dynamical tropopause" has become more popular recently as a tool to analyse dynamics and transport of STE events. It is defined by a specific value of PV. For conservative or weakly dissipative flows the extratropical dynamical tropopause is a material surface; this is an advantage when evaluating mass and constituent exchange across the tropopause.

Figure 9 shows the global distribution of tropopause pressure, using three different tropopause definitions, computed using NASA GMAO GEOS-5 meteorological analyses. The tropopauses computed using these definitions do not coincide, but are complementary to each other, allowing different characterizations of dynamics, mixing and chemistry. Interested readers are referred to the studies of Shapiro (1981), Wirth (2000), Zahn et al. (2004), and Randel et al. (2007). These papers explain the differences and similarities between the thermal, dynamical and chemical definitions of the tropopause boundary, and their use for global and regional estimations of STE. This is an area of active research, and assimilation of recently available constituent data from satellites such as EOS Aura is expected to shed light on UTLS transport and chemistry.

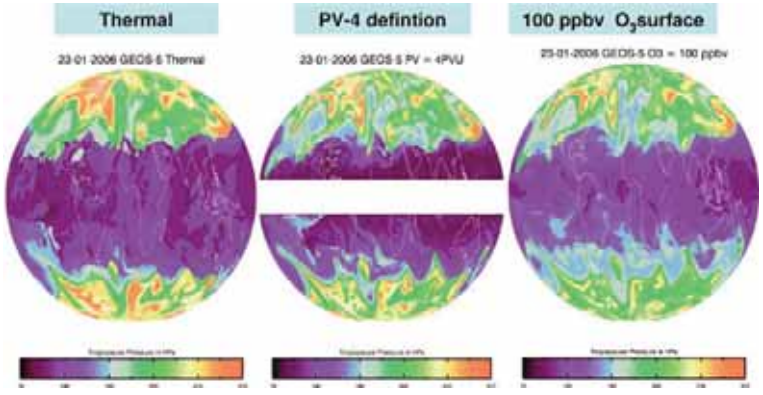



**Fig. 9** Global distributions of the tropopause pressure using three types of tropopause definition, and computed using NASA GMAO GEOS-5 meteorological analyses for 23-01-2006: thermal tropopause (*left panel*); dynamical (PV) tropopause (*middle panel*); chemical (100 ppbv $O_3$) tropopause (*right panel*)

The complexity of dynamics and chemistry in the UTLS has been highlighted in a number of studies. For chemical transformations and budgets one must include the interaction between organic and inorganic photochemical cycles, and the influence of surface and atmospheric sources. Reactions of trace gases on or within aerosol/cloud/ice particles are likely also to contribute significantly to the UTLS constituent budget, including ozone. There are two types of multiphase chemical reactions that one must account for when considering aerosols and other particles: (i) aqueous phase reactions in liquid aerosol and cloud droplets; and (ii) surface heterogeneous reactions on ice crystal and solid aerosol surfaces.

The key scientific issues for UTLS chemistry and transport are related to: (i) quantification of the chemical mechanisms and budgets of the halogen, chlorine, and bromine compounds that affect ozone concentrations in the UTLS; (ii) the upper tropospheric spatial and temporal distributions of the major chemical species in tropospheric ozone chemistry ($NO_x$, VOCs – volatile organic compounds), including the magnitude of chemical sources and sinks; (iii) understanding the role of organic chemistry in terms of sources and reservoirs for $HO_x$ and $NO_x$; (iv) the influence of transport processes on the concentrations of short-lived source gases of tropospheric origin; and (v) quantification of the concentrations of key stratospheric species ($O_3$, $HNO_3$) in the neighbourhood of the tropopause, including their downward fluxes.

In the lower stratosphere, transport processes (a key uncertainty) are affected by the different pathways and mechanisms by which air may enter this region. Relatively dry and ozone-rich air arriving from above (middle stratosphere) differs from the relatively moist, ozone-poor air and polluted air that enters the lower stratosphere from the upper troposphere.

Holton et al. (1995) illustrates the dynamical aspects of troposphere-stratosphere mass exchange across the polar, mid latitude and tropical tropopause. Along with large-scale ascent, entrainment of air into the tropical lower stratosophere via direct convective penetration from the upper troposphere has been observed, for instance, from aircraft observations of high CO concentrations. However, the cumulative importance of convective tropical tracer transport is still uncertain. More effort should be spent studying the processes by which irreversible mixing takes place. Turbulent mixing associated with the breaking of tropical inertia-gravity waves can be also important for formation of the thin low-ozone streamers that can enter the extratropical UTLS from the tropics.

In the last decade, observational investigations into mid and high latitude ozone depletion have made substantial progress in understanding the stratospheric overworld (isentropic levels above 380 K). However, some transport and mixing questions remain; in particular concerning exchange between tropics and mid latitudes and the vertical variation of such exchange. Transport studies of the UTLS based on the use of analysed winds highlight problems, likely associated with transport uncertainties, in the reproduction of observed intrusions of stratospheric and tropospheric air masses across the tropopause.

In mid latitudes, stratospheric air masses arrive in the upper troposphere during tropopause fold events (Shapiro 1980). In the upper troposphere, tracer filaments are formed through the stirring effects of synoptic eddies, and upper-level

frontogenesis, and also through large-scale diabatic downwelling. The dilution of these filaments and blobs with ambient air is controlled by efficient vertical and horizontal mixing known as CAT (clear sky atmospheric turbulence) phenomena. The polluted air from the PBL can arrive in the upper troposphere through convection and frontal circulations. To better understand STE, and the transport, chemistry and dynamics of the UTLS, it is desirable to perform combined multi-instrumental data assimilation analysis, including chemical data assimilation (see chapter *Constituent Assimilation*, Lahoz and Errera).

## 3 Summary

This chapter provides a brief overview of the chemistry and transport of atmospheric chemical species, highlighting current research challenges. These challenges can be addressed by data assimilation.

Recent observations are oriented toward making global chemical model predictions. Existing satellite-based multi-year data records provide opportunities to: (i) constrain the initial distributions of chemical species; (ii) evaluate the transport properties of air flows predicted by models; (iii) optimize key parametrizations of convective transport and diffusion; and (iv) constrain the year-to-year variations of surface emissions. To properly assimilate data into chemical models such as CTMs, issues related to stochastic and random errors of observations and models should be studied. The main subject of the next chapter in this book (*Representation and Modelling of Uncertainties in Chemistry and Transport Models*, Khattatov and Yudin) is the representation of uncertainties in the chemical models introduced in this chapter.

## References

Andrews, D., J.R. Holton and C.B. Leovy, 1987. *Middle Atmosphere Dynamics*, Elsevier Science & Technology, USA, 504 pp.

Brasseur, G.P., J.J. Orlando and G.S. Tyndall (eds.), 1999. *Atmospheric Chemistry and Global Change*, Oxford University Press, USA, 688 pp.

Brasseur, G.P. and S. Solomon, 2005. *Aeronomy of the Middle Atmosphere*. Chemistry and Physics of the Stratosphere and Mesosphere Series: Atmospheric and Oceanographic Sciences Library, Vol. 32, 3rd rev. and enlarged ed., XII. Springer, The Netherland, 646 pp.

Crutzen, P.J., 1974. Photochemical reactions initiated by and influencing ozone in unpolluted tropospheric air. *Tellus*, **26**, 48–57.

Giglio, L., I. Csiszar and C.O. Justice, 2006. Global distribution and seasonality of active fires as observed with the Terra and Aqua MODIS sensors. *J. Geophys. Res.*, **111**, G02016, doi:10.1029/2005JG000142.

Haynes, P.H. and E.F. Shuckburgh, 2000a. Effective diffusivity as a diagnostic of atmospheric transport. Part I: Stratosphere. *J. Geophys. Res.*, **105**, 22777–22794.

Haynes, P.H. and E.F. Shuckburgh, 2000b. Effective diffusivity as a diagnostic of atmospheric transport. Part II: Troposphere and lower stratosphere. *J. Geophys. Res.*, **105**, 22795–22810.

Held, I.M., 1982. On the height of the tropopause and the static stability of the atmosphere. *J. Atmos. Sci.*, **39**, 412–417.

Holton, J.R., P.H. Haynes, M.E. McIntyre, et al., 1995. Stratosphere-troposphere exchange. *Rev. Geophys.*, **33**, 403–439.

Horowitz, L.W., S. Walters, D.L. Mauzerall, et al., 2003. A global simulation of tropospheric ozone and related tracers: Description and evaluation of MOZART, version 2. *J. Geophys. Res.*, **108**, 4784, doi:10.1029/2002JD002853.

Khattatov, B.V., J.C. Gille, L.V. Lyjak, et al., 1999. Assimilation of photochemically active species and a case analysis of UARS data. *J. Geophys. Res.*, **104**, 18,715–18,737.

Lyjak, L. and V. Yudin, 2005. Diagnostics of the large-scale mixing properties from stratospheric analyses. *J. Geophys. Res.*, **110**, D17107, doi:10.1029/2004JD005577.

Pétron, G., C. Granier, B. Khattatov, et al., 2004. Monthly CO surface sources inventory based on the 2000–2001 MOPITT satellite data. *Geophys. Res. Lett.*, **31**, L21107, doi:10.1029/2004GL020560.

Plumb, R.A. and J.D. Mahlman, 1987. The zonally averaged transport characteristics of the GFDL general circulation/transport model. *J. Atmos. Sci.*, **44**, 298–327.

Randel, W.J., D.J. Seidel and L.L. Pan, 2007. Observational characteristics of double tropopause. *J. Geophys. Res.*, **112**, D07309, doi:10.1029/2006JD007904.

Shapiro, M., 1980. Turbulent mixing within tropopause folds as a mechanism for the exchange of chemical constituents between the stratosphere and troposphere. *J. Atmos. Sci*., **37**, 994–1004.

Shapiro, M., 1981. Frontogenesis and geostrophically forced secondary circulations in the vicinity of jet-stream frontal zone systems. *J. Atmos Sci.*, **38**, 955–973.

van der Werf, G.R., J.T. Randerson, L.Giglio, et al., 2006. Interannual variability in global biomass burning emission from 1997 to 2004. *Atmos. Chem. Phys.*, **6**, 3423–3441.

Wirth, V., 2000. Thermal versus dynamical tropopause in upper tropospheric balanced flow anomalies. *Q. J. R. Meteorol. Soc.*, **126**, 299–317.

Yudin, V.A., G. Pétron, J-F. Lamarque, et al., 2004. Assimilation of the 2000–2001 CO MOPITT retrievals with optimized surface emissions. *Geophys. Res. Lett.*, **31**, L20105, doi:10.1029/2004GL021037.

Yudin, V.A., S.P. Smyshlyaev, M.A. Geller and V.L. Dvortsov, 2000. Transport diagnostics of GCMs and implications for 2D Chemistry-Transport Model of troposphere and stratosphere. *J. Atmos. Sci.,* **57**, 673–699.

Zahn, A., C.A.M. Brenninkmeijer and P.F.J. van Velthoven, 2004. Passenger Aircraft Project CARIBIC 1997–2002. Part I: The extratropical chemical tropopause. *Atmos. Chem. Phys. Discuss.*, **4**, 1091–1117.

# Representation and Modelling of Uncertainties in Chemistry and Transport Models

**Boris Khattatov and Valery Yudin**

## 1 Introduction

Representation and analysis of uncertainties (errors) is at the core of any data assimilation system. The main aim of data assimilation is to reduce uncertainties of model predictions using observations. Under Gaussian error statistics for both parts of the assimilation system (data and forecast), and by making the assumption of zero bias, optimal estimation schemes can be derived. After insertion of data, the analysis error covariance can be evaluated by the optimal estimation formula for linear systems:

$$\mathbf{C}_a = (\mathbf{I} - \mathbf{KH})\mathbf{C}_f \tag{1}$$

where $\mathbf{C}_a$ and $\mathbf{C}_f$ are the analysis and the forecast error covariance matrices, respectively. $\mathbf{I}$ is the identity matrix, $\mathbf{H}$ is a linear operator which projects the analysed variables from the forecast space to location of observations and expresses them in terms of measured quantities, and $\mathbf{K}$ is the Kalman gain matrix defined previously in this book (see chapter *Mathematical Concepts of Data Assimilation*, Nichols). In this chapter, the error statistics of observations are assumed to be prescribed by observers, although in practice the combined tuning of the forecast and observational error covariances usually is performed in order to secure the optimality of the observing systems.

The subject of this chapter is representation of uncertainties in the various chemical transport models described in chapter *Introduction to Atmospheric Chemistry and Constituent Transport*(Yudin and Khatattov). In studies of atmospheric chemistry and transport, description of forecast errors depends on the resolution, spatial coverage, dimension, and complexity (e.g. number of reactive species) of models. Our introduction to the parametrization and modelling of errors for atmospheric composition is separately outlined for errors induced by the chemical coupling of

B. Khattatov (✉)
Fusion Numerics Inc, Boulder, CO, USA
e-mail: boris@fusionnumerics.com

W. Lahoz et al. (eds.), *Data Assimilation*, DOI 10.1007/978-3-540-74703-1_17,

species, and uncertainties due to the transport of tracers and emissions. Two conceptual schemes for the description and time evolution of random errors in modelling of atmospheric chemicals will be described:

– A box model approach suitable for trajectory modelling assuming negligible mixing;
– A 3-D chemistry-transport model (CTM) approach taking into account transport and diffusion effects.

With increasing computational power, parametrized representation of errors in chemistry-climate models can, in principle, be replaced by the ensemble-based methods, which are beyond the scope of this chapter. It is likely that Monte Carlo methods with an ensemble of ~30–100 members could be a practical avenue for quantitative error approximations when exploring the coupling between chemistry and transport.

## 2 Linear Formalism for Error Evolution in Box Chemical Models

Let vector $\mathbf{x}$ of length $J$ represent the state (volume mixing ratios of a number of chemicals) of a time-dependent chemistry model described in the chapter *Introduction to Atmospheric Chemistry and Constituent Transport* by Yudin and Khattatov. Let the non-linear operator $\mathcal{M}$ describe the transformation of vector $\mathbf{x}$between two consecutive time intervals $t$ to time $t+\Delta t$, that can be expressed as,

$$\mathbf{x}(t + \Delta t) = \mathcal{M}(\mathbf{x}(t)) \tag{2}$$

Let vector $\mathbf{y}$ contain $N$ observations of the state, i.e., observations of the chemical composition of the atmosphere. Usually, $N < J$, meaning that we do not have enough observations to constrain the complete model space. The connection between $\mathbf{y}$ (the observations) and $\mathbf{x}$ (the model values) can be established through the non-linear observational operator $\mathcal{H}$, which represents mapping of the state variables from the model space to the observational space (chapter *Mathematical Concepts of Data Assimilation*, Nichols):

$$\mathbf{y} = \mathcal{H}(\mathbf{x}) \tag{3}$$

Everywhere in this discussion we assume that the interpolation errors associated with operator $\mathcal{H}$ are negligible. The results are easily extended to the case when this is not true. Combining the above two equations, we get

$$\mathbf{y} = \mathcal{H}(\mathcal{M}(\mathbf{x})) \tag{4}$$

The data assimilation problem is then to find the "best" value of **x**, which inverts this equation for a given **y** allowing for observation errors and other prior information (Lorenc 1986). In most cases, the dimensions of vectors **x** and **y** will be different, and this problem will be either overdetermined or underdetermined. Therefore, inversion of Eq. (4) should be done in the statistical sense.

"Best" here means that the errors of the final analysis are minimal. An exact value of a physical quantity can rarely be determined. One can only say that this value lies within a certain range with a certain probability, and therefore all estimates of the best value of **x** obtained from the observed **y** are probabilistic in nature. A mathematically robust definition of the best or optimal **x** is, for instance, the value corresponding to the maximum of the probability density function (PDF) of **x** given observations **y**. This is the maximum likelihood definition.

The exact shapes of the PDFs in both **x** and **y** spaces are generally unknown. In order to solve the problem posed one needs to establish a relationship between the PDF of **x** and the PDF of **y**. Formal transformation of PDFs by the model from the parameter space **x** to the model space **y** is described by the Fokker-Kolmogorov equation (e.g. Jazwinski 1970), which is impossible to solve in most practical applications. This is one of the reasons why simplifications are needed in order to be able to solve practical problems in data assimilation.

One simplification is that the probability density functions can be approximated by Gaussian functions:

$$\mathrm{PDF}(\mathbf{x}) \sim \exp\left\{-0.5(\mathbf{x}-\widehat{\mathbf{x}})^T\mathbf{C}^{-1}(\mathbf{x}-\widehat{\mathbf{x}})\right\} \tag{5}$$

where $\widehat{\mathbf{x}}$ is the true (unknown) value of **x,** and **C** is the corresponding error covariance matrix. Its diagonal elements are the uncertainties (standard deviations) of $\widehat{\mathbf{x}}$ and the off-diagonal elements represent correlation between uncertainties of different elements of vector **x**. The covariance matrix **C** is defined as

$$\mathbf{C} = \langle(\mathbf{x}-\widehat{\mathbf{x}})(\mathbf{x}-\widehat{\mathbf{x}})^T\rangle \tag{6}$$

where angle brackets represent averaging over all available realizations of **x**.

We also assume that there exists a prior, independent estimate of **x**, or $\mathbf{x}_b$, often called the background, with corresponding background error covariance **B**. The solution minimizing the final analysis errors is given by a minimum of the following functional, where *T* denotes transpose (Lorenc 1986):

$$J(\mathbf{x}) = [\mathbf{y}-\mathcal{H}(\mathcal{M}(\mathbf{x}))]^T(\mathbf{O}+\mathbf{F})^{-1}[\mathbf{y}-\mathcal{H}(\mathcal{M}(\mathbf{x}))] + [\mathbf{x}-\mathbf{x}_b]^T\mathbf{B}^{-1}[\mathbf{x}-\mathbf{x}_b] \tag{7}$$

Here **O** is the observational error covariance matrix, **F** is the error covariance corresponding to operators $\mathcal{M}$ and $\mathcal{H}$, and **B** is the background error covariance matrix. The sum **O** + **F** is often represented as **R**. These error covariance matrices characterize our confidence in the measurements, the model and observation operator, and the a priori background estimate. $J(\mathbf{x})$ is often called the misfit or cost function.

In practical applications one has to find an appropriate way to compute the error covariances and to minimize $J(\mathbf{x})$. In most cases, in order to be able to do this we need to introduce the linear approximation. In the linear approximation we assume that for small perturbations of the parameter vector $\Delta\mathbf{x}$ the following is a good approximation:

$$\mathcal{M}(\mathbf{x} + \Delta\mathbf{x}) = \mathcal{M}(\mathbf{x}) + \mathbf{L}\Delta\mathbf{x} \tag{8}$$

In this expression $\mathbf{L}$ is a matrix, while $\mathcal{M}$ is, in general, a non-linear operator. Formally, $\mathbf{L}$ is a derivative of $\mathcal{M}$ with respect to $\mathbf{x}$:

$$\mathbf{L} = \mathrm{d}\mathcal{M}/\mathrm{d}\mathbf{x} \tag{9}$$

The linearization $\mathbf{L}$ of the original model $\mathcal{M}$ will be used in two ways. First, minimization of $J(\mathbf{x})$ often requires knowledge of the derivative of $J(\mathbf{x})$ with respect to $\mathbf{x}$. This, in turn, requires knowledge of $\mathrm{d}\mathcal{M}/\mathrm{d}\mathbf{x}$. Second, for small variations of $\mathbf{x}$ one can show that the transformation of error covariance matrix $\mathbf{C_x}$ in the parameter space to the error covariance matrix $\mathbf{C_y}$ in the model space is as follows:

$$\mathbf{C_y} = \mathbf{L}\,\mathbf{C_x}\mathbf{L}^T \tag{10}$$

This, in turn, allows one to establish a correspondence between the PDF of $\mathbf{x}$ in the parameter space and the PDF of $\mathbf{y}$ in the model space. If $\mathcal{M}$ represents the original non-linear model, matrix $\mathbf{L}$ is said to be the tangent-linear model and its transpose, $\mathbf{L}^T$, is said to be the adjoint of $\mathcal{M}$. The linearization matrix describes time evolution of small perturbations of the model state:

$$\delta\mathbf{x}(t + \Delta t) = \mathbf{L}\delta\mathbf{x}(t) \tag{11}$$

In the case of the photochemical box model described in the chapter *Introduction to Atmospheric Chemistry and Atmospheric Transport* (Yudin and Khattatov), Eq. (11) expands to

$$\begin{bmatrix} \mathrm{H} \\ \mathrm{OH} \\ \ldots\ldots \\ \ldots\ldots \\ \mathrm{O_3} \\ \mathrm{H_2O(a)} \end{bmatrix}_{t+\Delta t} = \begin{bmatrix} L_{11}\ L_{12}\ L_{13}\ \ldots\ldots \\ L_{21}\ L_{22}\ L_{23}\ \ldots\ldots \\ \ldots\ \ldots\ \ldots\ \ldots\ldots \\ \ldots\ \ldots\ \ldots\ \ldots\ldots \\ \ldots\ \ldots\ \ldots\ \ldots L_{\mathrm{NN}} \end{bmatrix} \begin{bmatrix} \mathrm{H} \\ \mathrm{OH} \\ \ldots\ldots \\ \ldots\ldots \\ \mathrm{O_3} \\ \mathrm{H_2O(a)} \end{bmatrix}_{t} \tag{12}$$

The linearization matrix $\mathbf{L}$ is, in general, a function of the time interval $\Delta t$. This is easy to understand if we take the extreme case of $\Delta t = 0$. In this case, the final perturbation is the same as the initial perturbation and $\mathbf{L}$ is the identity matrix, $\mathbf{I}$. As the time interval increases, the linearization matrix changes its structure (see Fig. 1).

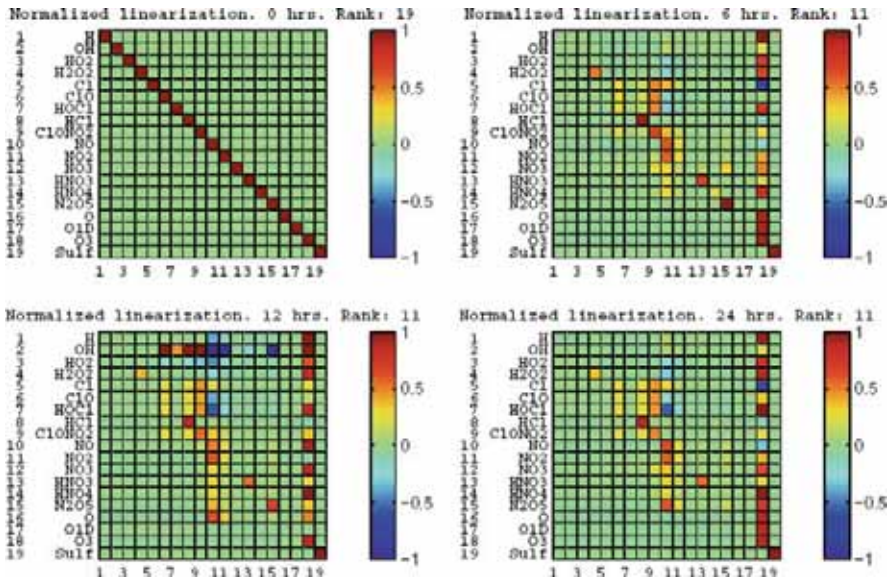



**Fig. 1** Example of time evolution of the linearization matrix, **L**. See text for details. With permission from Khatattov et al. (1999)

After a few hours of integration a pattern emerges in the distribution of the non-zero elements, with only a few columns containing most of the non-zero values. This demonstrates that a relatively small number of species determine concentrations of all constituents in the model at later times.

For a typical stratospheric chemical system, the matrix **L** is not invertible for $\Delta t$ longer than a few hours. The rank of **L**, i.e., number of linearly independent rows or columns, quickly decreases with time, thus making the matrix not invertible. The rank is shown in Fig. 1 on top of each plot. For this example, after just 6 h the rank decreases from 19 to 11 and becomes 9 after 4 days of integration.

This means that for this example, in general, only nine linear combinations of initial species concentrations completely define concentrations of all 19 constituents after 4 days. Formally, on day 4, matrix **L** represents a transformation from 19-dimensional space to 9-dimensional space. This is a multidimensional equivalent of multiplication by zero along some of the dimensions. No matter how large some concentrations were initially, in a few hours or a few days their impact might be completely negligible. This behaviour is due to a strong diurnal cycle and the short lifetime of some species in the model.

An interesting consequence of this result is that the past state of the modelled stratospheric chemical system can never be determined from present observations of the system, since **L** cannot be inverted. On the other hand, it means that one does not have to know concentrations of all species to predict the state at some later time. For this example, provided that the model is fairly realistic, only nine linear combinations of species concentrations spanning the orthogonal space of matrix **L** need to

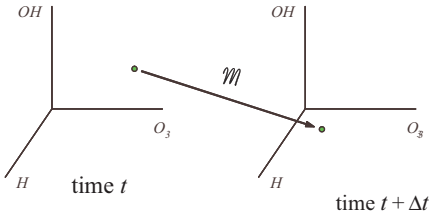



**Fig. 2** Action of photochemical model, $\mathcal{M}$. See text for details. See also Khattatov (2003)

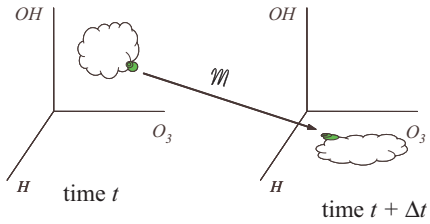



**Fig. 3** Transformation of PDFs. See text for details. See also Khattatov (2003)

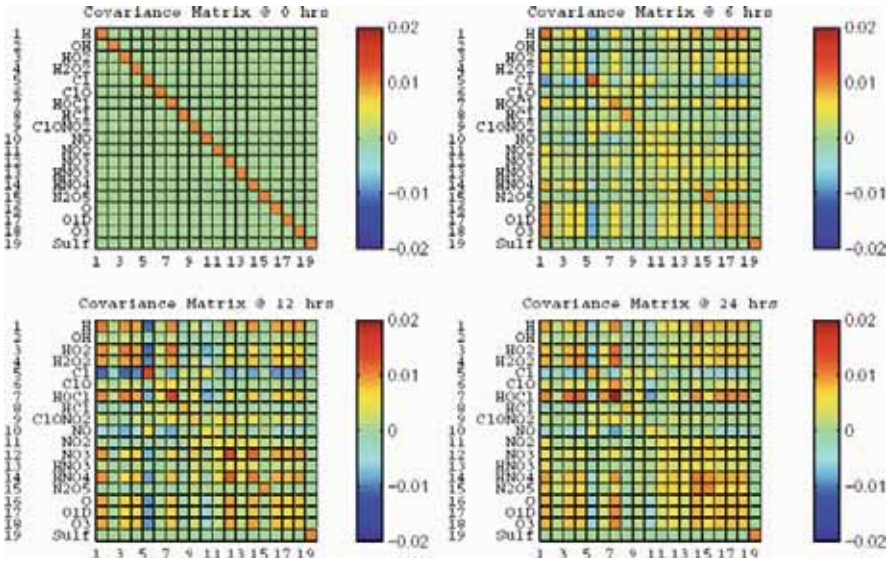



**Fig. 4** Time evolution of error covariance matrices. The panels refer to values of the covariance matrix at times in a numerical experiment: 0 h (*top left*), 6 h (*top right*), 12 h (*bottom left*) and 24 h (*bottom right*). With permission from Khatattov et al. (1999)

be known in order to predict concentrations of all 19 model constituents 4 days later. Computer codes for constructing and solving the described photochemical model as well as for computing the linearization and covariance matrices can be found at: http://acd.ucar.edu/~boris/research.htm.

One can think of the photochemical model $\mathcal{M}$ as a transformation from the $N$-dimensional space (19-dimensional for the example discussed in chapter *Introduction to Atmospheric Chemistry and Constituent Transport*by Yudin and Khatattov) of constituent concentrations at present time to some future time, as illustrated schematically in Fig. 2.

In most practical cases, the value of **x** at the initial time is not known precisely; instead one can specify a region of likely values of **x**. Instead of a point-to-point transformation; we now have a region-to-region transformation (Fig. 3). The "shapes" of these regions are described by probability density functions.

Evolution of the probability density functions is very hard to compute in practice due to the high dimensionality of the model space and the high computational requirements of the model operations. However, the Gaussian assumption and the linearization approximation allows one to use Eq. (10) to compute the evolution of the Gaussian error covariance matrices. An example of the temporal evolution of error covariance matrices computed this way is shown in Fig. 4.

## 3 Variance Evolution and Applications to Measurement Information Content

As shown in Fig. 1, as $\Delta t$ becomes larger than the lifetime of the shortest-lived chemical constituent in the model, the matrix **L** becomes rank deficient and, hence, non-invertible. In effect, it means that knowledge of the initial concentrations of some short-lived chemicals is irrelevant to establishing the state of the system at time $t+\Delta t$ since they are determined by concentrations of other chemicals.

Thus, one can pose the following questions (as adapted from Khattatov et al. 2001):

(1) Given concentrations and uncertainties of concentrations of a set of chemicals at time $t$, what can be inferred about their concentrations and uncertainties at time $t + \Delta t$?
(2) Which chemicals are the most important for determining the complete state of a chemical system and which are the least important?
(3) What are the most relaxed (i.e., maximum) measurement errors that guarantee specified prediction errors for a particular set of atmospheric chemicals?

To illustrate how one can address these questions we reproduce the approach described in Khattatov et al. (2001). The box photochemical model is initialized and run for several days using parameters (temperature, pressure, constituent concentrations) typical for the spring mid latitude stratosphere at 10 hPa (about 30 km in altitude). Concentrations of the following 18 species are predicted: H, OH, $HO_2$, $H_2O_2$, NO, $NO_2$, $NO_3$, $N_2O_5$, $HNO_3$, $HNO_4$, Cl, ClO, HOCl, HCl, $ClONO_2$, O, $O(^1D)$, and $O_3$; and concentrations of several others are held constant: CO, $CH_4$, $N_2O$, $H_2$, $H_2O$, and sulphate aerosol. The linearization matrices are automatically computed and stored for each time step of the model integration; the time step can

vary from milliseconds to 15 min. The final linearization matrix corresponding to a larger time interval can then be computed by multiplication of the intermediate matrices corresponding to individual time steps. For now we ignore uncertainties in photochemical/chemical reactions rates, numerical errors, and errors due to missing photochemical processes in the model

To show how uncertainties evolve in time we will now focus on variances, or diagonal elements of **C,** and ignore its off-diagonal elements. Let vector **v** contain the values of all the diagonal elements of **C**. It is easy to see from Eq. (10) that

$$\mathbf{v}_{t+\Delta t} = \mathbf{L}^2\ \mathbf{v}_t \tag{13}$$

where elements of the matrix $\mathbf{L}^2$are simply squared elements of **L**. For large enough $\Delta t$, both of these matrices become rank deficient and non-invertible in the conventional sense. It means that the same uncertainties $\mathbf{v}_{t+\Delta t}$ of the prediction can be obtained from different initial uncertainties $\mathbf{v}_t$. If we assume that initial uncertainties come from measurements, we can pose the following question: What are the maximum measurement uncertainties that lead to prediction errors smaller than a specified upper limit? If vector $\mathbf{v}_{\max}$ designates the maximum allowed prediction variance then, formally, the problem is to maximize each element of $\mathbf{v}_t$ subject to the following constraints:

$$\begin{aligned} &\mathbf{v}_{\max} \geq \mathbf{L}^2\mathbf{v}_t \\ &\mathbf{v}_t > 0 \end{aligned} \tag{14}$$

Since $\mathbf{v}_t$ is a vector, this is a multi-objective optimization problem which, in general, has multiple solutions. Various algorithms have been developed for finding practical solutions to such problems. Here, we apply the goal attainment method described in Gembicki (1974) to find a solution to Eq. (14).

We performed a 24-h model integration and computed the corresponding $18\times18$ linearization matrix. We then computed the singular value decomposition (SVD) spectrum of the linearization matrix, shown in blue in Fig. 5. The portion of the spectrum corresponding to the nine largest singular values is fairly flat while the tail of the spectrum drops abruptly. This means that projections of the vector of initial concentrations onto the corresponding nine eigenvectors contain most of the information needed to determine concentrations of all 18 constituents after 24 h.

Figure 5 also presents the SVD spectrum of the matrix $\mathbf{L}^2$ whose elements are the squared elements of matrix **L**. According to Eq. (11) this matrix determines the evolution of variances. As expected, the tail portion of this SVD spectrum is significantly flatter than that of matrix **L** and the "cut-off" value is not obvious. To illustrate the time evolution of relative errors in the model we performed a 10-day model integration, set the values of variances **v** at the beginning of the integration to correspond to 10% relative errors, and computed variances at each model time step. Results of these calculations are presented in Fig. 6.

From Fig. 6 one can see that relative errors change in a complicated way in response to the diurnal cycle and photochemical transformations between species.

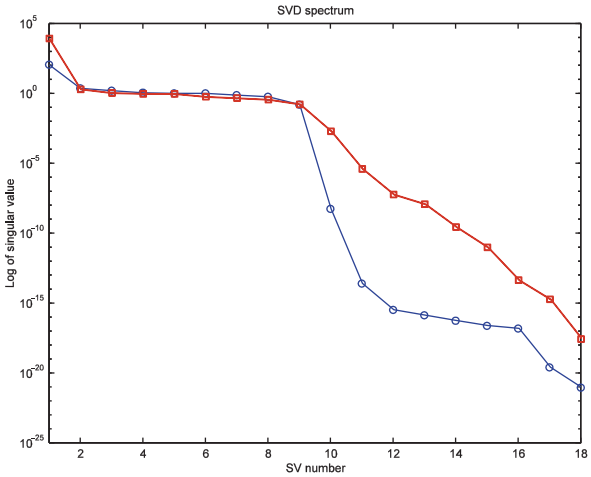



**Fig. 5** SVD spectrum of matrix **L** (in *blue*, *circles*) and $\mathbf{L}^2$ (in *red*, *squares*). With permission from Khattatov et al. (2001)

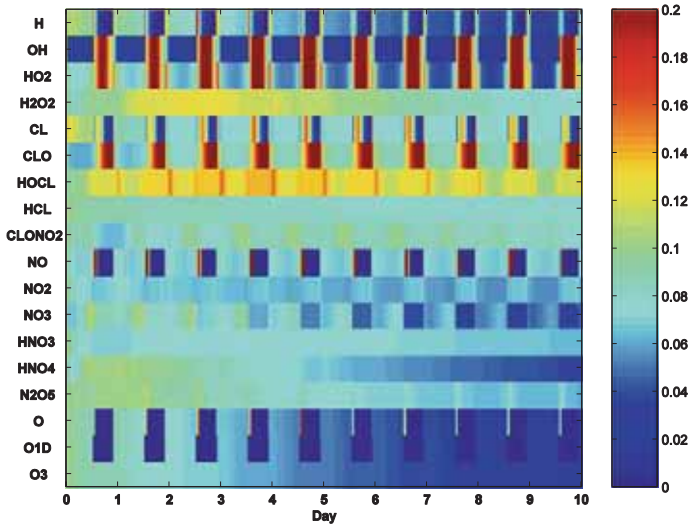



**Fig. 6** Time evolution of relative errors. With permission from Khattatov et al. (2001)

Fairly simple calculations represented by Eq. (13) allow one to assess easily how uncertainties associated with a particular set of measurements vary with time. The case where we assume that concentrations of all model constituents are known with a 10% error is clearly not realistic. However, calculations similar to the one performed may still provide some guidance as to how much information is contained in a particular set of measurements. Even if concentrations of some species are not measured at all, various a priori estimates are usually available and can be utilized provided that the corresponding variances are set to reasonably large values.

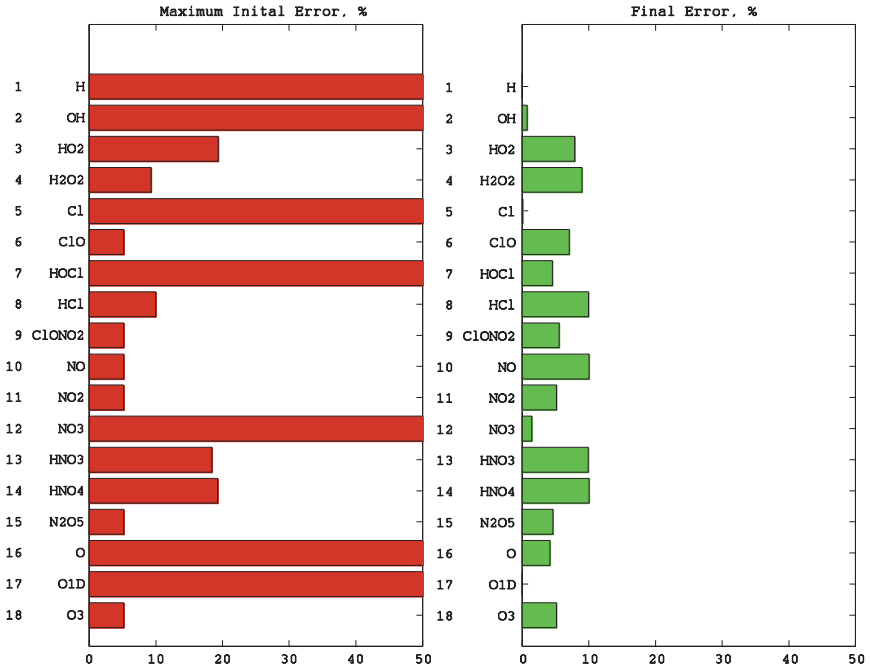



**Fig. 7** Maximum initial errors that guarantee final errors less than 10% (*left panel*). Actual final errors (*right panel*). With permission from Khattatov et al. (2001)

We will now concentrate on an "inverse" problem and assume that we have a set of requirements, i.e., the maximum allowed errors of prediction at the end of the 24-h period. What are the maximum possible errors of the measurements (at the beginning of the time interval) that guarantee maximum specified prediction errors? For illustrative purposes we assume that the relative prediction errors at the end of the 24-h interval have to be equal to or less than 10% and then attempt to find maximum initial errors that satisfy this criterion. As mentioned earlier, this problem is, in general, ill-posed and therefore semi-empirical algorithms have to be utilized to address it. A solution obtained with the goal attainment algorithm (Gembicki 1974) is shown in Fig. 7.

The bar graph on the right of Fig. 7 shows relative errors of constituent concentrations at the end of the 24-h model integration while the bar graph on the left of Fig. 7 shows the derived relative errors of initial concentrations. As required, the final relative errors of concentrations of all species do not exceed 10%, most chemicals showing smaller errors. The initial errors for some species (H, OH, Cl, HOCl, $NO_3$, O, $O(^1D)$) are so large that they go off the scale of the plot. The actual numerical values correspond to hundreds of percent and indicate that initial concentrations of these chemicals are irrelevant to determining the future state of the system for as long as the initial guess is of the correct order of magnitude. For the rest of the chemicals, initial errors range from about 5 to 20%. Clearly, the smaller the required initial

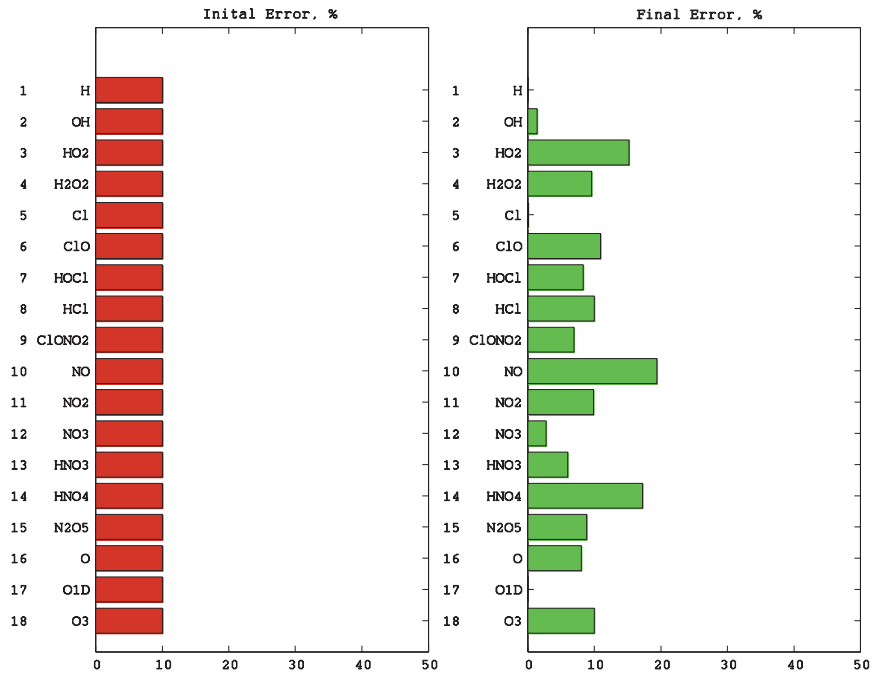



**Fig. 8** Final errors (*right panel*) for the case when all initial errors are set to 10% (*left panel*). With permission from Khattatov et al. (2001)

error the more important is the role of the corresponding chemical in determining the future system evolution. For comparison, Fig. 8 presents results of the forward error calculations for the case when the initial relative errors for all chemicals are set to 10%. As one can see, this does not guarantee that the final errors are 10% or less.

Uncertainties, or variances, can be considered to be a quantitative measure of the amount of useful information about the chemical system under consideration. The framework discussed above allows one to assess how this "information" changes with time. Assuming that the initial estimates of the system state come from measurements, this framework allows one to determine which chemicals should be measured and with what uncertainties in order to be able to make the best possible predictions. The above case for a typical stratospheric system is largely academic and its results confirm quantitatively what is already known, e.g., that concentrations of several key species or linear combinations of species (families) control the future evolution of the system. These results are encouraging and we believe that this framework will in practice be most useful when applied to complex and poorly studied chemical systems involving possibly hundreds of chemicals. Tropospheric chemistry in general, and boundary layer chemistry in particular, are examples of systems where this methodology can provide quantitative guidance and help to establish measurement priorities. Chapter *Observing System Simulation Experiments* (Masutani et al.) discusses similar concepts in the context of evaluating additions to the Global Observing System.

## 4 Error Representation in 3-D Chemistry-Transport Models

The approach described in Sect. 3 for computing error evolution based on the extended Kalman filter (EKF) assimilation technique described in Khattatov et al. (2001) is quite suitable for 0-D (box) photochemical models in combination with trajectory modelling. 3-D chemistry-transport models (CTMs) are much more powerful tools for studying the atmosphere as they account for a variety or transport and chemical processes within one model. Applying the same formalism described in Sect. 3 to CTMs, however, is prohibitively expensive due to the high dimensionality of the problem. For a CTM with a million grid points modelling 20 chemicals, the size of the error covariance matrix ($\sim 10^{14}$ elements) becomes impossible to handle in practice.

In this section we describe some attempts to characterize error evolution in CTMs that rely on earlier studies by Cohn (1993), Lyster et al. (1997), Ménard et al. (2000) and Ménard and Chang (2000), among others, to approximate the evolution of errors related to transport processes. In chemical data assimilation, model errors can be viewed as uncertainties in the specification of the CTM "prescribed" parameters (winds, diffusion, convection, reaction rates) and initial and boundary distributions. Indeed, analysed wind errors can directly control the variance distribution of the long-lived tracers, while estimated errors in the chemical constant rates and temperatures define uncertainties in the distribution of the short-lived constituents.

For simplicity of notation we will use the continuity tracer equation in the rectangular coordinate system with $x_k$-axes ($k = 1, 2, 3$) and $U_k$-wind components. We assume the non-divergent approximation for the stratospheric flow along the pressure surfaces. Generalization of our derivations for spherical geometry in a hybrid sigma-pressure vertical coordinate system is straightforward. Convection effects are ignored here but can also be introduced relatively easily. The continuity equation for constituent $q_i$ with chemical source $S_i$ can be written as follows:

$$\frac{\partial q_i}{\partial t} + \sum\nolimits_k U_k \frac{\partial q_i}{\partial x_k} - S_i = 0 \tag{15}$$

Assuming a decomposition of the model state ($q_i$), temperature *(T),* winds *($U_k$),* and photochemical sources *($S_i$)* on to the ensemble averaged values <*f*> and stochastic fluctuations *f'*, we can derive the forecast variance $V_i^f$ equation:

$$\begin{aligned} &\frac{\partial V_i^f}{\partial t} + \sum\nolimits_k \langle U_k \rangle \frac{\partial V_i^f}{\partial x_k} + \sum\nolimits_k \langle U_k' q_i' \rangle \frac{\partial \langle q_i \rangle}{\partial x_k} - \langle S_i' q_i' \rangle = 0 \\ &V_i^f = \frac{1}{2} \langle q_i'^2 \rangle \end{aligned} \tag{16}$$

Ignoring the flow-dependent terms in the left-hand-side of Eq. (16) and transforming the second order eddy terms on the left-hand-side, we obtain the variance equation discussed in Cohn (1993):

$$\frac{\partial V_i^f}{\partial t} + \sum\nolimits_k \langle U_k \rangle \frac{\partial V_i^f}{\partial x_k} = Q_i^t + Q_i^c = Q_i^m,\ Q_i^t = -\sum\nolimits_k \langle U_k' q_i' \rangle \frac{\partial \langle q_i \rangle}{\partial x_k}, Q_i^c = \langle S_i' q_i' \rangle \tag{17}$$

In this derivation, the model error terms $Q^m$ for stratospheric tracers are written explicitly and associated directly with the transport and chemical errors. Several approaches can be employed for the $Q^t$and $Q^c$ specification. The Monte Carlo or ensemble forecast simulations can be used as the most direct stochastic method to estimate the tracer perturbations $q'$. The linearized model can be used to build $<S_i' q'>$ and $<U_k' q'>$ for various perturbations in the wind system and photochemical parameters. For instance, for calculations of the $Q^c$-term we can use the linearization (written $L$ here) and photochemical Jacobian (written $J$ here) matrices (see, e.g., Khattatov et al. 1999) to relate the multivariate cross-species covariance to the chemical model error term $Q^c$ in the given model box:

$$Q_i^c(t+\tau) = \langle S_i q_i \rangle = \sum\nolimits_k J_{ik} C_{ik}^{qq}(t+\tau), C_{ik}^{qq}(t+\tau) = L C_{ik}^{qq}(t) L^T \tag{18}$$

The transport model errors can be parametrized in the spirit of macro-turbulence theory as follows:

$$Q_i^t = -\sum\nolimits_k \langle U_k' q_i' \rangle \frac{\partial \langle q_i \rangle}{\partial x_k} = \sum\nolimits_k D_k \left\{ \frac{\partial \langle q_i \rangle}{\partial x_k} \right\}^2,\ D_k = U_{*k} L_{mk} \tag{19}$$

The errors related to wind velocities ($U_{*k}$) and mixing length ($L_{mk}$) determine the transport contribution to the model errors. The horizontal wind error magnitude can be derived from the analysed wind errors. The continuity equation can be applied to estimate the vertical wind errors from the horizontal wind uncertainties.

The formulation of the model error dynamics described above has several useful properties. First, it supplies a natural "physical" mechanism for the model error growth from wind errors and state dependent spatial tracer gradients. The error growth is expected to be large in regions of strong tracer gradients and/or large wind uncertainties. Such regions should become prime targets for making new chemical observations. Second, this approach naturally introduces the concept of an anisotropic model for the state dependent error specification. It can be viewed as a theoretical basis for the empirical anisotropic forecast error correlation models discussed in Riishøjgard (1998). Third, this approach naturally links the adjoint approach (chapter *Variational Assimilation*, Talagrand) with the suboptimal versions of the Kalman filter framework. The practical implementation of this scheme uses a splitting algorithm for the transport and chemistry. It includes integration of equations for the constituents (Eq. 15) and their variances (Eq. 16), with adaptive calculations of the model errors at every time step (Eqs. 17, 18 and 19).

To illustrate the capabilities of this approach to represent errors in data assimilation, we used a 2-D CTM driven by the January 1992 non-divergent components of the United Kingdom Met Office winds. The chemical solver and its adjoint for 18 species have been described in Khattatov et al. (1999). The horizontal resolution of the CTM is 2.5°× 2.5°. The time step is 20 min. The monotonic mass

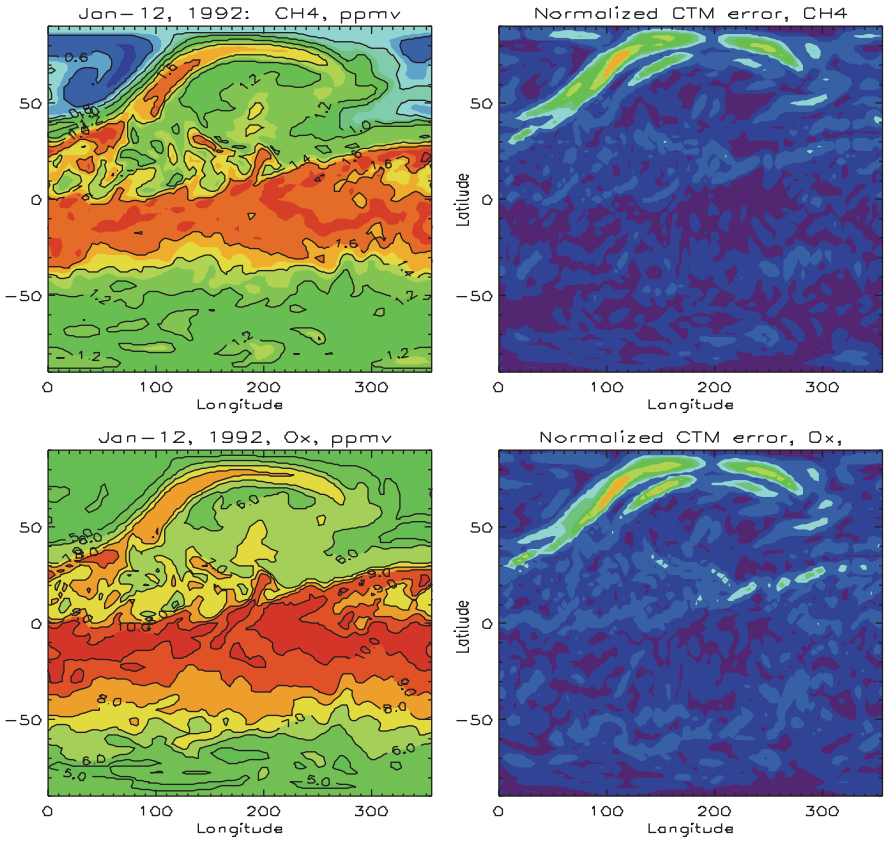



**Fig. 9** Global distributions of mixing ratios (*left*) and normalized model errors (*right*) for methane (*top*) and $O_x$ (odd oxygen, $O + O_3$) (*bottom*) at 32 km on 12 January 1992. *Blue* denotes relatively low values; *orange*/*red* denotes relatively high values. *x*-axis is longitude; *y*-axis is latitude

conserving flux form semi-Lagrangian (FFSL) transport scheme of Lin and Rood (1997) is employed for the tracer and its variance advection. The correlation length is also transported by the FFSL-scheme. The total period of integration was 1 month starting 1 January 1992. The initial conditions for constituents correspond to the zonal mean distribution of the SUNY-SPB 2-D photochemical model (Smyshlyaev et al. 1998; Yudin et al. 2000).

The analysed wind errors were selected arbitrarily, because in the standard Met Office wind product there is no information on the wind errors. We assumed a 15% value in the relative wind error for a series of different mixing correlation lengths ($L_{mk}$= 125, 250, 400 km). These two parameters (relative wind error and mixing correlation length) can be chosen in a more realistic fashion for future applications. Together, these two parameters control the rate of the model error growth and its distribution.

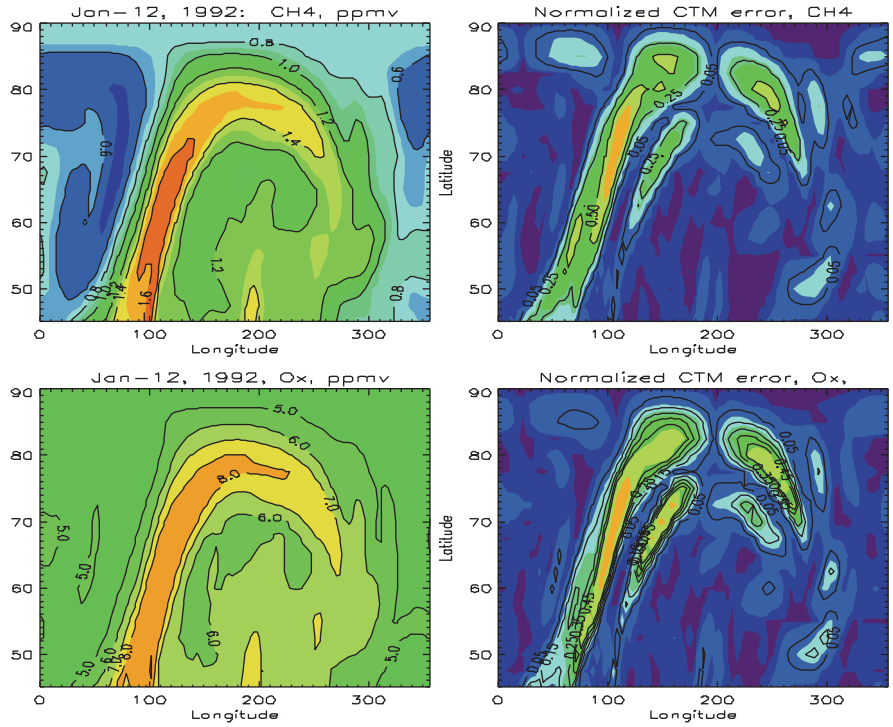



**Fig. 10** As Fig. 9, but zoomed in at the region of strong meridional gradients. *Blue* denotes relatively low values; *orange*/*red* denotes relatively high values

Figures 9 and 10 illustrate computed spatial distributions of $CH_4$ and odd oxygen and the corresponding model error source term for 12 January at 32 km. One can clearly see that, as expected, model error values are large in the regions where there is filamentary intrusion of the subtropical air into the mid and high latitudes, and where strong meridional tracer gradients are observed.

## 5 Discussion

The primary objective of this chapter is illustration of several techniques for computing approximations to the probability density functions (PDFs) in the case of chemistry-transport models (CTMs). The main problem is the large dimensionality of most practically interesting cases. This precludes computing the evolution of the PDF explicitly and forces one to adopt a number of simplifications. Primarily, these involve assuming Gaussian distribution of errors and the ability to linearize the underlying model for reasonably short time periods. Still, even under these assumptions, the extended Kalman filter (EKF) formalism commonly used for computing

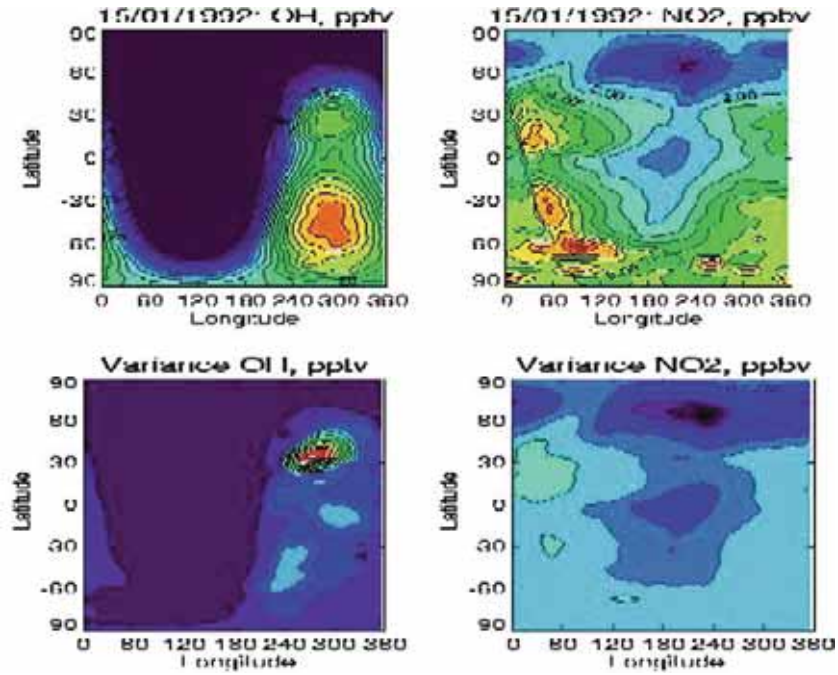



**Fig. 11** Global distributions of mixing ratios (*top panels*) and model variances (*bottom panels*) for OH (*left panels*, parts per trillion by volume, pptv) and $NO_2$ (*right panels*, parts per billion by volume, ppbv) at 32 km on 15 January 1992. *Blue* denotes relatively low values; *red* denotes relatively high values. *x*-axis: longitude; *y*-axis: latitude

the evolution of the error covariance matrices is impossible to implement in practice due to the high dimensionality of the underlying system.

If one decouples chemical and transport processes, it is feasible to explicitly compute the evolution of the error covariance matrices for the chemical processes. It is also possible to explicitly compute the evolution of the variance (the diagonal of the error covariance matrix) for the transport portion of the problem with seemingly satisfactory results.

In finding numerical solutions to the mass conservation equations in CTMs it is common to solve separately for chemical and transport processes. However, within the limitations of the framework discussed in this chapter, one cannot explicitly compute the evolution of the error covariance matrices by separating the two steps. One possible approach to approximating these matrices is to compute full error covariance matrices at the "chemical" step, then discard information about the error covariances and only pass on the diagonals of these matrices to the "transport" step. While clearly limited, this approach could bring us closer to being able to approximate the temporal evolution of the coupled transport-chemistry error covariance matrices.

Examples of the distributions of OH, and $NO_2$ (Fig. 11) and their variances (see experiment described in Sect. 4) show intrusion of mid latitude air into the winter

polar regions as well as the position of the solar terminator in the OH distributions at ~32 km altitude.

Note that, as expected, for short-lived species (OH, $NO_2$) the resulting variance fields in Fig. 11 do not resemble the shapes of the simulated concentration fields due to chemical coupling between species. For $NO_2$, the strong influence of the chemical loss terms of the forecast variance equation in the daytime regions results in substantial decrease of the concentration errors.

Significant research remains to be done (and in all likelihood accompanied by increases in computational power) to develop an adequate framework for proper treatment of full error covariance matrices in chemistry transport modelling. This chapter aims to give the reader some ideas of how to approach this problem in practice without attempting to provide a rigorous or complete formulation. It is likely that advanced ensemble filter methods combined with parallel computing represent a reasonable avenue for practical solution to the treatment of errors in this area of the atmospheric sciences.

**Acknowledgments** A large portion of this work has been funded by NASA's UARS grant UARS "Towards Interactive Three-dimensional chemical data assimilation".

## References

Cohn, S.E., 1993. Dynamics of short-term univariate forecast error covariances. *Mon. Weather Rev*., **121**, 3123–3149.

Gembicki, F.W., 1974. *Vector Optimization for Control with Performance and Parameter Sensitivity Indices*. Ph.D. Dissertation, Case Western Reserve University, Cleveland, OH.

Jazwinski, A.H., 1970. *Stochastic Processes and Filtering Theory*, Academic Press, New York, 376 pp.

Khattatov, B.V., 2003. Multivariate chemical data assimilation. In *Data Assimilation for the Earth System.* NATO Science Series: IV. Earth and Environmental Sciences 26, Swinbank, R., V. Shutyaev and W.A. Lahoz, Kluwer Academic Publishers, Dordrecht, The Netherlands, pp 279–288, 378pp.

Khattatov, B.V., J.C. Gille, L.V. Lyjak, G.P. Brasseur, V.L. Dvortsov, A.E. Roche and J. Waters, 1999. Assimilation of photochemically active species and a case analysis of UARS data. *J. Geophys. Res*., **104**,18715–18737.

Khattatov, B.V., L.V. Lyjak and J.C. Gille, 2001. On applications of photochemical models to the design of measurement strategies. *Geophys. Res. Lett*.,**28**, 2377–2381.

Lin, S.-J., and R.B. Rood, 1997: An explicit flux-form semi-Lagrangian shallow-water model on the sphere. *Q. J. R. Meteorol. Soc*., **123**, 2477–2498.

Lorenc, A.C., 1986. Analysis methods for numerical weather prediction. *Q. J. R. Meteorol. Soc*., **112**, 1177–1194.

Lyster, P.M., S.E. Cohn, R. Ménard, L.-P. Chang, S.-J. Lin and R. Olsen, 1997. An implementation of a two-dimensional filter for atmospheric chemical constituent assimilation on massively parallel computers. *Mon. Weather. Rev*., **125**, 1674–1686.

Ménard, R. and L.-P. Chang, 2000. Stratospheric assimilation of chemical tracer observations using a Kalman filter. Part II: Chi-square validated results and analysis of variance and correlation dynamics. *Mon. Weather Rev*., **128**, 2672–2686.

Ménard, R., S.E. Cohn, L.-P. Chang and P.M. Lyster, 2000. Stratospheric assimilation of chemical tracer observations using a Kalman filter. Part I: Formulation. *Mon. Weather Rev*.,**128**, 2654–2671.

Riishøjgaard, L.P., 1998. A direct way of specifying flow-dependent background error correlations for meteorological analysis systems. *Tellus*, **50A**, 42–57.

Smyshlyaev, S.P., V.L. Dvortsov, M.A. Geller and V.A. Yudin, 1998. A two-dimensional model with input parameters from a GCM: Ozone sensitivity to different formulations for the longitude temperature variation. *J. Geophys. Res.*, **103**, 28373–28387.

Yudin, V.A., S.P. Smyshlyaev, V.L. Dvortsov and M.A. Geller, 2000. Transport diagnostics of GCMs and implications for 2D Chemistry-Transport Model of Troposphere and Stratosphere. *J. Atmos. Sci.*, **57**, 673–699.

# Constituent Assimilation

**William Lahoz and Quentin Errera**

## 1 Introduction

*Background.* In the 1990s, following years of development of meteorological data assimilation by the Numerical Weather Prediction (NWP) community, the data assimilation methodology began to be applied to constituents, with a strong focus on stratospheric ozone (Fisher and Lary 1995; Rood 2005; Lahoz et al. 2007a). Because of its comparatively later application, constituent data assimilation is less mature than meteorological data (henceforth NWP) assimilation. Nevertheless, there has been substantial progress over the last 15 years, with the field evolving from initial efforts to test the methodology to later efforts focusing on products for monitoring ozone and other constituents. More recently, the production of ozone forecasts by a number of operational centres has become routine. A notable feature of the application of the data assimilation methodology to constituents has been the strong interaction between the NWP and research communities, for example, in the EU-funded ASSET project (Lahoz et al. 2007b). A list of acronyms can be found in the *Appendix*.

There are differences between NWP and constituent data assimilation that affect the way the assimilation is set up in the latter. These are (see also Eskes 2006; Lahoz et al. 2007a):

- Constituent data assimilation is less mature than NWP data assimilation (see chapter *Assimilation of Operational Data*, Andersson and Thépaut). An example of this concerns parametrizations of ozone chemistry due to Cariolle and Déqué (1986). They have been used to assimilate stratospheric ozone in the last 5 years or so, but it is only very recently that the performance of these schemes, and their associated errors, has been assessed in the data assimilation context (Geer et al. 2007);

W. Lahoz (✉)
Norsk Institutt for Luftforskning, Norwegian Institute for Air Research, NILU, Kjeller, Norway
e-mail: wal@nilu.no

W. Lahoz et al. (eds.), *Data Assimilation*, DOI 10.1007/978-3-540-74703-1_18,

- NWP is primarily an initial value problem, whereas tropospheric constituent data assimilation is determined by boundary conditions, emissions and removal processes. For stratospheric constituent data assimilation, although sources and sinks may need to be considered, the evolution of the model state is primarily controlled by the initial state;
- Improvements in NWP can be achieved by more accurate specification of dynamical variables such as temperature, winds and moisture. By contrast, in chemical weather (see chapter *Inverse Modelling and Combined State-Source Estimation for Chemical Weather*, Elbern et al.), in general concerning the troposphere, a better forecast is achieved mainly by a better description of sources and sinks. For stratospheric constituents, a better forecast can be achieved both by a better description of dynamical variables (and hence transport of the constituent), and by a better description of sources and sinks (if applicable);
- The time-scales relevant for NWP are order of days. For chemistry, there is a very wide range of time-scales, from decades (carbon dioxide) to seconds for very short-lived species (see chapter *Introduction to Atmospheric Chemistry and Constituent Transport*, Yudin and Khattatov). Residence times of species and removal time-scales vary substantially from one species to another;
- The spatial scales for chemical weather have a wider range than for NWP: detailed air quality forecasts are made from the hemispheric scale down to the scale of individual streets. Often, a hierarchy of chemical models is introduced to describe these different spatial scales;
- Chemical equation systems are stiff, i.e., they include reactions with rates varying by several orders of magnitude. This requires the use of sophisticated numerical integration schemes, called stiff solvers. Stiffness manifests itself in strong error correlations between species, and can cause error covariance matrices to become singular. Constituent data assimilation algorithms must aim to account for these features;
- Arguably, a combined initial value and source estimation approach is required for forecasting air quality. This is an extension of the initial value approach in NWP. Allied to these developments, inverse modelling techniques (see chapter *Inverse Modelling and Combined State-Source Estimation for Chemical Weather*, Elbern et al.) may be of use in improving the description of sources and sinks;
- The availability of useful satellite observations of tropospheric and stratospheric composition is still relatively limited compared to the availability of observations of dynamical variables for NWP. Retrieval algorithms for tropospheric constituents need further development and evaluation. Retrieval algorithms for stratospheric constituents are, however, reasonably well established, especially in comparison with the situation for tropospheric constituents;
- The Global Observing System (see chapter *The Global Observing System*, Thépaut and Andersson) for NWP is more mature than for constituents. This is reflected in that there are less operational instruments for constituents than for NWP. Many satellite constituent observations are classed as “research” or “pre-operational” (see chapter *Research Satellites*, Lahoz), which means that, compared to operational NWP observations, they are usually not available in

near-real-time; the reliability of data supply is often less robust; and observational errors may be larger, or less well understood and characterized;

- For NWP the numerical dimension of the problem is extremely large; the typical dimension of current NWP models is of order $10^7$, while the number of observations available over 24 h is currently of order $10^6$–$10^7$ (see chapter *The Global Observing System*, Thépaut and Andersson). For constituents, the number of data assimilated is generally an order of magnitude less than for NWP because fewer instruments are used, with fewer soundings per instrument. In both cases, however, the large dimension of the problem causes a range of practical difficulties, influencing the practical implementation of assimilation systems;
- The dimensionality of the state of stratospheric chemical models is much higher than that of the NWP models. Assuming the same number of grid points, stratospheric constituent models typically need to follow between 20 and 100 different species, i.e., variables, per grid point, as compared to under a dozen variables for a NWP model.

One important difference between NWP and constituent data assimilation is worth emphasizing. In principle, given accurate initial conditions, sources and sinks and accurate dynamics, it would be possible to model constituent distributions many months without constituent data assimilation. Furthermore, in chemistry, many situations can be modelled as a relaxation to an equilibrium state. This is very different to the chaotic system involved in dynamical data assimilation.

This does not mean that constituent data assimilation is unnecessary. Constituent data assimilation is needed to: (1) infer the constituent's initial conditions (we can only ever get these, imperfectly, from observations); (2) correct for imperfectly known reaction rates; (3) correct for imperfectly modelled chemistry (e.g. not enough species, not enough reactions described, or approximate parametrizations are needed); (4) correct for unknown source terms (e.g. tropospheric pollution, troposphere-stratosphere transport); and (5) most importantly of all at the moment, correct for errors in constituent transport, such as excessive Brewer-Dobson circulations in analysed wind fields, or errors in temperature fields. Constituent data assimilation can thus be regarded as a way of providing accurate initial conditions (point 1), and as a way of confronting models with observations in order to evaluate them and, in particular, correct model bias (points 2–5). The latter objective shows that constituent data assimilation is a different kind of problem compared to NWP data assimilation, where the goal is to get accurate initial conditions.

*Motivation.* The main constituents assimilated are humidity (chiefly in the troposphere) and ozone (chiefly in the stratosphere). Humidity (or water vapour) is assimilated in the troposphere by NWP centres, but only now is it starting to be assimilated in the stratosphere. This is chiefly due to its important role in the radiation budget of the atmosphere, especially in the upper troposphere/lower stratosphere (UTLS) region, because it provides information on the atmospheric circulation, because it is a source of $HO_x$ (involved in the catalytic destruction of ozone), and because it is a constituent of the Polar Stratospheric Clouds (PSCs) involved in polar ozone loss (Dessler 2000).

The main aims for assimilating stratospheric ozone include the development of ozone and UV-forecasting capabilities; the need to monitor stratospheric ozone to track the evolution of the stratospheric composition, mainly ozone and the gases that destroy it (WMO 2006), and assess compliance with the Montreal protocol; and the need to evaluate the performance of instruments measuring ozone, especially those providing long-term datasets (e.g. TOMS, GOME). The assimilation of ozone is also important for technical reasons, including: the constraints ozone observations provide on other constituents; the use of assimilation techniques to evaluate models and ozone observations; the development of computer code to assimilate instrument radiances sensitive to temperature and constituents; and the dynamical information provided by ozone tracer distributions.

Other stratospheric constituents besides ozone that are of interest include $H_2O$, $N_2O$, $CH_4$, $NO_2$, $HNO_3$, ClO, BrO and aerosol (see IGACO 2004 for a more complete list). Tropospheric constituents besides humidity are of interest for monitoring and forecasting air quality; examples include ozone, $NO_2$, CO, formaldehyde, $SO_2$ and aerosols.

In NWP, the main motivation for constituent assimilation has been the use of constituent information (in particular, water vapour and stratospheric ozone) to improve the weather forecast (see chapter *Assimilation of Operational Data*, Andersson and Thépaut). More recently, efforts in NWP have been motivated by the need to monitor and forecast air quality.

*Methodology.* Tests of the data assimilation methodology have involved many demonstrations that a number of techniques could be used to combine constituent information from a model and from observations. These models have been generally based on general circulation models (GCMs) or on chemical models, either chemistry-transport models (CTMs) or photochemical box models. Their chemistry representation varies from treating constituents as tracers to sophisticated photochemical packages incorporating aerosols and heterogeneous chemistry (see chapter *Introduction to Atmospheric Chemistry and Constituent Transport*, Yudin and Khattatov).

Early examples of the assimilation techniques demonstrated include nudging (Austin 1992), variational methods (Fisher and Lary 1995) and sequential methods based on variants of the Kalman filter (Khattatov et al. 1999) – the chapters in Part I, *Theory*, provide details of these data assimilation techniques. Austin used a CTM, whereas Fisher and Lary, and Khattatov et al. used a photochemical box model. Following on from these efforts, assimilation into chemical models has been used to test chemical theories (Lary et al. 2003; Marchand et al. 2003, 2004); to produce ozone analyses, either height-resolved or total column (Levelt et al. 1998; El Serafy et al. 2002; Eskes et al. 2003; Štajner and Wargan 2004; Massart et al. 2004, 2009; Segers et al. 2005; Wargan et al. 2005; Rösevall et al. 2007a, b); to produce analyses of other chemical species besides ozone, including $NO_2$, $CH_4$, $N_2O$, water vapour and aerosols (Khattatov et al. 2000; Fonteyn et al. 2000; Ménard et al. 2000; Ménard and Chang 2000; Errera and Fonteyn 2001; Chipperfield et al. 2002; El Amraoui et al. 2004; Errera et al. 2008; Thornton et al. 2009); and to design constituent measurement strategies (Khattatov et al. 2001) (Note that these examples are

not exhaustive.). Recent reviews include those by Lary (1999), Wang et al. (2001), Khattatov (2003), Rood (2005) and Lahoz et al. (2007a).

CTMs are used to forecast ozone and other constituents operationally; examples include the Royal Netherlands Meteorological Institute, KNMI (Eskes et al. 2002, 2005; El Serafy and Kelder 2003), the Global Modeling Assimilation Office, GMAO (Riishøjgaard et al. 2000; Štajner et al. 2001) and the Belgian Institute for Space Aeronomy, BIRA-IASB (http://bascoe.oma.be/archives). CTMs are also used to monitor observations; for example, Štajner et al. (2004) describes the use of data assimilation at the GMAO to monitor SBUV/2 data.

The above examples focus on the stratosphere. Models incorporating chemistry are increasingly being used for research on tropospheric pollution and air quality. Examples (not exhaustive) include the demonstration that data assimilation can improve analyses of tropospheric pollution (Elbern and Schmidt 2001), and that inverse modelling can provide estimates of tropospheric emissions like carbon monoxide (Müller and Stavrakou 2005) or methane (Meirink et al. 2006). More generally, inferring sources and sinks of constituents using inverse modelling provides information on transcontinental pollution (e.g. Pétron et al. 2004), air quality (e.g. Blond and Vautard 2004) and national greenhouse gas inventories (e.g. Bergamaschi et al. 2005).

Examples of GCM-based NWP models used for constituent data assimilation include the European Centre for Medium-Range Weather Forecasts (ECMWF) model, where ozone has been assimilated for forecasts (Dethof 2003) and reanalyses (Dethof and Hólm 2004), and where research has been done on the assimilation of limb infrared radiances sensitive to ozone and humidity (Bormann et al. 2005, 2007; Bormann and Healy 2006; Bormann and Thépaut 2007); and the UK Met Office model, where ozone has been assimilated for research (Jackson and Saunders 2002; Struthers et al. 2002; Jackson 2004, 2007; Lahoz et al. 2005, 2007a, b; Geer et al. 2006a, b, 2007; Mathison et al. 2007; Jackson and Orsolini 2008). More recently, water vapour has been also assimilated for research at both ECMWF and the Met Office (Lahoz et al. 2007a, b; Thornton et al. 2009).

At the GMAO, ozone assimilation into the GEOS-4 GCM started with SBUV/2, POAM-III and ILAS-II data (Štajner et al. 2006) and, more recently, includes EOS Aura OMI and MLS data (Štajner et al. 2008). Currently, at the GMAO there is near real time assimilation of ozone from Version 8 SBUV/2 retrievals of ozone layers into the GEOS-5 GCM (Rienecker et al. 2008). Polavarapu et al. (2005a, b), using the Canadian middle atmosphere model (CMAM), discuss the role of dynamics on analysed stratospheric constituents, including ozone. As will be discussed in Sect. 2, in the NWP-based approach the use of simplified chemistry is the norm. An exception is the CMAM model, where full chemistry is used.

Other constituents besides ozone and water vapour have been assimilated into GCM-based systems. For example, total column carbon monoxide, CO, observations from SCIAMACHY have been assimilated into the GEOS-4 GCM-based system (Tangborn et al. 2009); radiance observations from AIRS have been used in the ECMWF data assimilation to constrain $CO_2$ mixing ratios (Engelen et al. 2009).

Both chemical model and GCM-based data assimilation approaches have been used to evaluate models and observations, in particular concerning ozone (e.g. Geer et al. 2006a, b, 2007; Coy et al. 2007). Data assimilation not only corrects weaknesses in models, but also identifies model deficiencies such as biases, e.g., between model and observations, and between different observations. Bias estimation and correction is likely the greatest current challenge in data assimilation (Rood 2005) – see chapter *Bias Estimation* (Ménard).

The above body of work shows that constituent data assimilation is feasible, provides fields for monitoring the atmosphere; provides a way of evaluating models and observations, including bias; provides initial fields for constituent forecasts; and can be used to infer constituent emissions (in the context of inverse modelling). In this way, constituent data assimilation adds value to the information provided by the observations and the models. Key to the success of constituent data assimilation is that, as for NWP, it is an objective method based on mathematical principles (see chapter *Mathematical Concepts of Data Assimilation*, Nichols).

From the point of view of an end-to-end approach to Earth Observation, from mission pre-launch to mission post-launch, constituent assimilation also adds value to the information provided by observations and models. In mission pre-launch, data assimilation activities such as Observing System Simulation Experiments (OSSEs; see chapter *Observing Simulation System Experiments*, Masutani et al.) provide information on the quality of additions to the Global Observing System and the data assimilation systems incorporating this data. In mission post-launch, data assimilation activities provide information on the quality of constituent observations and models, and of constituent analyses, in a number of ways: evaluation of observations and models (e.g. Štajner et al. 2004; Geer et al. 2006a, b, 2007); Observing System Experiments (OSEs) to assess the incremental value of existing observations (e.g. Struthers et al. 2002); analyses of constituents such as ozone and humidity (e.g. Lahoz et al. 2007a, b); and ozone reanalyses such as those from ERA-40 (Dethof and Hólm 2004).

This chapter discusses current approaches used in constituent data assimilation. We first discuss GCM-based data assimilation; then chemical model data assimilation. We then discuss evaluation of models, observations and analyses, and applications. Finally, we discuss future developments and identify potential key drivers.

## 2 GCM-Based Approaches

### *2.1 Introduction*

An NWP model is a complex numerical model designed to simulate the evolution of the atmospheric state over the length of a weather forecast (typically for a few hours up to 2 weeks in the future). The dynamical core of the model is concerned with solving the Navier-Stokes equations (or an approximation thereto) that

govern the evolution of atmospheric winds and mass fields. The NWP dynamical core must solve for humidity, as the Navier-Stokes equations are formulated with moisture terms included. This means that mature humidity data assimilation code has already been developed in operational NWP systems. Additional stratospheric humidity data assimilation efforts must build on this code without unduly interfering with the assimilation of tropospheric humidity data. Details are provided in Sect. 2.2.

The equations are typically solved using finite difference or spectral methods. Numerical models include parametrizations of a range of atmospheric physical processes, including the formation of clouds, production of rainfall, interactions of the flow with orography and radiative transfer processes, and, increasingly, chemistry (see chapter *The Role of the Model in the Data Assimilation System*, Rood).

There is a strong common heritage linking NWP models with GCMs used for global climate simulations (e.g. Trenberth 1992). In some cases, the same basic model is run in different configurations for both NWP and climate simulations (e.g. the Met Office Unified Model; Davies et al. 2005). The most complex atmospheric GCMs are coupled with sophisticated models of the ocean and land surface, to form Earth System models.

The development of data assimilation techniques in NWP has been strongly focused by the pressure for continual improvements in weather forecasting. More sophisticated techniques have led to much better use of observations to provide the initial conditions for operational weather forecasts. A particular example is that modern variational data assimilation systems commonly assimilate satellite soundings as radiances rather than retrievals, allowing better use of the information.

NWP data assimilation has focused on observations of variables such as temperature, pressure, winds and humidity. Because of the vital role played by water vapour in the atmosphere, its assimilation has always played a key role in NWP. The treatment of water vapour, cloud and rainfall in NWP is a vast subject (see Hólm et al. 2002, and references therein), which we are not going to address in this chapter.

We focus on stratospheric humidity (tropospheric humidity is discussed in chapter *Assimilation of Operational Data*, Andersson and Thépaut), and on the stratospheric constituent that has received most attention over the past decade, ozone (Rood 2003, 2005).

## *2.2 Assimilation of Humidity*

*Background.* In this section, we highlight some of the key issues concerning the assimilation of stratospheric humidity. First, the stratosphere is very dry; while condensation of water vapour is commonplace in the troposphere, clouds (PSCs) only form in the stratosphere in the polar night, where extremely cold temperatures occur. The water vapour mixing ratio varies by many orders of magnitude, from a few percent (by mass) in the tropical lower troposphere to a few parts per million in the stratosphere.

A second key issue is the available observations of water vapour. The primary source of moisture measurements is the radiosonde network. Radiosondes carry sensors that are primarily designed to measure the high relative humidity (*RH*) typical of the lower and middle troposphere. Where the humidity is low and temperature cold, as in the stratosphere, the measurements become less accurate (relatively, if not absolutely). Thus, routine radiosonde humidity measurements are of little or no use in the stratosphere, even if the sondes reach that level. More recently, satellite data have become more widely available, and are now used as an integral part of the operational assimilation of moisture information (e.g. ATOVS and SSM/I). However, the operational nadir soundings have relatively poor vertical resolution.

*Implementation.* The large variation in humidity between the surface and the stratopause (4–5 orders of magnitude), together with different priorities in the troposphere (description of precipitation and identification of clouds) and the stratosphere (description of tracer distributions), means that it is difficult to specify a control variable (see chapter *Mathematical Concepts of Data Assimilation*, Nichols) suitable for use throughout the domain of models that span this region.

Dee and da Silva (2003) introduce a "pseudo-relative humidity" ($RH^*$), defined by scaling the mixing ratio $q$ by the saturation mixing ratio of the background field. An advantage of this approach is that a univariate $RH^*$ analysis preserves $q$ in the absence of moisture observations. By contrast, using unmodified *RH* as a control variable implies a change in scaling if the temperature is changed, leading to changes in $q$ in the absence of moisture observations. In the presence of multivariate observations (e.g. temperature and moisture), this approach produces analysed humidity values that are close to those produced by a *RH* analysis.

In a parallel development, Hólm et al. (2002) introduced a normalized *RH* control variable, in which *RH* is divided by (an approximation of) the background variability. The new control variable has background errors that are more nearly Gaussian and homogeneous. Using normalized *RH*, the assimilation scheme also takes better account of the large variability in the background error covariance matrix. This should improve the interpretation of humidity data, and the mapping of information from radiances into temperature and humidity fields. Initial tests have been encouraging.

Further developments are currently under way at a number of NWP centres (e.g. ECMWF, Met Office), with the aim of developing an approach to moisture assimilation that performs well in both troposphere and stratosphere, in dry conditions and close to saturation (see Lahoz et al. 2007a, b; Thornton et al. 2009).

## *2.3 Assimilation of Ozone*

*Background.* The main motivation for the inclusion of ozone data assimilation in operational NWP has been to take better account of ozone (in particular stratospheric ozone) when assimilating satellite radiance data, mainly from nadir sounding instruments (see chapter *Assimilation of Operational Data*, Andersson

and Thépaut). Radiance assimilation has been shown to improve the overall skill of weather forecasts (Saunders et al. 1999; McNally et al. 2006). Many of the channels used for atmospheric temperature sounding are at least partially sensitive to ozone, so improvements in the accuracy of ozone profiles can lead to more accurate temperature inversions.

At the same time, the assimilated ozone data can be used by the model radiation scheme, potentially leading to better radiative forcing of the model. Model radiation schemes take into account the absorption and emission of both short-wave (visible and near-UV) and long-wave (infrared) radiation by a number of atmospheric constituents. In the stratosphere, ozone is the dominant contributor to radiative heating, but the values are generally taken from ozone climatologies (e.g. Fortuin and Kelder 1998). An estimate of the true ozone distribution is likely to improve these calculations.

At ECMWF, ozone is already included in the forward modelling of satellite radiances. Experiments at ECMWF, using analysed ozone in heating rate calculations, found that variations in ozone amounts of $\sim$10% could result in changes in analysed UTLS temperatures of 2–4 K (Cariolle and Morcrette 2006). Model runs with comprehensive chemistry and fully interactive ozone show significant temperature differences of up to 3 K in the upper stratosphere and lower mesosphere, compared with those with climatological ozone (Sassi et al. 2005). A prognostic ozone field allows the modelling of feedbacks between radiation, chemistry and dynamics, and this is expected to improve forecasts, especially over longer time-scales. However, work by Morcrette (2003) suggests that coupling of the analysed ozone with the radiation scheme does not always bring improvement, and Cariolle and Morcrette (2006) state that in order to adequately represent the ozone radiative heating in the UTLS, ozone profiles with a vertical resolution of $\sim$1 km need to be assimilated. Recent experiments at the Met Office have shown that the inclusion of ozone-radiation feedbacks leads to an increase in the quality of tropospheric temperature, wind and geopotential height forecasts (Mathison et al. 2007). However, these changes are small and as yet not well understood, and the greatest impact of the ozone-radiation feedback is on analysed and forecast temperatures near the stratopause. In recent assimilation experiments in Canada, de Grandpré et al. (2009) use a coupled model that includes a comprehensive stratospheric chemistry scheme. They show that incorporating the radiative feedback from ozone can improve temperature predictability throughout the stratosphere.

An additional motivation for ozone assimilation is that the motion of ozone in the atmosphere could give useful dynamical information. Daley (1995) pointed out the feasibility of estimating the wind field from constituent observations, given sufficiently dense, frequent and accurate measurements. Riishøjgaard (1996) demonstrated the use of ozone measurements to reconstruct the flow field in a barotropic vorticity equation model. Peuch et al. (2000) demonstrated the dynamical impact of total ozone column observations in OSSEs using a 4D-Var data assimilation system. Recently, Semane et al. (2009) have provided evidence that assimilation of ozone observations from EOS Aura MLS together with operational observations can improve lower stratospheric wind fields. However, the use of ozone data to

infer dynamical information is not without its problems. An inappropriately specified background error covariance matrix can lead to unrealistic impacts of ozone measurements on the wind fields. So, in practice, many ozone assimilation systems treat ozone as a univariate variable, i.e., its background errors are uncorrelated with those of other variables.

A further motivation for ozone assimilation is UV forecasting. Burrows et al. (1994) set up a system for operational UV forecasts in Canada. First, a field of total column ozone over the Northern Hemisphere is calculated using climatological total ozone column data, modified using regression relationships with a range of meteorological forecast fields (including vorticity, temperature and geopotential height) in the upper troposphere and stratosphere. Second, the total column ozone is corrected to fit ozone measurements over Canada. Finally, the clear-sky UV index is calculated using the solar zenith angle and day of the year. Other operational centres have developed similar systems (Austin et al. 1994, for the Met Office). An operational ozone data assimilation system could be used to replace the first two steps of the procedure, with potentially better accuracy. The Australian Bureau of Meteorology already does something similar (Lemus-Deschamps et al. 2005), using a simplified analysis and forecast of TOVS total column ozone. This system, and that used at the National Centers for Environmental Prediction, NCEP (Long 2003) have the benefit of using a radiative transfer model to calculate the surface UV, rather than the empirical methods used in Canada and the UK.

*Representation of chemistry.* The minimum number of species typically needed for a good representation of chemistry range from ∼25–50 in the stratosphere to ∼50–100 in the troposphere. This reflects the increased complexity of chemistry in the troposphere compared to the stratosphere.

In the stratosphere, ozone has a life-time ranging from ∼100 days (lower stratosphere) to less than 1 day (upper stratosphere) (Dessler 2000). Except in the upper stratosphere, these time-scales are relatively long compared to the length of a typical weather forecast, which is of the order of days. So, in that context, the full treatment of chemical sources and sinks has not been a priority. Indeed, the use of a complex representation of ozone chemistry in an NWP system would be judged an unjustified overhead. Instead, the usual approach has been to implement simplified representations of ozone production and loss processes.

In early data assimilation systems, any representation of chemistry was omitted and ozone was treated as a passive tracer. Because ozone behaves as a passive tracer in the lower stratosphere (except under ozone hole conditions), this approach can provide useful information on the stratospheric ozone distribution (Polavarapu et al. 2005a, b). More recent developments have incorporated simple linear parametrizations of the chemical sources and sinks of ozone, typically known as Cariolle schemes (Cariolle and Déqué 1986; McLinden et al. 2000; McCormack et al. 2004; Cariolle and Teyssèdre 2007).

In the Cariolle scheme, the rate of change of ozone due to photochemistry ($C$) is written as a first-order Taylor series expansion:

$$C = a + b(\chi - \chi_0) + c(T - T_0) + d(\Phi - \Phi_0). \quad (1)$$

The first term in Eq. (1), $a$, is the equilibrium production minus loss, at the appropriate level and latitude. The second term accounts for differences between the current ozone amount $\chi$ and its equilibrium value, and the third for differences in the temperature, $T$. The last term allows for solar radiation by considering the effect of the total ozone column $\Phi$ above the point under consideration. The coefficients *a, b, c* and *d* in Eq. (1), as well as the equilibrium values, are derived from a full chemistry model (usually a 2-D model), so the parametrized photochemistry is highly dependent on the particular model used. Geer et al. (2007) compare results from a range of linear chemistry ozone parametrizations and highlight some large differences.

The Cariolle schemes, contrary to some perceptions that they are non-rigorous, are actually based on sound photochemical arguments (see McCormack et al. 2006, for more details). Equation (1) springs directly from a linearized expansion of the fundamental odd-oxygen photochemical production and loss rate equations. This was done initially for pure oxygen (Chapman) photochemistry (Lindzen and Goody 1965), and subsequently extended to reactions involving nitrogen, hydrogen and chlorine species (Blake and Lindzen 1973; Stolarski and Douglass 1985).

The scheme described by Eq. (1) does not take into account heterogeneous ozone chemistry, which is dominant under ozone hole conditions (Dessler 2000). To remedy this shortcoming, the approach expressed in Eq. (1) is modified to include a "cold tracer" to parametrize ozone loss due to heterogeneous processes (Hadjinicolau et al. 1997; Eskes et al. 2003). The cold tracer approach is not the only means by which heterogeneous ozone loss is represented in ozone data assimilation. Cariolle and Teyssèdre (2007) describe a version of the Cariolle scheme that represents this ozone loss without using a cold tracer, and ECMWF uses a version with this approach, too (Dethof 2003; Dragani and Dee 2008).

The relaxation rate $\tau = -1/b$ can be used to quantify the importance of photochemistry effects. As shown by Geer et al. (2006a, 2007), the values of $\tau$ confirm that in the lower stratosphere ($\tau \sim 100$ days) the photochemistry could be neglected, but in the upper stratosphere ($\tau \sim 0.5$ days) the photochemistry is very important. But, it follows that, if the photochemical coefficients and equilibrium values are not realistic, the ozone data will quickly relax to an incorrect value, ignoring information from observations. In such circumstances, the parametrized chemistry scheme will seriously degrade the assimilated ozone fields in the upper stratosphere, and it may be preferable to omit the chemistry.

Results reported in the "ASSET analysis intercomparison project" (Geer et al. 2006a) where ozone analyses from several GCMs and CTMs are compared for a fixed time period, show that, for current ozone data assimilation systems, with good ozone observations and no chemistry one can get a good representation of the ozone field even when the photochemistry time-scales are fast. However, above 0.5 hPa (about 60 km in altitude), where the ozone diurnal cycle is no longer negligible, only analyses with a detailed representation of mesospheric chemistry capture it. Finally, provided that there are no observational gaps, the complexity of the chemical scheme tends to have little effect on the quality of the ozone analyses. However, these results also show that observational gaps can seriously degrade the ozone

analyses. Arguably, in the upper stratosphere (fast chemical time-scales), a better solution than omitting chemistry would be to bias correct the Cariolle scheme (see, e.g., Coy et al. 2007).

*Implementation.* The first implementation of an ozone assimilation system for operational NWP was at the NCEP (Caplan et al. 1997; Derber et al. 1998). A univariate ozone assimilation was included in the operational ECMWF 4D-Var system in April 2002, and was also part of the 3D-Var system for the ERA-40 reanalysis (Dethof and Hólm 2004; Uppala et al. 2005). ECMWF also currently provide analyses and forecasts of ozone (Dragani and Dee 2008). Of necessity, ozone assimilation systems for NWP are limited to using measurements that are available close to real time. This effectively means data from SBUV/2 (retrievals) and HIRS (channel 9 radiances), both carried by the NOAA polar-orbiter satellites. However, ozone data from research satellites can also be available in close to real time: e.g., ECMWF have assimilated operationally ozone profile data from MIPAS on board ESA's Envisat.

Ozone data that have been assimilated operationally recently by ECMWF include NOAA-16 SBUV/2 partial columns (April 2002–October 2008) and SCIAMACHY total column ozone provided by KNMI (September 2004–December 2008). Following the results from experimental tests, in November 2007 ECMWF started the assimilation of NOAA-17 and NOAA-18 SBUV/2 partial columns and in June 2008 of OMI total column ozone (although this was blacklisted on 27 January 2009 due to instrument problems). Also, ECMWF hope to start assimilation of height-resolved EOS MLS ozone data during 2009 (R. Dragani, personal communication). Finally, at ECMWF, TOMS total column ozone data have been assimilated for reanalyses (Dethof and Hólm 2004), and GOME ozone profiles provided by the Rutherford Appleton Laboratory (RAL) are being assimilated in the ERA-interim reanalysis project (ECMWF 2007). Further details of ozone assimilation at ECMWF can be found in Dragani and Dee (2008).

Ozone assimilation has also been developed at the Met Office, first using the analysis correction scheme (Connew 1999; Struthers et al. 2002), and later 3D-Var (Jackson 2004, 2007; Geer et al. 2006b; Jackson and Orsolini 2008) – see Fig. 1. Other NWP centres, e.g., GMAO and KNMI have taken the approach of developing an ozone analysis in a CTM driven by assimilated wind and temperature data (see Sect. 3 in this chapter). However, as mentioned in Sect. 1, the GMAO have also developed a GCM-based ozone assimilation system. This uses PSAS with GEOS-4 (Štajner et al. 2008), and a Gridpoint Statistical Interpolation (GSI) scheme, a variant of 3D-Var, with GEOS-5 (Rienecker et al. 2008). The GSI scheme is applied in grid-point space to facilitate the implementation of anisotropic inhomogeneous background error covariances.

Some of the satellite instruments used in ozone assimilation give only restricted vertical coverage; for example, HIRS channel 9 is most sensitive to the lower-stratosphere ozone maximum, while SBUV/2 retrievals give some profile information above the ozone peak in the mid stratosphere. For non-operational systems (and, increasingly, operational systems such as that of ECMWF) that assimilate research satellite data from platforms such as ESA's Envisat and NASA's EOS Aura, the

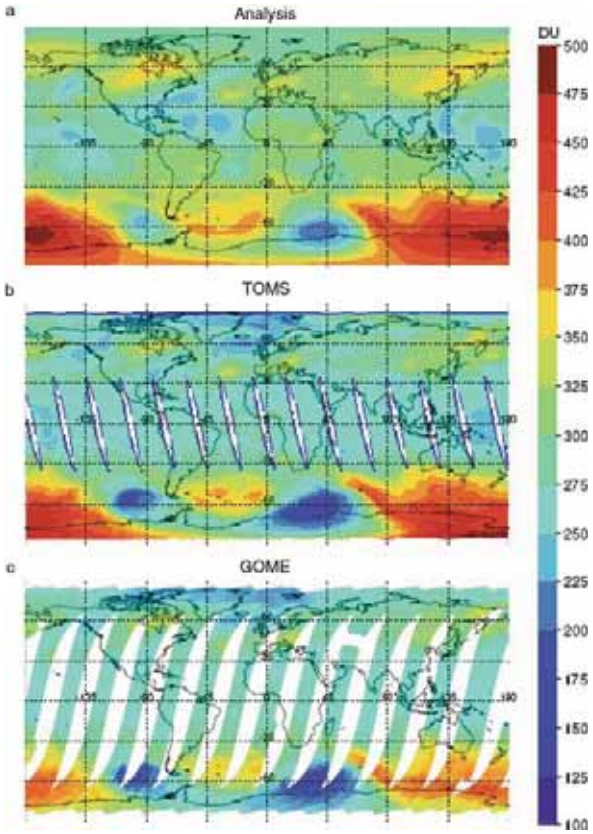



**Fig. 1** Total ozone column on 26 September 2002 (Dobson Units) from (**a**) the 1200 UTC troposphere-stratosphere Met Office analysis with the column ozone below 200 hPa replaced by an ozone climatology; (**b**) TOMS; (**c**) GOME. *Red* indicates relatively high values; *blue* indicates relatively low values. Based on Geer et al. (2006b). The *x*-axis is longitude; the *y*-axis is latitude. © Royal Meteorological Society

situation is better than with traditional operational satellite data (e.g. SBUV/2, HIRS channel 9 radiances). In this case both nadir and limb sounders are used, with the latter providing better vertical resolution because of their viewing geometry.

There is recent evidence that adding height-resolved ozone data improves ozone analyses in an NWP system. In the intercomparison of ozone analyses described by Geer et al. (2006a), it is shown that assimilation of height-resolved MIPAS ozone data improves the ECMWF NWP ozone analyses. This improvement is attributed to the benefit coming from the relatively high vertical resolution of MIPAS, and the fact that before this only limited ozone data were assimilated (namely, SBUV/2 ozone layers and GOME total column ozone). A similar improvement is seen in the Met Office system, where assimilation of height-resolved EOS MLS ozone data reduces analyses errors compared to the situation when only SBUV/2 ozone layers are assimilated (Jackson 2007). These results suggest a way forward toward improved use of ozone data in NWP systems. Along these lines, benefit could be expected from the assimilation of height-resolved ozone data from the Metop IASI instrument, and from the AIRS instrument aboard the EOS Aqua platform. See

Dragani and Dee (2008) for a discussion of experiments at ECMWF involving the assimilation of AIRS infrared channels in the ozone band.

While ozone assimilation systems have focused almost exclusively on satellite data, it would also be possible to use ground based ozone measurements. The main reasons why they are not generally used is first their scarcity and second that they have not been routinely exchanged alongside other meteorological data. Ozonesondes are expensive to make – much more expensive than radiosondes, themselves under economic pressure. As a result, ozonesondes tend to be flown routinely once a week from a very limited number of stations, plus during certain research campaigns, such as *MATCH* (Streibel et al. 2006). While the scarcity of ground-based ozone data means that it is not worthwhile assimilating them routinely, they are a very valuable data set for the validation of ozone assimilation systems. There are a larger number of Dobson measurements of total column ozone, but these have no profile information, as well as being sparse compared to satellite measurements.

## 3 Chemical Model Approaches

For constituent assimilation, there are several good reasons for avoiding the use of NWP models, and instead using what we refer to as the chemical model approach. First, NWP models are complex and generally expensive in terms of computer resources. Second, they tend to focus on the dynamics of the atmosphere, so that, typically, only constituents that interact with the dynamics are represented. This is the case for ozone and water vapour (see Sect. 2). In NWP models, chemistry is commonly parametrized to simplify the system, so that in some cases (to be discussed later) this set-up can be inappropriate.

If the goal is not to improve the weather forecast, other types of model are more appropriate for constituent assimilation. In particular, (i) photochemical box models along an air parcel trajectory, and (ii) three-dimensional CTMs (see chapters *Introduction to Atmospheric Chemistry and Constituent Transport*, Yudin and Khattatov; *Representation and Modelling of Uncertainties in Chemistry and Transport Models*, Khatattov and Yudin). In both these cases, the dynamical problem is simplified because the dynamical fields are pre-calculated from a NWP-based system. In the first case, the trajectory and the atmospheric conditions (temperature, pressure) along it are given and a photochemical box model is used to calculate the evolution of the composition in the transported air parcel. In the second case, wind and temperature fields are prescribed and used to advect the constituents in the model. The chemical scheme used by CTMs varies in complexity and depends on the final application. If the assimilation system focuses on long-lived species (chemistry time-scales >> transport time-scales), e.g. methane in the lower stratosphere, chemistry can be neglected. If the assimilation system focuses on ozone, where both chemistry and transport can be important in the stratosphere, a parametrized chemical scheme can be sufficient. If the assimilation system focuses on reactive, i.e.,

short-lived species (chemistry time-scales << transport time-scales), e.g. $NO_2$ in the stratosphere, then explicit calculation of the chemical interactions is generally necessary. The first two cases are cheaper in computer time than the third one. The cost of computer time is another important factor to consider in constituent assimilation.

In general, there is more variability in the data assimilation set-up of chemical model systems than in that for NWP systems. This is also reflected in the number of applications of the former. Currently, chemical model assimilation systems are used to: (i) derive information on unobserved species; (ii) provide analyses of tropospheric pollution; (iii) support the evaluation of satellite instruments; (iv) monitor stratospheric composition; and (v) forecast stratospheric ozone. To attain these goals, several methods are used: successive correction; optimal interpolation (OI), the Kalman filter (KF) and variants thereof; 3-D and 4-D variational methods (3D- and 4D-Var); and 3D-PSAS (physical space statistical analysis scheme). By contrast, most current NWP systems are based on variational methods (3D- and 4D-Var), although there are currently efforts underway at a number of institutions to evaluate the performance of the Ensemble Kalman filter, EnKF, for NWP (Lorenc 2003; Houtekamer et al. 2005). These studies concluded that EnKF is not currently competitive for NWP compared to 4D-Var, although continuing developments in EnKF may allow the method to match the performance of 4D-Var. Chapters in Part I, *Theory*, provide details of these data assimilation methods.

In the following part of this section, we review the different methods and systems used in constituent data assimilation with chemical models. We will also point out the major differences between these systems and the systems based on NWP models. For example, CTM-based systems tend to not consider radiance assimilation, which is generally the case in operational NWP systems (This is not due to a fundamental limitation of CTMs, which can theoretically be used with complicated observation operators – see, e.g., Müller et al. 2004.). For CTM-based systems, the observations are previously inverted to provide profiles or total column. In the case of profiles, the observation operator is reduced to the spatial interpolation of the model values at the observation location. In the case of columns, the model values are integrated over the model layers before performing the spatial interpolation. A second important point concerns the case where CTMs use a full photochemical scheme. In this case, the number of constituent control variables is much greater than in an NWP system. To give an example, a modern stratospheric CTM includes $\sim$50 chemical species while the current operational ECMWF NWP system includes only two constituents (humidity and ozone).

Three methods are commonly used in constituent data assimilation with chemical models: 4D-Var, approximations to the Kalman filter (generally involving parametrizations of the error covariances), and PSAS, which can be viewed as an approach to solve the Kalman filter, or as the dual of 3D- or 4D-Var, depending on whether the time dimension is included (3D- PSAS, the dual of 3D-Var, is, to our knowledge, the only form of PSAS to have been used so far on constituent assimilation). Each of these methods has advantages and disadvantages. The feasibility of 4D-Var has been demonstrated in NWP systems. Its main advantage is that it considers observations over a time window that is generally much longer than the

model time step: typically 24 h for chemical models, while the CTM time step is of the order of 30 min or less. For non-linear systems (as is generally the case for the atmosphere), this feature of 4D-Var, together with the non-diagonal nature of the adjoint operator (see chapter *Variational Assimilation*, Talagrand) which transfers information from observed regions to unobserved regions, reduces the weight of the background error covariance matrix in the final 4D-Var analysis compared to the KF analysis (for linear systems, the general equivalence between 4D-Var and the KF implies that the same weight is given to all data in both systems). In the case of constituent assimilation where a full photochemistry scheme is considered, the properties of the adjoint operator allow unobserved species to be constrained by observed species. This constraint can be expected when observed and unobserved species chemically interact with a time-scale of the order of the assimilation window or less. A special property of the 4D-Var analysis is that in the middle of the assimilation window it uses all of the observations simultaneously, not just those before the analysis. Because of this, 4D-Var is said to be a smoothing algorithm.

In contrast with the above advantages of 4D-Var, three weaknesses must be mentioned. First, its numerical cost is very high compared to approximate versions of the KF, and to PSAS, so that, in general, its implementation requires a supercomputer. The cost of 4D-PSAS (the dual of 4D-Var), like the cost of 4D-Var, is determined by the cost of the repeated integrations of the assimilating model and its adjoint (see, e.g., Courtier 1997; Louvel 2001); thus, its cost (if implemented for stratospheric constituent data assimilation) would not be significantly lower compared to that of 4D-Var. Second, its formalism cannot determine the analysis error directly; rather it has to be computed from the inverse of the Hessian matrix (again, this procedure is prohibitive in both CPU and memory). Finally, in contrast with NWP 4D-Var systems, past assimilation experiments using CTMs have not been based on the incremental method (Bouttier and Courtier 1999) and thus cannot take advantage of its benefits, e.g., solving the analysis at a reduced resolution, thereby reducing the computational cost.

The first assimilation study of constituent observations based on 4D-Var was presented by Fisher and Lary (1995). They used a trajectory box model with a reduced stratospheric chemistry scheme involving $O_3$, O, NO, $NO_2$ and $N_2O_5$. They assimilated $O_3$ and $NO_2$ data from the MLS and CLAES instruments on board NASA's Upper Atmosphere Research Satellite (UARS). They also performed an assimilation experiment using synthetic, i.e., simulated, data that showed ozone observations were able to constrain the other species. This study also introduced the concept of the influence function which, with the help of the adjoint model, measures the influence of a species at time $t>t_0$ on other species at the initial time, $t_0$.

Elbern et al. (1997) built a 4D-Var system using a trajectory box model to study tropospheric pollution and test synthetic observations. This study, motivated by a need to improve air quality forecasts, showed that morning ozone assimilation provides good initial conditions to forecast ozone in the afternoon and gives sufficient information to constrain unobserved species (e.g. $NO_2$). This work was extended by considering a three-dimensional regional CTM and the concept of twin experiments (Elbern and Schmidt 1999). With this set-up, they found that analyses of unobserved

$NO_2$ depended on the quality of the initial guess of the $NO_2$ field above the surface. Finally, Elbern and Schmidt (2001) studied the assimilation of real observations during a summer episode with enhanced ozone levels. While the short-term forecasts (from 6 to 12 h) benefit from the optimized initial conditions, this paper also indicates what is needed to improve the longer forecasts. The formalism developed in this latest effort includes a background covariance matrix that takes into account the anisotropy and inhomogeneity of the constituent fields and a better knowledge of the emission rates. With this set-up, it is in principle possible to estimate in the same assimilation cycle both the initial conditions and the model parameters, in other words, perform 4D-Var and inverse modelling at the same time. More details can be found in chapter *Inverse Modelling and Combined State-Source Estimation for Chemical Weather* (Elbern et al.).

Errera and Fonteyn (2001) built a 4D-Var assimilation system for stratospheric chemical observations. This system is based on a three-dimensional CTM with a detailed chemical scheme including 41 species and 144 reactions. Observations are taken from the CRISTA instrument. These include long-lived species ($CH_4$, $N_2O$ and CFC-11) and species with relatively shorter lifetimes ($O_3$, $HNO_3$, $ClONO_2$ and $N_2O_5$) in comparison to the time-scale of the assimilation window (24 h). Comparison with independent observations shows good agreement for observed species (e.g. 7% for ozone against HALOE; less than 15% for $HNO_3$ against ATMOS), and for $NO_x$ ($=NO + NO_2$) and HCl, two constituents that are not observed by CRISTA (in both cases less than 25% against HALOE). It was also shown that the HCl field is influenced by the assimilation of $ClONO_2$ observations.

Because of the strong temperature-dependence of the chemistry of short-lived species such as $NO_2$ and $NO_3$, their variability could provide information on temperature. One possible application is the use of temperature as a control variable in a chemical data assimilation system. Along these lines, the variational system built by Marchand et al. (2003, 2004) has been used to extract temperature information from GOMOS $NO_3$ observations (Lahoz et al. 2007b; Marchand et al. 2007).

The two other methods commonly used for constituent data assimilation are approximate versions of the KF, and PSAS. The KF method is formulated so that the analyses uncertainties are determined directly and can be propagated to the next assimilation time step. The PSAS set-up at the GMAO includes a method to compute an approximation of the forecast error covariance matrix.

Approximate versions of the KF, and PSAS, are based on the hypothesis of model linearity. Thus, the time window over which observations can be considered should be chosen carefully to ensure that the linearity hypothesis is satisfied. Khattatov et al. (1999) provided evidence that for a stratospheric photochemical box model, the linear approximation essential to applicability of the Extended Kalman filter (EKF) and 4D-Var is valid up to $\sim$10 days. This behaviour is explained as the combination of two factors: (i) concentrations of many modelled short-lived constituents are largely determined by concentrations of a few relatively long-lived constituents such as ozone, and parameters such as total active chlorine or nitrogen; and (ii) within the data assimilation set-up, linear approximations are generated at every solver time step and the matrices corresponding to such linear transformations are

multiplied to obtain a matrix approximating the evolution of the system over a 10-day period. Due to the nature of the stiff solvers, these time steps vary by orders of magnitude and get very small when the changes in concentration for some species are most rapid.

Lyster et al. (1997) developed a Kalman filter system for a 2-D advection model on an isentropic surface. Although particular effort was made to optimize the CPU time, such a system was not found to be practical due to the large computer resources required. Ménard et al. (2000), using the same model as Lyster et al. (1997) for the assimilation of $CH_4$ data, found that the standard KF formalism propagated the analysis covariance matrix inaccurately, with rapid loss of variance and an increase in the error correlations. To remedy this shortcoming, they formulated an alternative formalism to the KF system. This alternative formalism, described in companion papers by Ménard et al. (2000) and Ménard and Chang (2000), estimates model parameters using a robust method based on $\chi^2$ diagnostics which compares the observation minus forecast (OmF) residuals with those calculated by the Kalman filter (see also Sect. 4). The method is used to estimate three covariance parameters (representativeness error, model error, and initial error – see chapters in Part I, *Theory*). Because correlation length-scale parameters are found to be insensitive to the $\chi^2$ diagnostics, they are estimated using a maximum-likelihood method. The $\chi^2$ diagnostics have been used in other studies to estimate data assimilation system parameters; statistics from the OmF time series are also used to estimate these parameters.

Khattatov et al. (2000) used the $\chi^2$ diagnostics with a three-dimensional CTM that assimilated ozone data. The multidimensional nature of the problem meant that some simplification was required to comply with limitations in computer resources, both in terms of CPU and memory. Khattatov et al. (2000) also showed that the value of $\chi^2$ primarily depends on the value of the error growth and not on the correlation distance. The same authors also found that the root-mean-square of the OmF differences is mainly sensitive to the correlation length in the case where the spatial density of observations is high.

The $\chi^2$ diagnostics methodology has been applied successfully in constituent data assimilation (e.g. Chipperfield et al. 2002; Fierli et al. 2002; Lary et al. 2003 and, with some modifications, by El Amraoui et al. 2004 and Baier et al. 2005). Chipperfield et al. (2002) also introduced a method to constrain unobserved long-lived species (e.g. $N_2O$), in which an observed long-lived species (e.g. $CH_4$) is used to preserve a compact tracer-tracer relationship between both constituents. Finally, Eskes et al. (2003) developed a KF approach to produce near real time ozone analyses and 5-day forecasts. To comply with limited computer resources and the constraints of an operational service, Eskes et al. (2003) introduced several approximations in the KF method. For example, they used observation minus forecast (OmF) statistics to estimate the horizontal error correlations, the observation errors and the forecast errors.

As can be seen from the above examples, approximate versions of Kalman filter methods are very popular for constituent assimilation. This popularity is due to their low demand for computer resources in comparison to 4D-Var. An alternative

to approximate versions of the KF is the PSAS method used at the GMAO. It has the advantage that it solves the analysis in the observation space, which, for constituent assimilation, is typically much smaller in size than the model space. It thus reduces the computer resources needed. This approach is used in the Goddard Earth Observation System (GEOS) ozone data assimilation system described by Štajner et al. (2001). This system, based on a 3-D CTM with parametrized ozone chemistry, also uses the $\chi^2$ diagnostics to estimate the system parameters. The system became operational in 1999, providing stratospheric ozone analyses using SBUV/2 and TOMS (Štajner et al. 2001). Other combinations of ozone datasets have been assimilated in experimental versions of the GMAO CTM-based system: SBUV/2 and POAM-III (Štajner and Wargan 2004), SBUV/2 and MIPAS (Wargan et al. 2005).

Finally, as well as considering the performance of the NWP-based and chemical model approaches, one also needs to address the relative costs. While cost differences depend on the complexity of different model components, one can still highlight some key factors.

First, it is significantly cheaper to use a transport model than a coupled chemistry/dynamics model, if dynamical fields are available already. In a test with the Met Office Unified Model, the dynamics took ∼25% of the total model time, while advection of three tracers took 6% (A. Malcolm, personal communication). The advection of a single tracer is relatively simple and cheap compared with the sophistication required by the dynamics of the Met Office model. Similarly, the cost of the univariate assimilation of a single constituent will be simpler and cheaper than the proportionate cost of a dynamical variable that is treated multivariately. Furthermore, the smaller data volume of constituent observations makes constituent data assimilation relatively cheaper than data assimilation of dynamical variables (e.g. temperature, winds, humidity).

On the other hand, costs of the constituent data assimilation include the cost of the required chemistry model. While this could be simple (or even non-existent for constituents such as stratospheric methane), a complex chemical model is likely to be a major component of a sophisticated chemical data assimilation system. While we have outlined a range of cost considerations, it is worth stressing that the costs are highly dependent on the type of data assimilation method, transport model, and chemistry employed.

## 4 Evaluation of Models, Observations and Analyses

Both NWP-based and chemical model data assimilation approaches (see Sects. 2 and 3) are used to evaluate models and observations, in particular concerning ozone (e.g. Štajner et al. 2004; Geer et al. 2006a, b, 2007; Coy et al. 2007). Data assimilation not only corrects weaknesses in models, but also identifies model deficiencies such as biases (e.g. between model and observations; between different observations). In this section we provide further details – see also chapter *Evaluation of Assimilation Algorithms* (Talagrand).

A crucial element of data assimilation is the evaluation of the quality of the observations, the model and the analyses, and the test of several assumptions built into data assimilation algorithms, e.g., Gaussian errors; unbiased observations and models. Several diagnostics have been developed to do this (Talagrand 2003). Broadly speaking, these consist of: self-consistency tests, and independent tests. We first discuss self-consistency and independent tests in general. We then provide illustrative examples of how constituent data assimilation can be used to evaluate satellite instruments.

*Self-consistency tests.* Self-consistency tests provide useful information for evaluating the quality of the data assimilation ingredients and the assumptions built into assimilation algorithms. Histograms of OmA (observation minus analysis) and OmF (observation minus forecast) differences are computed for a range of spatial and temporal scales to test whether the observations, forecast and analysis fields, and their errors, are consistent with each other. For example, the OmA histogram should be more peaked than that for OmF, as the analyses should be closer to the assimilated observations than the forecast. Furthermore, the OmF histogram should be Gaussian, if both the observation and forecast are assumed to have Gaussian errors. Time averages of the standard deviation of OmA can also be used to test whether the assimilation system is consistent with the concept of the Best Linear Unbiased Estimate, *BLUE* (Talagrand 2003). Other tests check whether there are biases between observation and forecast, or between observation and analysis. Application of these tests is discussed in Errera and Fonteyn (2001), Štajner et al. (2001), Struthers et al. (2002) and Segers et al. (2005). See Fig. 2 for an example. Tests for Gaussian errors can also include tests of skewness and kurtosis (Geer et al. 2006b).

Time series of OmA and OmF differences test whether the observation, forecast and analysis fields, and their errors, are consistent with each other. A well-behaved data assimilation system will have time series with mean OmA and OmF values that are close to zero and do not vary much over time. If this is not true, a bias between the model and the data (or a subset of the data) is present. Also, if the standard deviation about the mean of the OmA time series is larger than the observational error, this indicates that the system is not properly set up. For example, the observation and background error covariance matrices, **R** and **B**, respectively, could be poorly characterized. Desroziers et al. (2005) suggest a simple method to evaluate **R** and **B** separately; Chapnik et al. (2006) describe a way of quantifying errors and biases of both model and observations in the process of tuning a data assimilation scheme for internal consistency.

Time series of OmA and OmF differences can also be used to monitor the performance of satellite instruments; changes in their values can indicate a change in the instrument algorithm, or a degradation of the instrument. For example, Štajner et al. (2004) uses the OmF time series provided by the GEOS ozone data assimilation system to validate the NOAA-14 SBUV/2 retrieval algorithm. Furthermore, at the start of a data assimilation experiment, it can take some time for the system to spin-up; this spin-up time is shown by the time it takes for OmA or OmF differences to converge towards a constant value (Struthers et al. 2002).

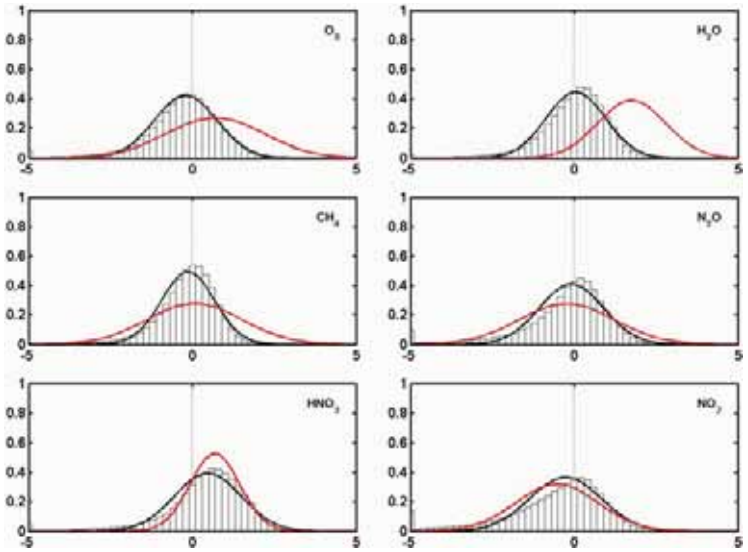



**Fig. 2** Evaluation of analyses using histograms of OmF differences (normalized by the observation error) averaged for the stratosphere, the globe and August 2003 for six stratospheric constituents: $O_3$ (*top left*), $H_2O$ (*top right*), $CH_4$ (*middle left*), $N_2O$ (*middle right*), $HNO_3$ (*bottom left*) and $NO_2$ (*bottom right*). The constituent observations are from ESA MIPAS off-line retrievals. The frequency of the histograms is normalized to lie between 0 and 1. The *black line* is a Gaussian fit to the histograms; the *red line* is a Gaussian fit from a model run without assimilation. The results support the assumption of Gaussian errors in the observations and the forecast, and show the analyses are closer to the observations than simulations from the model run without assimilation. The experiments were performed at BIRA-IASB. With permission from Lahoz et al. (2007a)

If the OmF differences have a Gaussian distribution, its inner product normalized by its covariance is a random variable that has a $\chi^2$ distribution with $p$ degrees of freedom, where $p$ is the number of observations. This result can be used to test whether the OmF differences are consistent with assumptions made in the assimilation algorithm, and to monitor the observations (Ménard et al. 2000; Ménard and Chang 2000; Štajner et al. 2004).

If the data (observation and background) errors are Gaussian, the minimum of the penalty function, $J_{\min}$, follows a $\chi^2$ distribution with $p$ degrees of freedom, and must be equal on average to $p/2$. This last result is also true if the errors are not Gaussian, but the assimilation scheme remains linear. Thus, in these cases, $J_{\min}/p$ should on average be 0.5 (Talagrand 2003). In practice, $J_{\min}/p$ is often significantly different from 0.5. This discrepancy can arise from an incorrect estimate of **B** or **R** (mainly the representativeness error in the case of **R**).

Comparison of observations with short-term forecasts is also used to evaluate the consistency of the observations and the model prior to assimilation of the observations. This quality control is an important element of data assimilation algorithms (see, e.g., Lorenc and Hammon 1988).

Several robust correlations between pairs of long-lived tracers have been observed in the atmosphere (Plumb and Ko 1992). A particular example is the correlation between $CH_4$ and $N_2O$ (Chipperfield et al. 2002). When two or more long-lived tracers are assimilated, the quality of the analyses can be assessed through the consistency of the tracer-tracer correlations.

*Independent tests.* These tests involve comparison of analyses with data that are independent from the analyses, i.e., data not assimilated to provide the analyses. Independent datasets used to evaluate ozone analyses include ozonesondes (Logan 1999) or satellite data which is not commonly assimilated (e.g. from the UARS HALOE instrument, Russell et al. 1993). Independent data can provide information on whether the analyses are realistic and can help attribute biases to observations, forecast and analysis; note that self-consistency tests cannot be used to perform this attribution. Estimating the bias in the analyses by comparison against independent data is only possible when the error characteristics of the latter are well known. Application of these tests is discussed in Khattatov et al. (2000), Struthers et al. (2002) and Segers et al. (2005). See Fig. 3 for an example.

When analyses are compared against independent data it is important to take account of the observation characteristics of each dataset. This can be accomplished by making use of averaging kernel information, which accounts for the information content, including the vertical resolution, of the observations (Migliorini et al. 2004). This is difficult in practice, as the averaging kernel information is not always readily supplied by the measuring instrument specifications.

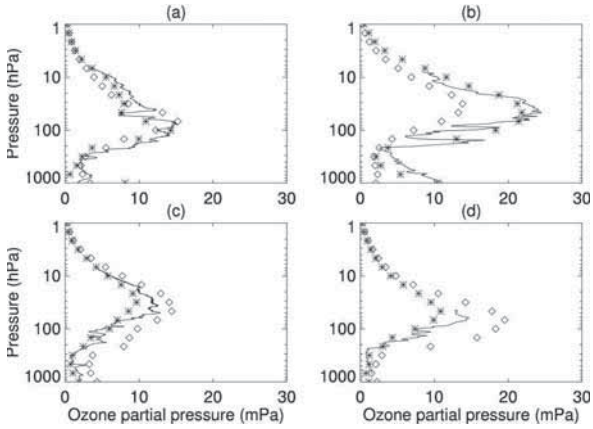



**Fig. 3** Evaluation of ozone analyses using independent data at four locations: (**a**) Ny Ålesund (78.9°N, 11.9°E) on 27th April 1997; (**b**) Payerne (46.8°N, 7.0°E) on 25th April 1997; (**c**) Lauder (45.05°S, 169.7°E) on 16th April 1997; and (**d**) South Pole (90°S) on 18th April 1997; all plots at 1200 UTC. The analyses (*stars*) are compared against ozonesonde data (*line*) that have not been used in the assimilation. The ozone data used to initialize the assimilation are shown as *diamonds.* The results show reasonable agreement between the analyses and the ozonesondes, and the lack of influence of the initial ozone conditions after the spin-up period. Units are mPa. With permission from Struthers et al. (2002)

In general, comparison against independent data is much more significant than comparison against the assimilated observations. Thus, independent data are the ultimate arbiter of the quality of analyses. In Sect. 5 we discuss the quality of humidity and ozone analyses from NWP- and CTM-based assimilation systems, based on the intercomparison of analyses between themselves and against independent data. We also mention briefly early efforts to carry out these intercomparisons for other stratospheric constituents.

*Illustrative examples.* The use of constituent data assimilation to evaluate instruments is numerous. Assimilation of long-term data set can be used to detect and characterize changes in the observation errors (e.g. Štajner et al. 2004 – see above).

In the next two examples, data assimilation has been used to evaluate two scientific instruments onboard Envisat: GOMOS and MIPAS. GOMOS is a stellar occultation instrument that measures, among other species, stratospheric nighttime profiles of $O_3$, $NO_2$ and, for the first time, $NO_3$. This last species has a very short life-time. During the daytime, its concentration is close to zero because it is photolysed in the presence of sunlight. During the night, its chemistry is very simple and strongly coupled to $O_3$ and $NO_2$. Marchand et al. (2004) have assimilated GOMOS $O_3$ and $NO_2$ in a photochemical model using a variational approach. Figure 4 shows the $NO_3$ analysis plotted against the corresponding GOMOS observations averaged over 7 isentropic levels: 735, 900, 990, 1,100, 1,210, 1,350 and 1,510 K. $NO_3$ observations that correspond to $NO_2$ observations with an higher error than 30% are not taking into account. Based on this comparison, Marchand et al. (2004) validate the self-consistency of GOMOS $O_3$, $NO_2$ and $NO_3$ measurements, and the nighttime $NO_3$ chemistry (see Fig. 4).

Within the validation framework for MIPAS, Vigouroux et al. (2007) have compared MIPAS $N_2O$ and $HNO_3$ with ground based FTIR measurements for 2003. They use a co-location criterion of 1,000 km around ground-based stations. In order to increase the number of co-locations, they also use MIPAS $N_2O$ and $HNO_3$ analyses produced by the Belgian Assimilation System for Chemical ObsErvation,

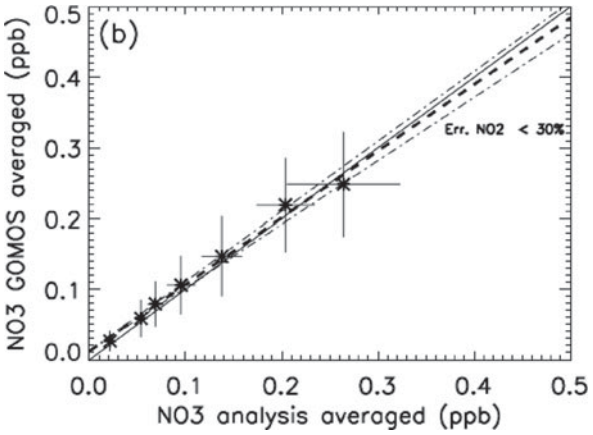



**Fig. 4** GOMOS $NO_3$ measurement against analysed $NO_3$ averaged over isentropic levels. The standard deviations of the isentropic means of GOMOS $NO_3$ and of mean analysed $NO_3$ are indicated by *vertical* and *horizontal lines*, respectively. With permission from Marchand et al. (2004)

BASCOE (previously the Belgian Assimilation System for Chemical Observations from Envisat). This paper also discusses under what conditions these analyses can be considered a good proxy for MIPAS observations. In the case of $N_2O$, the agreement between BASCOE analyses and the MIPAS and FTIR data is excellent. Comparison with FTIR shows a bias ranging from –5 to +1%, and standard deviations ranging from 2 to 7%. Compared to the MIPAS random errors, these values are not significant. BASCOE appears to have more difficulty in producing proxies for MIPAS $HNO_3$ profiles but the estimated standard deviations, less than 10% between BASCOE and FTIR, appear reasonable.

# 5 Applications

## *5.1 Tropospheric Pollution*

Efforts to apply assimilation systems to the more complex system of the troposphere (Monks 2003) draw heavily on the NWP and chemistry model heritage from the assimilation of stratospheric constituents. These efforts are being directed at tackling a number of technical problems specific to the troposphere, including: more chemical variables than in the stratosphere; extracting chemical information in the presence of clouds; the need for higher spatial resolution to capture mesoscale phenomena such as fronts and pollution plumes; and the need to reassess balance at higher spatial resolution.

In contrast to stratospheric constituent data assimilation and NWP data assimilation, the evolution of the tropospheric model state is not primarily controlled by the initial state. Instead, emissions are a strong controlling factor, and exert a direct influence over short time-scales (ranging from seconds to days). Furthermore, currently, emission rates are not sufficiently well known. Thus, emission rates must be considered as another parameter to be optimized in the data assimilation process. Tropospheric data assimilation must also take account of the differences in spatial scale between satellite data retrievals and point-like emissions. For more details see chapter *Inverse Modelling and Combined State-Source Estimation for Chemical Weather* (Elbern et al.)

During the last decade, tropospheric CTMs (global and regional) have become increasingly more accurate. The number of observations available has also increased and this has favoured the development of tropospheric assimilation systems for the study of tropospheric pollution.

Lamarque et al. (2002) derived global tropospheric ozone column (TOC) from the assimilation of ozone profiles from UARS MLS and ozone total column from TOMS. This study was based on the suboptimal KF method and used the MOZART CTM. Daily analyses of TOC show that this method is useful for the study of the transport of pollutants such as biomass burning plumes.

Using the same system, Lamarque and Gille (2003) assimilated carbon monoxide. In this case, the observations are taken directly from the troposphere, and are

provided by the MOPITT instrument. This study takes into account the bias between observations and the forecast field in the CTM. This is done by the implementation of a bias estimator in the suboptimal KF. Results show that the method significantly improves the assimilated field, with a reduction of the global mean OmF statistic.

Finally, the group at the University of Köln (e.g. Elbern et al. 1997; Elbern and Schmidt 1999, 2001) has developed an assimilation system for air quality forecasts. Using a regional CTM and 4D-Var, Elbern and Schmidt (2001) showed that assimilation of surface ozone can improve the forecast (note that Lahoz et al. 2007b; Elbern et al. 2007 discuss more recent developments). However, these improvements are limited by uncertainties in model parameters such as boundary values, deposition velocities and surface emission rates. In fact, in tropospheric assimilation, these parameters affect the analyses on the time-scale of the assimilation window. All things being equal, this makes tropospheric constituent assimilation systems harder to implement than stratospheric constituent assimilation systems, where such parameters have little effect on the analyses at the time-scale of the assimilation window.

## *5.2 Analyses of Constituents*

Objective evaluation of analyses can be obtained by the intercomparison of analyses produced using different data assimilation systems. If the systems assimilate a common observational dataset, differences between the analyses can be attributed to differences in the models and/or the assimilation system. Furthermore, by confronting these analyses against others and against independent data (i.e., not assimilated) it is possible to both gain an understanding of their strengths and weaknesses, and to make new developments. Finally, these intercomparisons provide more information (and faster) than if each participant assessed their own system independently.

In this section we use the analyses intercomparison approach to assess the accuracies of humidity analyses in the stratosphere-mesosphere (Lahoz et al. 2007a, b; Thornton et al. 2009), and the accuracy of ozone analyses in the stratosphere-mesosphere (Geer et al. 2006a). Intercomparison of analyses of stratospheric constituents other than humidity and ozone are currently underway. For example, Errera et al. (2007, 2008) discusses the performance of $NO_2$ analyses using the BASCOE chemical model and observations from MIPAS and GOMOS.

*Humidity analyses.* Accuracies of sample humidity analyses in the stratosphere-mesosphere are discussed by Lahoz et al. (2007a, b) and, more recently, Thornton et al. (2009). We summarize the results from Thornton et al. below.

For 1 month period (29 August–29 September 2003), MIPAS water vapour profiles were assimilated into four models: ECMWF and Met Office, GCM-based; BASCOE (from BIRA-IASB) and MIMOSA (from Service d'Aeronomie), CTM-based. The resultant analyses were compared against the original MIPAS observations and with independent water vapour data: HALOE, SAGE-II and POAM-III.

Met Office results were not considered in the analyses comparison due to their poor performance. Because many of the problems associated with the Met Office humidity assimilation may be linked to the specification of the background error covariances, tests with different specifications of the background error covariances were implemented. This is discussed in more detail below.

Comparison of the ECMWF, BASCOE and MIMOSA analyses with the independent data highlights areas where the analyses are either realistic or require further improvement. In general, these three analyses compare favourably with the UARS climatology (http://code916.gsfc.nasa.gov/Public/Analysis/UARS/urap/home.html). The main features of the stratospheric water vapour field are captured (Fig. 5), for example, the tropical water vapour minimum, the Southern Hemisphere polar vortex water vapour minimum and the vertical distribution of water vapour associated with the Brewer-Dobson circulation (see also SPARC 2000). In the mesosphere, the analyses are wetter than the UARS climatology and reflect the wet bias of the MIPAS observations relative to other satellite data in this region. Thornton et al. provides detailed quantitative information on the performance of the ECMWF, BASCOE and MIMOSA water vapour analyses.

The region of the stratosphere-lower mesosphere that the assimilation schemes find hardest to simulate is the Southern Hemisphere polar vortex between 100 and 20 hPa. Most of the analyses have difficulty correctly capturing the moisture minimum within the vortex core, and the strong horizontal and vertical humidity gradients at the vortex boundary, resulting in large dry biases in this region. The water vapour minimum in the tropical lower stratosphere associated with temperature minima at the tropopause, and the horizontal transport of this dry air to higher latitudes, is also difficult for the assimilation systems to represent, with most systems showing a dry bias. BASCOE was found to have a particularly large dry bias in the Southern Hemisphere polar vortex; this is most likely explained by an overactive PSC parametrization scheme. The MIPAS observations only had a limited effect, as many were rejected due to their large deviation from the erroneously dry background field.

In the upper stratosphere, the ECMWF, BASCOE and MIMOSA analyses have assimilated the MIPAS observations well, with small biases and standard deviations. Larger biases exist when compared to independent observations, especially HALOE data, which can in part be explained by the relatively large MIPAS bias relative to HALOE in this region. The strong latitudinal water vapour gradient at 2 hPa associated with the Brewer-Dobson circulation does, however, produce a peak in the biases in the upper stratosphere. The MIPAS observations have also been well assimilated in the lower mesosphere and small biases are found against independent data below 0.5 hPa for both ECMWF and BASCOE.

Although the MIMOSA analyses often have smaller biases with respect to the MIPAS observations, the analyses are unrealistically noisy. This most likely reflects the lack of data quality control and the background error covariance dependence on potential vorticity, PV (El Amraoui et al. 2004). The water vapour analyses of ECMWF and BASCOE are smoother and the relatively low horizontal resolution of the BASCOE grid amplifies this homogeneity.

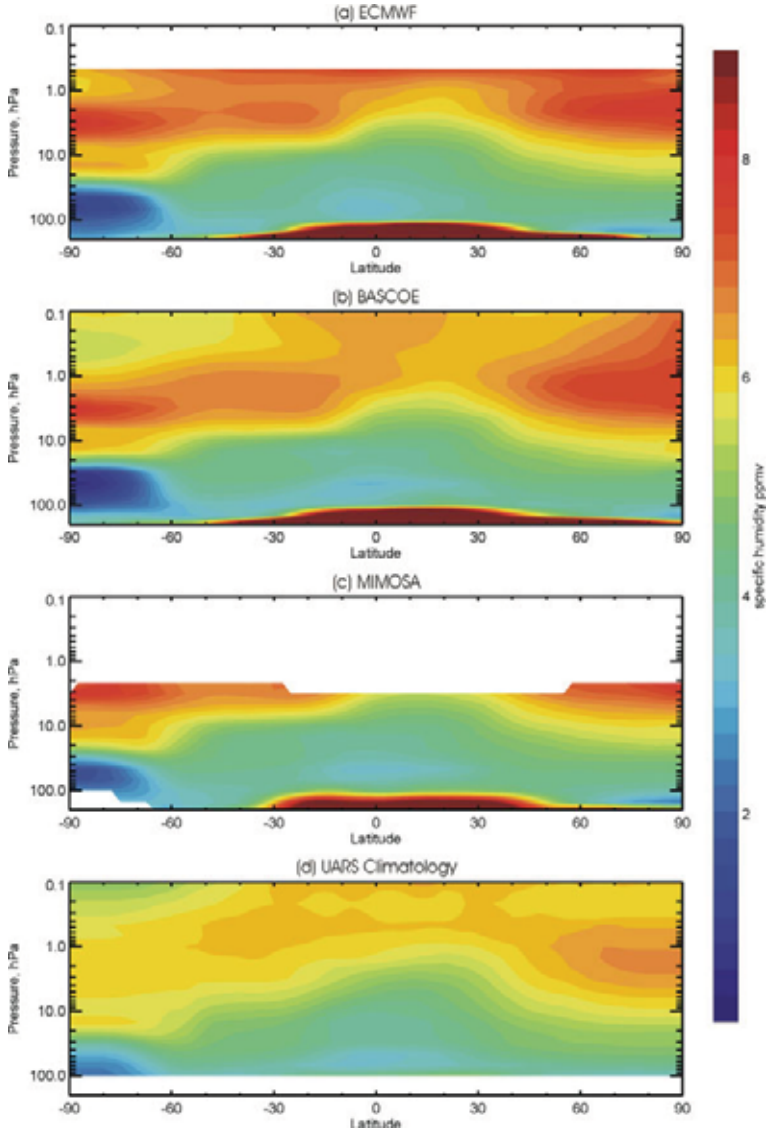



**Fig. 5** Monthly zonal mean specific humidity analyses for September 2003 for (**a**) ECMWF, (**b**) BASCOE, and (**c**) MIMOSA; (**d**) UARS climatology. MIPAS water vapour profiles have been assimilated in the ECMWF, BASCOE and MIMOSA analyses. *Blue* denotes relatively low specific humidity values; *red* denotes relatively high specific humidity values. Units: parts per million by volume, ppmv. Based on Thornton et al. (2009)

As discussed above, the Met Office stratospheric water vapour analyses studied in Thornton et al. were poor, with large dry biases at most latitudes and altitudes relative to all observation types considered. Investigations have shown that the poor

assimilation of MIPAS profiles is related to an unrealistic humidity background error covariance matrix, rather than to any dynamical feature of the model. The humidity background error covariances were found to have excessively deep vertical error correlations and error variances that were larger than the background humidity values. Modification of the error covariance matrix failed to sufficiently improve the assimilation capability.

The water vapour analyses comparison in Thornton et al. has highlighted the following: (1) the role of the background error covariance matrix is crucial in producing a realistic mid atmosphere water vapour analysis; (2) quality control of the observations assimilated can avoid poor observations degrading the analyses; and (3) the assimilation schemes compared (ECMWF, BASCOE, MIMOSA) have succeeded in producing reasonable mid atmosphere water vapour analyses, although the schemes have difficulty in reproducing regions with strong humidity gradients.

The Met Office has investigated the impact of varying the control variable in the assimilation of MIPAS humidity data. The objective is to develop a humidity control variable that has the desirable properties that it is usable in both the troposphere and the stratosphere; has approximately Gaussian background errors; that temperature and humidity increments are decoupled; and that allows realistic vertical error correlations. To achieve this, the Met Office have combined the ideas of Dee and da Silva (2003) and Hólm et al. (2002), and defined a normalized relative humidity variable.

Lahoz et al. (2007b) describe three different experiments by the Met Office where the humidity control variable is either relative humidity (*RH*), normalized *RH* or normalized specific humidity. All three experiments show fairly reasonable specific humidity profiles for levels below 5 hPa. However, at higher levels the fit to the MIPAS observations is less good, with the analyses being consistently too dry. The experiment with the normalized specific humidity control variable has a more reasonable lower mesospheric specific humidity, but is still too dry when compared to the MIPAS observations.

Recently, Eckermann et al. (2008), using a prototype system based on the operational US Navy models, have assimilated water vapour (and ozone) to high altitudes (water vapour to mesosphere, ozone to stratopause). Interestingly, they avoid the water vapour noise problems mentioned in Thornton et al. (2009).

*Ozone analyses.* The accuracy of ozone analyses from NWP- and CTM-based systems is discussed in detail in the intercomparison by Geer et al. (2006a). It is shown that the best performing analyses are capable of producing very good agreement with ozonesonde, HALOE and MIPAS ozone data. From the lower stratosphere to the lower mesosphere (100–0.5 hPa), these analyses show biases less than $\pm 10\%$ with respect to HALOE ozone data and ozonesondes. Standard deviations can be less than 10% above 50 hPa and less than 20% in the lower stratosphere (100–50 hPa). This shows that current assimilation techniques are capable of producing ozone analyses that have good agreement with independent data (see Fig. 6).

The enhanced skill of the best performing analyses can usually be attributed to better modelling of ozone chemistry or transport processes. The worse performing systems could often be easily improved by following similar modelling techniques.

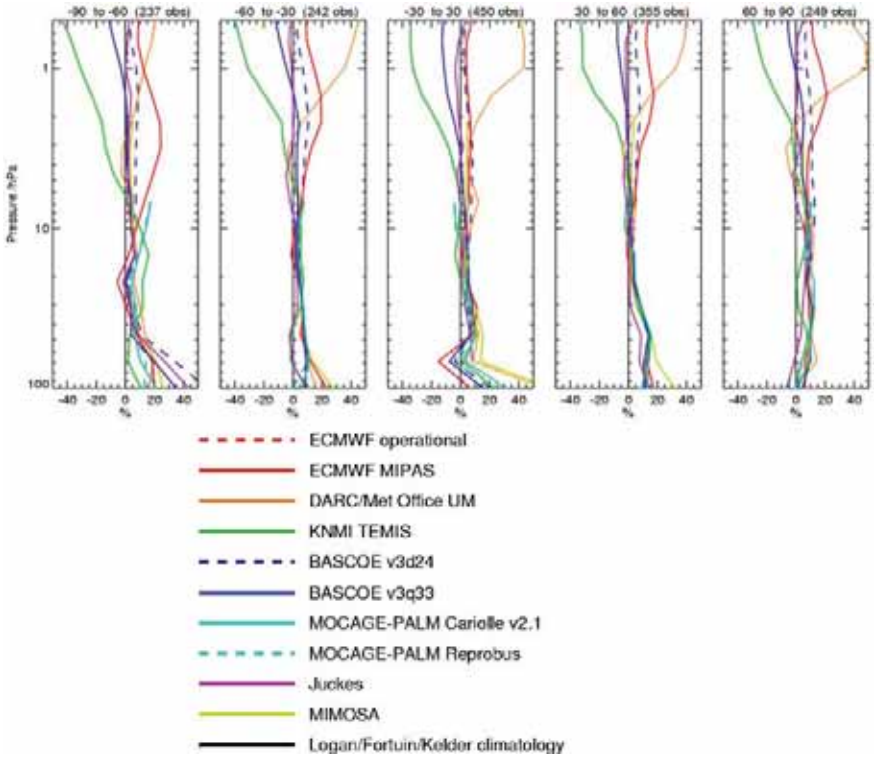



**Fig. 6** *Top*: Mean of analysis minus HALOE differences (in percent), normalized by climatology, for the period 18 August–30 November 2003. *Bottom*: Colour key for *top part* of figure. The numbers in *brackets* indicate the HALOE/analysis coincidences within each latitude bin. Based on Geer et al. (2006a)

For example, this can apply to regions where there are limitations with the ozone data assimilated, where as shown by Geer et al. (2006a), CTMs and GCMs with chemistry generally do better. The intercomparison finds few differences that can be attributed to the assimilation technique or the model used (GCM or CTM). It would require focused experiments, rather than an intercomparison, to reveal such differences. Overall, the study by Geer et al. (2006a) shows that the first priority for ozone data assimilation systems is to improve the modelling of ozone chemistry and transport.

The work of Geer et al. (2006a, b) on the quality of ozone analyses has highlighted the importance of observational and model bias in data assimilation. Besides providing information on observational bias, data assimilation can provide information on, and be affected by, model bias. For example, Geer et al. (2006b), using the Met Office Unified Model, found that vertical transport of ozone in the tropical pipe, and transport in the Brewer-Dobson circulation, is much too fast as a result of known problems in the tracer transport scheme. This was manifested in that ozone forecasts above the ozone peak (10 hPa) tended to be biased high against the MIPAS

values (negative OmF values), and ozone forecasts around the ozone peak tended to be biased low against the MIPAS values (positive OmF values).

The Brewer-Dobson circulation is also degraded by problems with the assimilation of dynamical variables (Douglass et al. 2003; Schoeberl et al. 2003; Tan et al. 2004). This reflects that it is very hard for data assimilation to handle slow processes, on time-scales much longer than typical assimilation cycles. Problems with stratospheric tracer transport are seen in many data assimilation systems (Oikonomou and O'Neill 2006), and this remains a major focus of investigation.

Work by Monge-Sanz et al. (2007) shows that ECMWF ERA-interim reanalyses (ECMWF 2007) can be used to provide realistic stratospheric transport over multi-annual time-scales with an off-line CTM; in particular, the CTM's age of air agrees reasonably well with observations. The improvement, in comparison with forcing the CTM with ERA-40 reanalyses or troposphere-stratosphere analyses from the Met Office, is attributed mainly to the use of 4D-Var and an improved balance operator, together leading to more balanced flow and reduced mixing in the subtropics. In addition, an improved implementation of the bias correction of satellite radiances is thought to have helped reduce the analysed strength of the Brewer-Dobson circulation.

Finally, several papers (Levelt et al. 1998; Chipperfield et al. 2002; Juckes 2006, to name a few) show analysed constituent datasets that are closer to independent data than the assimilated observations or the simulated fields, thereby providing evidence that the data assimilation method can add value to constituent information, either from observations or from a model. Jackson (2007) shows that assimilation of EOS MLS ozone data reduces mean analyses errors in the lower stratosphere. Compared to control simulations where no ozone data are assimilated, mean errors (evaluated against HALOE ozone data) dropped by 5–25% in the Southern Hemisphere extra-tropics, and by ∼10% in the Northern Hemisphere extra-tropics; mean errors (evaluated against ozonesondes) dropped by ∼50% in the tropical UTLS.

Along these lines, Struthers et al. (2002) demonstrate that the combined assimilation of UARS MLS ozone profiles and GOME total column ozone gives analysed constituent datasets that are closer to independent data than either of the analyses derived from the assimilation of UARS MLS ozone profiles, or of GOME total column ozone. Thus, in this case, combined assimilation has added value to the single assimilation of these ozone datasets. Note, however, that this is not always the case, as there could be inconsistencies in the assimilation system, for instance in the treatment of biases (Rood 2005). Thus, there is scope for improving the use of observations in constituent data assimilation.

## *5.3 Stratospheric Ozone Monitoring*

Monitoring the stratosphere is done routinely by satellite instruments in order to track the evolution of the stratospheric composition, mainly ozone and the gases that destroy it (WMO 2006). Currently, products from different data assimilation groups are used to help this monitoring effort and assess protocols.

ECMWF use their NWP operational system to monitor satellite ozone data by passive data assimilation, i.e., the ozone data are passed through the assimilation system and evaluated, but are not allowed to affect the analyses. For example, Dethof (2004) describes the monitoring of ozone profiles from the MIPAS and GOMOS instruments, and total column ozone from the SCIAMACHY instrument. As of February 2009, ECMWF monitored the following ozone products: partial columns from the three SBUV/2s; total column from SCIAMACHY, OMI, SEVIRI and GOME-2; and ozone profiles from GOMOS (R. Dragani, personal communication). If the monitored data prove satisfactory, they are moved to active assimilation into the ECMWF operational system, and thus are allowed to affect the meteorological analyses (as well as the ozone analyses).

NCEP have set up an operational ozone monitoring and forecasting system within the NCEP Global Forecasting System (GFS). They use the CHEM2D-OPP chemistry module (McCormack et al. 2006). As of September 2007, the system assimilated several ozone products, including SBUV/2 partial ozone columns from NOAA-16 and NOAA-17, and total column ozone from OMI (Long et al. 2007). Operational assimilation of NOAA-18 SBUV/2, and total column ozone from OMI and GOME-2 is expected to begin in late 2009. Parallel tests assimilating NOAA-19 SBUV/2, and ozone profiles from OMI and EOS Aura MLS will continue through 2010 (C. Long, personal communication).

Ozone is not assimilated operationally at the Met Office, but recent work has been carried out using ozone data assimilation to investigate the impact of different ozone representations on tropospheric weather forecasts (Mathison et al. 2007), and to estimate stratospheric ozone loss (Jackson and Orsolini 2008). Currently, the work on 3D-Var is being extended to 4D-Var, and ozone profile data from GOME-2 are being assimilated. Finally, during 2010 the Met Office hope to develop a new ozone control variable, using a concept similar to the humidity variable developed by Hólm et al. (2002) (D. Jackson, personal communication).

Since 2000, KNMI produce near real time total ozone assimilation (Eskes et al. 2003). This system is constrained by total ozone observations provided by a variety of satellite instruments (TOMS, SBUV, GOME, SCIAMACHY, OMI, GOME-2 depending on the time period) and has delivered global maps of total ozone since August 1995 (http://www.temis.nl). This database is being used to evaluate the change of total ozone since the 1960s (WMO 2006).

Stratospheric constituent assimilation using a full chemistry model and 4D-Var is underway at DLR and BIRA-IASB. In the framework of the ESA-funded PROMOTE project, these two institutions will provide reanalyses of stratospheric ozone from 1992 (i.e., soon after the launch of the UARS satellite) to the present, using ozone data from different sensors (see the "Stratospheric Ozone Profile Record" project, http://www.gse-promote.org for more details). In addition to ozone, they expect to provide analyses of several parameters related to ozone chemistry: $ClO_x$, $NO_x$, PSCs, ozone depletion rate and $Cl_y$ (total available chlorine). These reanalyses and analyses will be used by international organizations such as SPARC (Stratospheric Processes And their Role in Climate) in the framework of the Chemistry-Climate Model Validation (CCMVal) and WMO-GAW (World

Meteorological Organization – Global Atmospheric Watch) projects to assist in the evaluation of compliance with the Montreal protocol.

## *5.4 Ozone Forecasting*

Ozone forecasts are useful for predicting high UV-flux events. They can be used to warn populations near the Antarctic when the ozone hole moves above these areas. They can also be used to plan observation campaigns. Ozone forecasts are operational at ECMWF since 2002 (Dethof 2003), and became operational at GMAO and KNMI in, respectively, 1999 (Štajner et al. 2001) and 2000 (Eskes et al. 2003). The ECMWF system is GCM-based; these KNMI and GMAO systems are CTM-based. There is currently a GCM-based system at GMAO used for near real time assimilation of ozone (Rienecker et al. 2008).

The ECMWF products have been based on different ozone datasets, depending on their availability; see Sects. 2.3 and 5.3 for the status of operational ozone assimilation at ECMWF as of February 2009. The KNMI products are based on total column ozone measurements from multiple instruments: GOME on ERS-2, SCIAMACHY on Envisat, GOME-2 on MetOp, OMI on EOS Aura, the NASA TOMS instruments, and the SBUV instruments on the NOAA satellites. GMAO products are based on TOMS total column ozone and SBUV/2 partial column ozone measurements. The ECMWF system is based on its NWP system, and includes parametrized ozone chemistry. The KNMI CTM-based system, and the GMAO CTM- and GCM-based systems also use parametrized ozone chemistry. The CTM-based systems are forced by off-line winds and temperature from the ECMWF (KNMI) and GEOS (GMAO) models. The ozone forecasts from these CTM-based systems are produced using wind and temperature forecasts from the ECMWF (KNMI) and GEOS (GMAO) models.

BIRA-IASB also set up an ozone forecasting service using the BASCOE system (http://bascoe.oma.be/archives). The system is based on a CTM with full chemistry and a scheme that explicitly calculates the microphysics of PSCs. The CTM is forced by off-line winds and temperature from the ECMWF model. The constraining observations are MIPAS near real time ozone profiles as well as five other chemical species ($NO_2$, $HNO_3$, $N_2O$, $CH_4$ and $H_2O$). In addition to ozone 10-day forecasts, this service also produced forecasts of $ClO_x$, $N_2O$, $HNO_3$ and $ClONO_2$ volume mixing ratio, and PSC surface area density. This service was operational for $1\frac{1}{2}$ years, and ended in March 2004 when delivery of MIPAS near real time profiles was interrupted due to problems with the MIPAS instrument. This difficulty with the MIPAS instrument highlights the weakness of using near real time products from research instruments for operational services.

Eskes et al. (2002) estimate that useful ozone forecasts can be obtained up to about 1 week for the extra-tropics with the KNMI system. In the tropics, the forecast skill is less good (useful forecasts out to $\sim$2 days) due, mainly, to the lack of tropospheric chemistry in the KNMI CTM. Two examples illustrate the skill of the KNMI

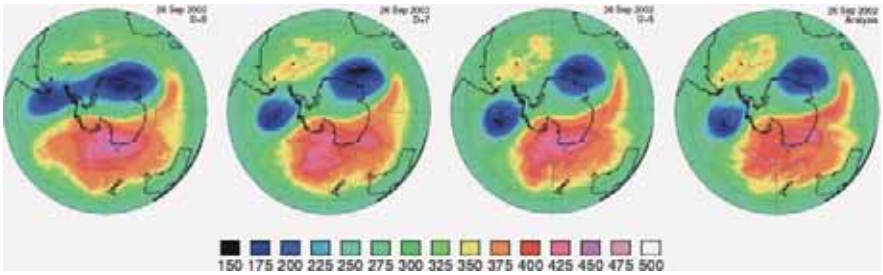



**Fig. 7** Ozone total column on 26 September 2002, provided by the KNMI operational ozone assimilation system. From *left* to *right*: 9-, 7-, 5-day forecasts, and the corresponding analysis. With permission from Eskes et al. (2005)

system. The first concerns low ozone events (also known as ozone mini-holes) that are observed during winter over the Atlantic and Northern Europe, and last for 1–2 days (Orsolini and Nikulin 2006). These events are due to dynamical transport of low ozone from the subtropics to the extra-tropics. For these events, 5-day ozone forecasts are found to be qualitatively good; 3-day forecasts are found to be quantitatively equivalent to the analyses, the latter being close to the observations (GOME total column ozone). The second example concerns the Antarctic polar vortex split of September 2002. During this unprecedented event, associated with a stratospheric warming (Eskes et al. 2005), the vortex split into two parts before decaying. As a result of this, the ozone hole also split into two parts. Figure 7 shows the ozone total column on 26 September over Antarctica calculated by the KNMI analysis and 5-, 7- and 9-day forecasts of the total ozone column. The analysis for this day shows the ozone hole split with two distinct regions of low total column ozone (values less than 200 Dobson Units, DU). For this event, forecasts out to 7 days perform well, and differences from the analyses are small. The 9-day forecast captures elements of the ozone hole split.

These two cases highlight the maturity of the KNMI ozone forecast service. However, the high accuracy of the forecasts would not have been possible without high quality dynamical fields, in this case from ECMWF. The success of the KNMI forecasts shows that the underlying dynamical processes were well captured by the ECMWF NWP system (Simmons et al. 2005).

## 6 Future Directions

Constituent data assimilation has developed enormously during the last 15 years to a position where incorporation of constituents in NWP (especially ozone) is routine. This effort has benefited from collaboration between operational and research institutions to identify shortcomings in the different assimilation approaches, for example within the EU-funded ASSET project (Lahoz et al. 2007b) and, in particular, the ASSET ozone intercomparison project (Geer et al. 2006a). The importance

of maintaining and developing these collaborations has been noted (McLaughlin et al. 2005).

Two approaches have been commonly used in constituent data assimilation: GCM-based NWP models and chemical models, either CTMs or photochemical box models. Recently, the NWP and CTM approaches have started to be combined in coupled NWP/CTM data assimilation, e.g., in collaboration between Environment Canada, other Canadian partners and BIRA-IASB, where the chemical scheme of the BASCOE CTM is coupled to the Canadian GEM-strato GCM; early results are promising (Ménard et al. 2007; see also de Grandpré et al. 2009). The CMAM data assimilation set-up at Met Service Canada (MSC) described by Polavarapu et al. (2005a, b) uses a GCM with full chemistry and can also be described as a coupled system.

Key drivers in constituent data assimilation for the future are likely to include the need to monitor the environment (e.g. stratospheric ozone; tropospheric pollution); the need to comply with international treaties such as the Montreal protocol; and the need to comply with environmental legislation concerning, e.g., air quality. This is illustrated by the PROMOTE project (http://www.gse-promote.org), one of the GMES service elements set up by ESA. PROMOTE is a user-oriented project, which aims to use the assimilation of constituent data to provide services on global ozone, greenhouse gases and air quality.

Another area of increasing importance will be the relationship between chemistry and climate. While this is naturally mainly the focus of coupled chemistry-climate models (see Eyring et al. 2006 and references therein), it does increase the importance of the compilation of assimilated constituent data for the study of recent climate variations and evaluation of climate simulations; climate/chemistry interactions will thus be one of the leading drivers for the development of coupled chemistry/dynamics assimilation systems. The inclusion of ozone in the recent ERA-40 reanalysis (Dethof and Hólm 2004) illustrates the importance of these considerations. The EC and ESA initiative on GMES illustrates the perceived importance on more general environmental monitoring. The ECMWF-led GEMS project (Hollingsworth 2005; Hollingsworth et al. 2008), part of GMES, illustrates the widening the scope of data assimilation to include not just atmospheric dynamics but a widening range of atmospheric constituents. The EU-funded MACC project, due to start in 2009, will build on EU and ESA investments in GEMS and PROMOTE.

In developing further constituent data assimilation, choices will have to be made concerning issues such as the type of model, the complexity of the chemistry component in the model and the assimilation set-up. These choices will depend on the application (see, e.g., Eskes 2006; Lahoz 2006). Challenges concerning issues such as bias, what datasets to assimilate, the usefulness of satellite observations of tropospheric constituents, the need for ancillary datasets (e.g. cloud and aerosol information), representation of the model physics and chemistry, the suitability of the NWP approach to constituent data assimilation and air quality forecasts, and the nature and evolution of the Global Observing System will have to be tackled.

**Acknowledgments** Thanks to R. Dragani, I. Štajner, C. Long, S. Eckermann, H. Eskes, S. Polavarapu and D. Jackson for providing updated information on constituent assimilation efforts at ECMWF, GMAO, NCEP, NRL, KNMI, Canada and the Met Office (UK), respectively.

## References

Austin, J., 1992. Towards the four-dimensional variational assimilation of stratospheric chemical constituents. *J. Geophys. Res.*, **97**, 2569–2588.

Austin, J., B.R. Barwell, S.J. Cox, et al., 1994. The diagnosis and forecast of clear sky ultraviolet levels at the Earth's surface. *Met. Apps.*, **1**, 321–336.

Baier, F., T. Erbertseder, O. Morgenstern, et al., 2005. Assimilation of MIPAS observations using a three-dimensional global chemistry-transport model. *Q. J. R. Meteorol. Soc.*, **131**, 3529–3542.

Bergamaschi, P., M. Krol, F. Dentener, et al., 2005. Inverse modelling of national and European $CH_4$ emissions using the atmospheric zoom model TM5. *Atmos. Chem. Phys.*, **5**, 2431–2460.

Blake, D. and R.S. Lindzen, 1973. Effect of photochemical models on calculated equilibria and cooling rates in the stratosphere. *Mon. Weather Rev.*, **101**, 783–802.

Blond, N. and R. Vautard, 2004. Three-dimensional ozone analyses and their use for short term ozone forecasts. *J. Geophys. Res.*, **109**, 10.1029/2004JD004515.

Bormann, N. and S. Healy, 2006. A fast radiative transfer model for the assimilation of infrared limb radiances from MIPAS: Accounting for horizontal gradients. *Q. J. R. Meteorol. Soc.*, **132**, 2357–2376.

Bormann, N., S. Healy and M. Hamrud, 2007. Assimilation of MIPAS limb radiances in the ECMWF system. Part II: Experiments with a 2-dimensional observation operator. *Q. J. R. Meteorol. Soc.*, **133**, 329–346.

Bormann, N., M. Matricardi and S.B. Healy, 2005. RTMIPAS: A fast radiative transfer model for the assimilation of infrared limb radiances from MIPAS. *Q. J. R. Meteorol. Soc.*, **131**, 1631–1653.

Bormann, N. and J.-N. Thépaut, 2007. Assimilation of MIPAS limb radiances in the ECMWF system. Part I: Experiments with a 1-dimensional observation operator. *Q. J. R. Meteorol. Soc.*, **133**, 309–327.

Bouttier, F. and P. Courtier, 1999. Data assimilation concepts and methods. ECMWF training notes, March 1999. Available from http://www.ecmwf.int.

Burrows, W.R., M. Vallée, D.I. Wardle, et al., 1994. The Canadian operational procedure for forecasting total ozone and UV radiation. *Met. Apps.*, **1**, 247–265.

Caplan, P., J. Derber, W. Gemmill, et al., 1997. Changes to the 1995 NCEP operational medium-range forecast model analysis – forecast system. *Weather Forecasting*, **12**, 581–594.

Cariolle, D. and M. Déqué, 1986. Southern hemisphere medium-scale waves and total ozone disturbances in a spectral general circulation model. *J. Geophys. Res.*, **91**, 10825–10884.

Cariolle, D. and J.-J. Morcrette, 2006. A linearized approach to the radiative budget of the stratosphere: influence of the ozone distribution. *Geophys. Res. Lett.*, **33**, L05806, doi:10.1029/2005GL025597.

Cariolle, D. and H. Teyssèdre, 2007. A revised linear ozone photochemistry parameterization for use in transport and general circulation models: Multi-annual simulations. *Atmos. Chem. Phys.*, **7**, 2183–2196.

Chapnik, B., G. Desroziers, F. Rabier and O. Talagrand, 2006. Diagnosis and tuning of observational error statistics in a quasioperational data assimilation setting. *Q. J. R. Meteorol. Soc.*, **132**, 543–565.

Chipperfield, M.P., B.V. Khattatov and D.J. Lary, 2002. Sequential estimation of stratospheric chemical observations in a three-dimensional model. *J. Geophys. Res.*, **107**, 10.1029/2002JD002110.

Connew, P., 1999. Chemical data assimilation using the UKMO Unified Model. *Proceedings of the SODA Workshop on Chemical Data Assimilation*, 9–10 December 1998, KNMI, De Bilt, Netherlands.

Courtier, P., 1997. Dual formulation of four-dimensional variational assimilation. *Q. J. R. Meteorol. Soc.*, **123**, 2449–2461.

Coy, L., D.R. Allen, S.D. Eckermann, et al., 2007. Effects of model chemistry and data biases on stratospheric ozone assimilation. *Atmos. Chem. Phys.*, **7**, 2917–2935.

Daley, R., 1995. Estimating the wind field from chemical constituent observations: Experiments with a one-dimensional extended Kalman filter. *Mon. Weather Rev.*, **123**, 181–198.

Davies, T., M.J.P. Cullen, A.J. Malcolm, et al., 2005 A new dynamical core for the Met Office's global and regional modelling of the atmosphere. *Q. J. R. Meteorol. Soc.*, **131**, 1759–1782.

Dee, D. and A. de Silva, 2003. The choice of variable for atmospheric moisture analysis. *Mon. Weather Rev.*, **131**, 155–171.

de Grandpré, J., R. Ménard, Y.J. Rochon, et al., 2009. Radiative impact of ozone on temperature predictability in a coupled chemistry-dynamics data assimilation system. *Mon. Weather Rev.*, **137**, 679–692.

Derber, J., H.-L. Pan, J. Alpert, et al., 1998. Changes to the 1998 NCEP operational MRF Model Analysis/Forecast system. Available from http://www.nws.noaa.gov/om/tpb/449/449body.htm.

Desroziers, G., L. Berre, B. Chapnik and P. Poli, 2005. Diagnosis of observation, background and analysis-error statistics in observation space. *Q. J. R. Meteorol. Soc.*, **131**, 3385–3396.

Dessler, A.E., 2000. *The Chemistry and Physics of Stratospheric Ozone*. Academic Press, London, 209pp.

Dethof, A., 2003. Assimilation of ozone retrievals from the MIPAS instrument onboard ENVISAT. *ECMWF Tech Memo 428.*

Dethof, A., 2004. Monitoring and assimilation of MIPAS, SCIAMACHY and GOMOS retrievals at ECMWF. ESA Contract 17585/03/IOL: Technical support for global validation of ENVISAT data products.

Dethof, A. and E. Hólm, 2004. Ozone assimilation in the ERA-40 reanalysis project. *Q. J. R. Meteorol. Soc.*, **130**, 2851–2872.

Douglass, A.R., M.R. Schoeberl, R.B. Rood and S. Pawson, 2003. Evaluation of transport in the lower tropical stratosphere in a global chemistry and transport model. *J. Geophys. Res.*, **108**, Art. No. 4259.

Dragani, R. and D. Dee, 2008. Progress in ozone monitoring and assimilation. *ECMWF Newsletter*, **116**, 35–42, Summer 2008. Available from http://www.ecmwf.int.

Eckermann, S.D., K.W. Hoppel, L. Coy, et al., 2008. High-altitude data assimilation experiments for the Northern Summer Mesosphere season of 2007. *J. Atmos. Sol Terr. Phys.*, **71**, 531–551.

ECMWF, 2007. ECMWF Newsletter No. 100 – Winter 2006/07. Available from http://www.ecmwf.int.

El Amraoui, L., P. Ricaud, J. Urban, et al., 2004. Assimilation of ODIN/SMR $O_3$ and $N_2O$ measurements in a three-dimensional chemistry transport model. *J. Geophys. Res.*, **109**, 10.1029/2004JD004796.

Elbern, H. and H. Schmidt, 1999. A four-dimensional variational chemistry data assimilation scheme for Eulerian chemistry transport modeling. *J. Geophys. Res.*, **104**, 18583–18598.

Elbern, H. and H. Schmidt, 2001. Ozone episode analysis by four-dimensional variational chemistry data assimilation. *J. Geophys. Res.*, **106**, 3569–3590.

Elbern, H., H. Schmidt and A. Ebel, 1997. Variational data assimilation for tropospheric chemistry modeling. *J. Geophys. Res.*, **102**, 15967–15985.

Elbern, H., A. Strunk, H. Schmidt and O. Talagrand, 2007. Emission rate and chemical state estimation by 4-dimensional variational inversion. *Atmos. Chem. Phys.*, **7**, 3749–3769.

El Serafy, G.Y. and H.M. Kelder, 2003. Near-real-time approach to assimilation of satellite-retrieved 3D ozone fields in a global model using a simplified Kalman filter. *Q. J. R. Meteorol. Soc.*, **129**, 3099–3120.

El Serafy, G.Y., R.J. van der A, H. Eskes and H.M. Kelder, 2002. Assimilation of 3D ozone field in global chemistry transport models using the full Kalman filter. *Adv. Space Res.*, **30**, 2473–2478.

Engelen, R.J., S. Serrar and F. Chevallier, 2009. Four-dimensional data asimilation of atmospheric $CO_2$ using AIRS observations. *J. Geophys. Res.*, **114**, D03304, doi: 10.1029/2008JD010739.

Errera, Q., S. Bonjean, S. Chabrillat, et al., 2007. BASCOE assimilation of ozone and nitrogen dioxide observed by MIPAS and GOMOS: Comparison between the two sets of analyses. ESA Special Publication SP-636.

Errera, Q., F. Daerden, S. Chabrillat, et al., 2008. 4D-Var assimilation of MIPAS chemical observations: Ozone and nitrogen dioxide analyses. *Atmos. Chem. Phys.*, **8**, 6169-6187.

Errera, Q. and D. Fonteyn, 2001. Four-dimensional variational chemical data assimilation of CRISTA stratospheric measurements. *J. Geophys. Res.*, **106**, 12253–12265.

Eskes, H.J., 2006. The integration of atmospheric chemistry observations by next generation global/hemispheric and regional and NWP models. In *Chemical Data Assimilation for the Observation of the Earth's Atmosphere.* ACCENT/WMO Expert Workshop in support of IGACO, Barrie, L.A., J.P. Burrows, P. Monks and P. Borrell (eds.), WMO Tech, Report 1360, GAW Report 169, pp 44–49.

Eskes, H.J., A. Segers and P.F.J. van Velthoven, 2005. Ozone forecasts of the Stratospheric Polar Vortex-Splitting Event in September 2002. *J. Atmos. Sci.*, **62**, 812–821.

Eskes, H.J., P.F.J. van Velthoven and H.M. Kelder, 2002. Global ozone forecasting based on ERS-2 GOME observations. *Atmos. Chem. Phys.*, **2**, 271–278.

Eskes, H.J., P.F.J. van Velthoven, P.F.M. Valks and H.M. Kelder, 2003. Assimilation of GOME total-ozone satellite observations in a three-dimensional tracer-transport model. *Q. J. R. Meteorol. Soc.*, **129**, 1663–1681.

Eyring, V., N. Butchart, D.W. Waugh, et al., 2006. Assessment of temperature, trace species, and ozone in chemistry-climate model simulations of the recent past. *J. Geophys. Res.*, **111**, 10.1029/2006JD007327.

Fierli, F., A. Hauchecorne, S. Bekki, et al., 2002. Data assimilation of stratospheric ozone using a high-resolution transport model. *Geophys. Res. Lett.*, **29**, 10.1029/2001GL014272.

Fisher, M. and D.J. Lary, 1995. Lagrangian 4-dimensional variational data assimilation of chemical species. *Q. J. R. Meteorol. Soc*., **121**, 1681–1704.

Fonteyn, D., Q. Errera, M. DeMazière, et al., 2000. 4D-Var assimilation of stratospheric aerosol satellite data. *Adv. Space Res.*, **26**, 2049–2052.

Fortuin, J.P.F. and H. Kelder, 1998. An ozone climatology based on ozonesonde and satellite measurements. *J. Geophys. Res.*, **103**, 31709–31734.

Geer, A.J., W.A. Lahoz, S. Bekki, et al., 2006a. The ASSET intercomparison of ozone analyses: Method and first results. *Atmos. Chem. Phys.*, **6**, 5445–5474.

Geer, A.J., W.A. Lahoz, D.R. Jackson, et al., 2007. Evaluation of linear ozone photochemistry parametrizations in a stratosphere-troposphere data assimilation system. *Atmos. Chem. Phys.*, **7**, 939–959.

Geer, A.J., C. Peubey, R. Bannister, et al., 2006b. Assimilation of stratospheric ozone from MIPAS into a global general circulation model: The September 2002 vortex split. *Q. J. R. Meteorol. Soc.*, **132**, 231–257.

Hadjinicolau, P., J.A. Pyle, M.P. Chipperfield and J.A. Kettleborough, 1997. Effect of interannual meteorological variability on mid latitude O3. *Geophys. Res. Lett*., **24**, 2993–2996.

Hollingsworth, A., 2005. Global Earth-system modelling using space and in situ data. *ECMWF Seminar Proceedings*, September 2005, Reading, UK. Available from http://www.ecmwf.int.

Hollingsworth, A., R.J. Engelen, C. Textor, et al., 2008. Toward a monitoring and forecasting system for atmospheric composition: The GEMS project. *Bull. Amer. Meteorol. Soc.*, **89**, doi: 10.1175/2008BAMS2355.1.

Hólm, E., E. Andersson, A. Beljaars, et al., 2002. Assimilation and modelling of the hydrological cycle: ECMWF's status and plans. *ECMWF Tech Memo 383*.

Houtekamer, P.L., L.M. Herschel, G. Pellerin, et al., 2005. Atmospheric data assimilation with an ensemble Kalman filter: Results with real observations. *Mon. Weather Rev.*, **133**, 604–620.

IGACO, 2004. The changing atmosphere. An integrated global atmospheric chemistry observation theme for the IGOS partnership. ESA SP-1282, Report GAW No. 159 (WMO TD No. 1235), September 2004; Implementation up-date, December 2004. Available from http://www.igospartners.org/docsTHEM.htm.

Jackson, D.R., 2004. Improvements in data assimilation at the Met Office. *Forecasting Research Technical Report No. 454*, Met Office.

Jackson, D.R., 2007. Assimilation of EOS MLS ozone observations in the met office data assimilation system. *Q. J. R. Meteorol. Soc.*, **133**, 1771–1788.

Jackson, D.R. and Y.J. Orsolini, 2008. Estimation of Arctic ozone loss in winter 2004/05 based on assimilation of EOS MLS and SBUV/2 observations. *Q. J. R. Meteorol. Soc.*, **134**, 1833–1841.

Jackson, D.R. and R. Saunders, 2002. Ozone data assimilation: Preliminary system. *Forecasting Research Technical Report No. 394*, Met Office.

Juckes, M.N., 2006. Evaluation of MIPAS ozone fields assimilated using a new algorithm constrained by isentropic tracer advection. *Atmos. Chem. Phys*., **6**, 1549–1565.

Khattatov, B.V., 2003. Multivariate chemical data assimilation. In *Data Assimilation for the Earth System.* NATO Science Series: IV. Earth and Environmental Sciences 26, Swinbank, R., V. Shutyaev and W.A. Lahoz, Kluwer Academic Publishers, Dordrecht, The Netherlands, pp 279–288, 378pp.

Khattatov, B.V., J.C. Gille, L.V. Lyjak, et al., 1999. Assimilation of photochemically active species and a case analysis of UARS data. *J. Geophys. Res.*, **104**, 18715–18737.

Khattatov, B.V., J.-F. Lamarque, L.V. Lyjak, et al., 2000. Assimilation of satellite observations of long-lived chemical species in global chemistry transport models. *J. Geophys. Res.*, **105**, 29135–29144.

Khattatov, B., L. Lyjak and J. Gille, 2001. On applications of photochemical models to the design of measurement strategies. *Geophys. Res. Lett.*, **28**, 2377–2380.

Lahoz, W.A., 2006. Chemical data assimilation: Choices and challenges. In *Chemical Data Assimilation for the Observation of the Earth's Atmosphere.* ACCENT/WMO Expert Workshop in support of IGACO, Barrie, L.A., J.P. Burrows, P. Monks and P. Borrell (eds.), WMO Tech, Report 1360, GAW Report 169, pp 106–110.

Lahoz, W.A., R. Brugge, D.R. Jackson, et al., 2005. An observing system simulation experiment to evaluate the scientific merit of wind and ozone measurements from the future SWIFT instrument. *Q. J. R. Meteorol. Soc.*, **131**, 503–523.

Lahoz, W.A., Q. Errera, R. Swinbank and D. Fonteyn, 2007a. Data assimilation of stratospheric constituents: A review. *Atmos. Chem. Phys.*, **7**, 5745–5773.

Lahoz, W.A., A.J. Geer, S. Bekki, et al., 2007b. The Assimilation of Envisat data (ASSET) project. *Atmos. Chem. Phys.*, **7**, 1773–1796.

Lamarque, J.-F. and J.C. Gille, 2003. Improving the modeling of error variance evolution in the assimilation of chemical species: Applications to MOPITT data. *Geophys. Res. Lett.*, **30**, 10.1029/2003GL016994.

Lamarque, J.-F., B.V. Khattatov and J.C. Gille, 2002. Constraining tropospheric ozone column through data assimilation. *J. Geophys. Res.*, **107**, 10.1029/2001JD001249.

Lary, D.J., 1999. Data assimilation: A powerful tool for atmospheric chemistry. *Phil. Trans. R. Soc. Lond.*, **A357**, 3445–3457.

Lary, D.J., B. Khattatov and H.Y. Mussa, 2003. Chemical data assimilation: A case study of solar occultation data from the ATLAS 1 mission of the Atmospheric Trace Molecule Spectroscopy Experiment (ATMOS). *J. Geophys. Res.*, **108**, 10.101029/2003JD003500.

Lemus-Deschamps, L., S. Grainger, L. Rikus, et al., 2005. Australian UV and ozone forecasting system. Available from http://www.bom.gov.au/bmrc/mdev/expt/uvindex/uvi.shtml.

Levelt, P.F., B.V. Khattatov, J.C. Gille, et al., 1998. Assimilation of MLS ozone measurements in the global three-dimensional chemistry transport model ROSE. *Geophys. Res. Lett.*, **25**, 4493–4496.

Lindzen, R.S. and R. Goody, 1965. Radiative and photochemical processes in mesospheric dynamics: Part I, models for radiative and photochemical processes. *J. Atmos. Sci.*, **22**, 341–348.

Logan, J.A., 1999. An analysis of ozonesonde data for the troposphere: Recommendations for testing 3-D models, and development of a gridded climatology for tropospheric ozone. *J. Geophys. Res.*, **104**, 16115–16149.

Long, C.S., 2003. UV Index forecasting practices around the world. *SPARC Newsletter no. 21*, June 2003.

Long, C.S., S. Zhu and R. Treadon, 2007. Assimilation of multiple ozone products into the NCEP operational forecast model. Presentation at the SPARC Data Assimilation Workshop, Toronto, September 2007, abstract available from http://atlas-conferences.com/c/a/u/e/10.htm.

Lorenc, A.C., 2003. The potential of the ensemble Kalman filter for NWP: A comparison with 4D-Var. *Q. J. R. Meteorol. Soc.*, **129**, 3183–3204.

Lorenc, A.C. and O. Hammon, 1988. Objective quality control of observations using Bayesian methods: Theory and practical implementation. *Q. J. R. Meteorol. Soc*., **114**, 515–543.

Louvel, S., 2001. Implementation of a dual variational algorithm for assimilation of synthetic altimeter data in the oceanic primitive equation model MICOM. *J. Geophys. Res.*, **106**, 9199–9212.

Lyster, P.M., S.E. Cohn, R. Ménard, et al., 1997. Parallel implementation of a Kalman filter for constituent data assimilation. *Mon. Weather Rev.,* **125**, 1674–1686.

Marchand, M., S. Bekki, L. Denis and J.-P. Pommereau, 2003. Test of the nightime polar stratospheric $NO_2$ decay using wintertime SAOZ measurements and chemical data assimilation. *Geophys. Res. Lett.*, **30**, 10.1029/2003GL017582.

Marchand, M., S. Bekki, A. Hauchecorne and J.-L. Bertaux, 2004. Validation of the self-consistency of GOMOS $NO_3$, $NO_2$ and $O_3$ data using chemical data assimilation. *Geophys. Res. Lett.*, **31**, 10.1029/2004GL019631.

Marchand, M., S. Bekki, F. Lefèvre and A. Hauchecorne, 2007. Temperature retrieval from stratospheric $O_3$ and $NO_3$ GOMOS data. *Geophys. Res. Lett.*, **34**, L24809, doi: 10.1029/2007GL030280.

Massart, S., D. Cariolle and V.-H. Peuch, 2004. Towards an improvement of the atmospheric ozone distribution and variability by the assimilation of satellite data. *C.R. Geosci.*, **15**, 1305–1310.

Massart, S., C. Clerbaux, D. Cariolle, et al., 2009. First steps towards the assimilation of IASI ozone data into the MOCAGE-PALM system. *Atmos. Chem. Phys*., **9**, 5073–5091.

Mathison, C., D.R. Jackson and M. Keil, 2007. Methods of improving the representation of ozone in the Met Office model. *NWP Tech. Report No. 502*, Met Office.

McCormack J.P., S.D. Eckermann, L. Coy, et al., 2004. NOGAPS-ALPHA model simulations of stratospheric ozone during the SOLVE2 campaign, *Atmos. Chem. Phys*., **4**, 2401–2423.

McCormack, J.P., S.D. Eckermann, D.E. Siskind and T.J. McGee, 2006. CHEM2D-OPP: A new linearized gas-phase ozone photochemistry parameterization for high-altitude NWP and climate models. *Atmos. Chem. Phys.*, **6**, 4943–497.

McLaughlin, D., A. O'Neill, J. Derber and M. Kamachi, 2005. Opportunities for enhanced collaboration within the data assimilation community. *Q. J. R. Meteorol. Soc.*, **131**, 3683–3693.

McLinden C.A., S.C. Olsen B. Hannegan, et al., 2000. Stratospheric ozone in 3-D models: A simple chemistry and the cross-tropopause flux. *J. Geophys. Res.*, **105**, 14653–14665.

McNally, A.P., P. Watts, J. Smith, et al., 2006. The assimilation of AIRS radiance data at ECMWF. *Q. J. R. Meteorol. Soc.,* **132**, 935–958.

Meirink, J.F., H.J. Eskes and A.P.H. Goede, 2006. Sensitivity analysis of methane emissions derived from SCIAMACHY observations through inverse modelling. *Atmos. Chem. Phys.*, **6**, 9405–9445.

Ménard, R., S. Chabrillat, C. Charette, et al., 2007. Coupled chemistry-dynamics data assimilation. Presentation at the SPARC Data Assimilation Workshop, Toronto, September 2007, abstract available from http://atlas-conferences.com/c/a/u/e/25.htm.

Ménard, R. and L.-P. Chang, 2000. Stratospheric assimilation of chemical tracer observations using a Kalman filter, Part II: Chi-squared validated results and analysis of variance and correlation dynamics. *Mon. Weather Rev.*, **128**, 2672–2686.

Ménard, R., S.E. Cohn, L.P. Chang and P.M. Lyster, 2000. Stratospheric assimilation of chemical tracer observations using a Kalman filter, Part I: Formulation. *Mon. Weather Rev.*, **128**, 2654–2671.

Migliorini, S., C. Piccolo and C.D. Rodgers, 2004. Intercomparison of direct and indirect measurements: MIPAS versus sonde ozone profiles. *J. Geophys. Res.*, **109**, 10.1029/2004JD004988.

Monge-Sanz, B.M., M.P. Chipperfield, A.J. Simmons and S.M. Uppala, 2007. Mean age of air and transport in a CTM: Comparison of different ECMWF analyses. *Geophys. Res. Lett.*, **34**, 10.1029/2006GL028515.

Monks, P.S., 2003. Tropospheric photochemistry. In *Handbook of Atmospheric Sciences*, Hewitt, C.N. and A.V. Jackson (eds.), Blackwell Science, Oxford, pp 156–187.

Morcrette, J.-J., 2003. Ozone radiation interactions in the ECMWF forecast system. *ECMWF Tech Memo 375*.

Müller, M.D., P.K. Bhartia and I. Štajner, 2004. Assimilation of SBUV version 8 radiances into the GEOS DAS. *Proceedings of the Quadrennial Ozone Symposium, GOS, Kos*, Greece, June 2004. Available from ftp://gmaoftp.gsfc.nasa.gov/pub/papers/ivanka/ozone_papers/QOS04.Mueller.pdf.

Müller, J.-F. and T. Stavrakou, 2005. Inversion of CO and $NO_x$ emissions using the adjoint of the IMAGES model. *Atmos. Chem. Phys.*, **5**, 1157–1186.

Oikonomou, E. and A. O'Neill, 2006. An evaluation of water vapour and ozone in the ERA-40 reanalysis compared with UARS data. *J. Geophys. Res.*, **111**, 10.1029/2004JD005341.

Orsolini, Y.J. and G. Nikulin, 2006. A low-ozone episode during the European heat wave of August 2003. *Q. J. R. Meteorol. Soc.*, **132**, 667–680.

Pétron, G., C. Granier, B. Khattatov, et al., 2004. Monthly CO surface sources inventory based on the 2000–2001 MOPITT satellite data. *Geophys. Res. Lett.*, **31**, 10.1029/2004GL020560.

Peuch, A., J.-N. Thépaut and J. Pailleux, 2000. Dynamical impact of total-ozone observations in a four-dimensional variational assimilation. *Q. J. R. Meteorol. Soc.*, **126**, 1641–1659.

Plumb, R.A. and M.K.W. Ko, 1992. Interrelationships between mixing ratios of long-lived stratospheric constituents. *J. Geophys. Res.*, **97**, 10145–10156.

Polavarapu, S., S. Ren, Y. Rochon, et al., 2005a. Data assimilation with the Canadian middle atmosphere model. *Atmos. Ocean*, **43**, 77–100.

Polavarapu, S., T.G. Shepherd, Y. Rochon and S. Ren, 2005b. Some challenges of middle atmosphere data assimilation. *Q. J. R. Meteorol. Soc.*, **131**, 3513–3527.

Rawlins, F., S.P. Ballard, K.J. Bovis, et al., 2007. The met office global four-dimensional variational data assimilation scheme. *Q. J. R. Meteorol. Soc.*, **133**, 347–362.

Rienecker, M.M., M.J. Suarez, R. Todling, et al., 2008. The GEOS-5 data assimilation system – documentation of versions 5.01, 5.1.0 and 5.2.0. *NASA Tech. Memo.*, NASA/TM-2008-104606, Vol. 27, 102 pp. Available from http://gmao.gsfc.nasa.gov/pubs/docs/GEOSS_104606_Vol27.pdf.

Riishøjgaard, L.-P., 1996. On four-dimensional variational assimilation of ozone data in weather-prediction models. *Q. J. R. Meteorol. Soc.*, **122**, 1545–1572.

Riishøjgaard, L.-P., I. Štajner and G.-P. Lou, 2000. The GEOS ozone data assimilation system. *Adv. Space Res.*, **25**, 1063–1072.

Rood, R.B., 2003. Ozone assimilation. In *Data Assimilation for the Earth System.* NATO Science Series: IV. Earth and Environmental Sciences 26, Swinbank, R., V. Shutyaev and W.A. Lahoz, Kluwer Academic Publishers, Dordrecht, The Netherlands, pp 263–277, 378pp.

Rood, R.B., 2005. Assimilation of stratospheric meteorological and constituent observations: A Review. *SPARC Newsletter no. 25*, July 2005.

Rösevall, J.D., D.P. Murtagh and J. Urban, 2007a. Ozone depletion in the 2006/2007 Arctic winter. *Geophys. Res. Lett.*, **34**, L21809, doi:10.1029/2007GL030620.

Rösevall, J.D., D.P. Murtagh, J. Urban and A.K. Jones, 2007b. A study of polar ozone depletion based on sequential assimilation of satellite data from the ENVISAT/MIPAS and Odin/SMR instruments. *Atmos. Chem. Phys.*, **7**, 899–911.

Russell III, J.M., L.L. Gordley, J.H. Park, et al., 1993. The Halogen occultation experiment. *J. Geophys. Res.*, **98**, 10777–10797.

Sassi, F., B.A. Boville, D. Kinnison and R.R. Garcia, 2005. The effects of interactive ozone chemistry on simulations of the middle atmosphere. *Geophys. Res. Lett.*, **32**, 10.1029/2004GL022131.

Saunders, R., M. Matricardi and P. Brunel, 1999. An improved fast radiative transfer model for assimilation of satellite radiance observations. *Q. J. R. Meteorol. Soc.,* **125**, 1407–1426.

Schoeberl, M.R., A.R. Douglass, Z. Zhu and S. Pawson, 2003. A comparison of the lower stratospheric age spectra derived from a general circulation model and two data assimilation systems. *J. Geophys. Res.*, **108**, 10.1029/2002JD002652.

Segers, A.J., H.J. Eskes, A.R.J. van der, et al., 2005. Assimilation of GOME ozone profiles and a global chemistry-transport model. *Q. J. R. Meteorol. Soc.*, **131**, 477–502.

Semane, N., V.-H. Peuch, S. Pradier, et al., 2009. On the extraction of wind information from the assimilation of ozone profiles in Météo-France 4D-Var operational NWP suite. *Atmos. Chem. Phys.*, **9**, 4855–4867.

Simmons, A.J., M. Hortal, G. Kelly, et al., 2005. ECMWF analyses and forecasts of stratospheric winter polar vortex breakup: September 2002 in the Southern Hemisphere and related events. *J. Atmos. Sci.*, **62**, 668–689.

SPARC, 2000. SPARC: Assessment of upper tropospheric and lower stratospheric water vapour. WCRP-113, WMO/TD No. 1043, SPARC Report No. 2, Kley, D., J.M. Russell and C. Phillips (eds.), 2000.

Štajner, I., L.-P. Riishøjgaard and R.B. Rood, 2001. The GEOS ozone data assimilation system: Specification of error statistics. *Q. J. R. Meteorol. Soc.*, **127**, 1069–1094.

Štajner, I. and K. Wargan, 2004. Antarctic stratospheric ozone from the assimilation of occultation data. *Geophys. Res. Lett.*, **31**, 10.1029/2004GL020846.

Štajner, I., K. Wargan, L.-P. Chang, et al., 2006. Assimilation of ozone profiles from the improved limb atmospheric spectrometer-II: Study of Antarctic ozone. *J. Geophys. Res.*, **111**, 10.1029/2005JD006448.

Štajner, I., K. Wargan, S. Pawson, et al., 2008. Assimilated ozone from EOS-Aura: Evaluation of the tropopause region and tropospheric columns. *J. Geophys. Res.*, **113**, D16532, doi: 10.1029/2007JD008863.

Štajner, I., N. Winslow, R.B. Rood and S. Pawson, 2004. Monitoring of observation errors in the assimilation of satellite ozone data. *J. Geophys. Res.*, **109**, 10.1029/2003JD006309.

Stolarski, R. and A.R. Douglass, 1985. Parameterization of the photochemistry of stratospheric ozone including catalytic processes. *J. Geophys. Res.*, **90**, 10709–10718.

Streibel, M., M. Rex, P. von der Gathen, et al., 2006. Chemical ozone loss in the Arctic winter 2002/2003 determined with Match. *Atmos. Chem. Phys.*, **6**, 2783–2792.

Struthers, H., R. Brugge, W.A. Lahoz, et al., 2002. Assimilation of ozone profiles and total column measurements into a global general circulation model. *J. Geophys. Res.*, **107**, 10.1029/2001JD000957.

Talagrand, O., 2003. *A posteriori* validation of assimilation algorithms. In *Data Assimilation for the Earth System.* NATO Science Series: IV. Earth and Environmental Sciences 26, Swinbank, R., V. Shutyaev and W.A. Lahoz (eds.), Kluwer Academic Publishers, Dordrecht, The Netherlands, pp 85–95, 378pp.

Tan, W.-W., M.A. Geller, S. Pawson and A. de Silva, 2004. A case study of excessive subtropical transport in the stratosphere of a data assimilation system. *J. Geophys. Res.*, **109**, Art. No. D11102.

Tangborn, A., I. Štajner, M. Buchwitz, et al., 2009. Assimilation of SCIAMACHY total column CO observations: Global and regional analysis of data impact. *J. Geophys. Res.*, **114**, 10.1029/2008JD010781.

Thornton, H., D.R. Jackson, S. Bekki, et al., 2009. The ASSET intercomparison of stratosphere and lower mesosphere humidity analyses. *Atmos. Chem. Phys*., **9**, 995–1016.

Trenberth, K. (ed.), 1992. *Climate System Modeling*, Cambridge University Press, Cambridge, 788pp.

Uppala, S.M., P.W. Kållberg, A.J. Simmons, et al., 2005. The ERA-40 re-analysis. *Q. J. R. Meteorol. Soc.*, **131**, 2961–3012.

Vigouroux, C., M. De Mazière, Q. Errera, et al., 2007. Comparison between ground-based FTIR and MIPAS $N_2O$ and $HNO_3$ profiles before and after assimilation in BASCOE. *Atmos. Chem. Phys.*, **7**, 377–396.

Wang, K.-Y., D.J. Lary, D.E. Shallcross, et al., 2001. A review on the use of the adjoint method in four-dimensional atmospheric-chemistry data assimilation. *Q. J. R. Meteorol. Soc.*, **127**, 2181–2204.

Wargan, K., I. Štajner, S. Pawson, et al., 2005. Assimilation of ozone data from the Michelson interferometer for passive atmospheric sounding. *Q. J. R. Meteorol. Soc.*, **131**, 2713–2734.

WMO, 2006. Scientific Assessment of Ozone Depletion, 2006. World Meteorological Organization, Global Ozone Research and Monitoring Project, Report No. 50. Available from http://www.wmo.ch/web/arep/reports/ozone_2006/ozone_asst_report.html.

# Inverse Modelling and Combined State-Source Estimation for Chemical Weather

**Hendrik Elbern, Achim Strunk, and Lars Nieradzik**

## 1 Introduction

### *1.1 General Remarks*

Air quality data assimilation aims to find a best estimate of the control parameters (see chapters in Part I, *Theory*) for those processes of the atmosphere which govern the chemical evolution of biologically relevant height levels, typically located in the the lowermost atmosphere. As in data assimilation (see chapters in Part I, *Theory*), we have to resort to numerical models to complement usually sparse observation networks; these models serve as system constraints. Several research groups are developing data assimilation methods similar to those applied to meteorological applications. Techniques range from nudging to advanced spatio–temporal methods such as four-dimensional variational (4D-Var) data assimilation and various simplifications of the Kalman filter (KF).

The different observation methodologies (see chapters in Part II, *Observations*) imply heterogeneities in terms of accuracy, spatial representativity and density, sampling frequency, and various retrieval techniques. Similarly for the range of models and the information they provide. The appropriate approach to bring together (fuse) and analyse this observational and model information involves advanced data assimilation and inverse modelling techniques.

Remotely sensed Earth Observation data, primarily from polar orbiting platforms (see chapters in Part II, *Observations*), are scattered in space and time, providing only a very little fraction of the data at a single point in time. Therefore, a prerequisite for a full exploitation of these sensors is the use of numerical models for spatio–temporal interpolation by assimilation of data.

H. Elbern (✉)
Research Centre Jülich, Rhenish Institute for Environmental Research at the University of Cologne and at ICG-2, Cologne, Germany; Helmholtz Virtual Institute for Inverse Modelling of Atmospheric Chemical Composition (IMACCO), Cologne, Germany
e-mail: he@eurad.uni-koeln.de

W. Lahoz et al. (eds.), *Data Assimilation*, DOI 10.1007/978-3-540-74703-1_19,

## *1.2 Features of Tropospheric Chemical Data Assimilation*

In tropospheric chemistry data assimilation, a variety of aspects has to be considered; these differ considerably from traditional atmospheric data assimilation. As usual in atmospheric chemistry, the number of parameters per grid point is much higher than in meteorology. For example, a state-of-the-art air quality model prognoses more than 50 constituents in the gas phase only. If aerosol dynamics and chemistry are included, this number can easily double.

Furthermore, the underlying chemistry models are only an approximation to the most important constituents. For example, hydrocarbons, referred to as volatile organic compounds (VOCs) occur in a variety of components which cannot be accounted for in a complete way. Also, aerosol particles differ in their size, shape, chemical composition and complexity of reactions, which makes them hard to account for in models.

When the focus of the chemistry analysis is the Earth's surface, shorter temporal scales and smaller spatial scales become increasingly important. Local air quality is forced by local emissions, with background values controlled by transport processes at larger spatial scales. As a consequence, a sequence of spatial scales should be covered from the long range, even intercontinental transport of pollutants, down to a representation of emissions from point and line sources like chimney stacks and streets. In practice, different chemical regimes co-exist at the smaller spatial scales. Sinks act by surface uptake from soil and vegetation, again imposing a much finer spatial pattern, analogous to that exhibited by typical mesoscale meteorological features.

Figure 1 presents an example of the different spatial scales involved. Note the resolution factor of 27 between both graphics. Simulations of tropospheric chemistry must be able to span this range of spatial scales.

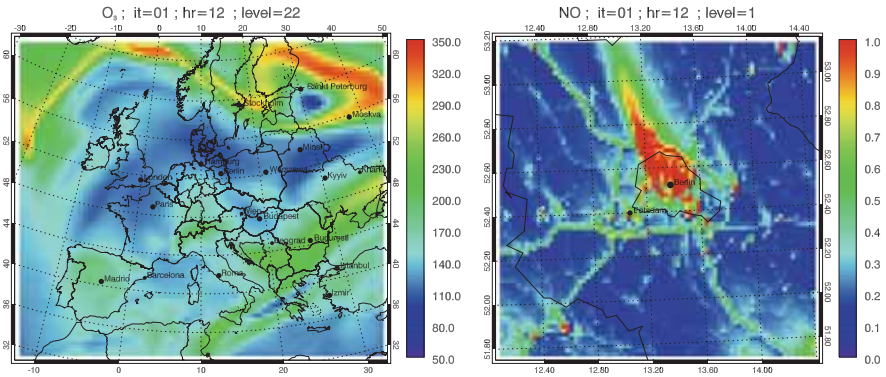



**Fig. 1** Simulated ozone state for July 20, 1998. Ozone field at 10 km height over Europe with 54 km horizontal resolution (*left panel*), and nitrogen oxide (NO) field of the greater Berlin (Germany) area with 2 km resolution (*right panel*)

For atmospheric chemistry data assimilation, as in meteorology, initial state variables of the model are usually the parameters optimized. Hence, these initial values are implicitly assumed to be the least well known parameters and a critical factor for improved analysis or forecast skill. Because tropospheric chemistry–transport models (CTMs) solve an initial–boundary value problem which is strongly dependent on surface parameters, the restriction to initial value optimization only is no longer justified. Furthermore, it is well known that under favourable conditions, freshly emitted surface pollutants can easily enter the free and upper troposphere. Therefore, better knowledge of emission strengths and meteorological conditions is likely to be at least as important as knowledge of initial values.

Errors in emission rate estimates can be considered as one of the main sources of uncertainty in predictions of pollution. Consequently, emission rates must be considered as an optimization parameter. Figure 2 demonstrates the limited memory of analyses based on initial value optimization only, and the deficiencies of only using emission rate optimization. In the second case, to compensate for an emission bias, compliance with observations can only be approximated by over-tuning of the model, with resulting forecast failures. Thus, combined and balanced initial

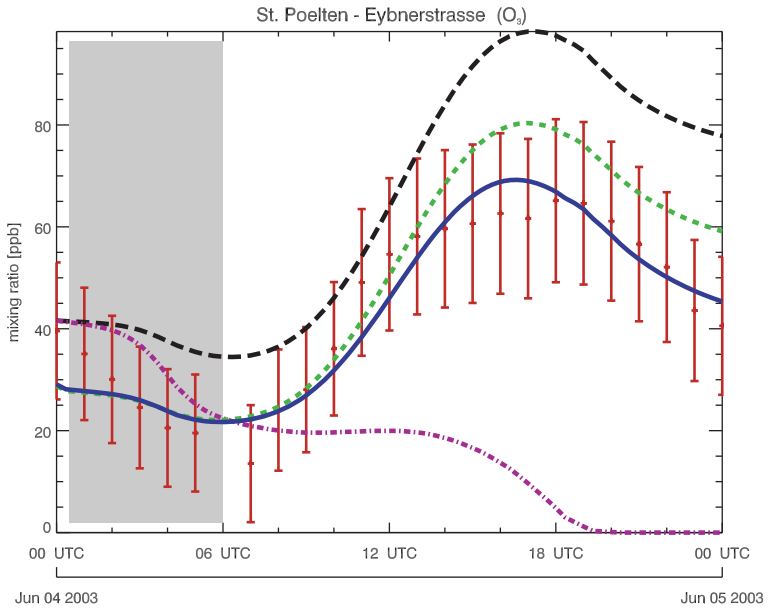



**Fig. 2** Ozone forecasts at the measurement site of St. Poelten (Austria) after assimilation of data inside the time interval 0000–0600 UTC, June 4, 2003 (*shaded area*). The forecasts are based on initial value optimization (*green dotted line*), emission rate optimization (*pink dash-dotted line*), and combined emission rate/initial value optimization (*bold blue line*). The control run without data assimilation is shown for reference (*black dashed line*). Observations are shown in *red*, with error bars indicated by *red vertical lines*

value/emission rate optimization appears to provide the best solution. Using similar arguments, deposition velocities, if poorly known, could also be considered as a variable to be optimized.

A thorough assessment of uncertainties in ozone forecasts due to uncertainties in various input parameters has been provided by various studies, e.g., Hanna et al. (1998), Hanna et al. (2001), Schmidt and Martin (2003). While parameters like photolysis rates and meteorological conditions are important, emissions are still the most important control parameters in the state vector (see chapters in Part I, *Theory*).

In summary, we seek a data assimilation algorithm, which is able to combine heterogeneous observations (scattered in space and time, and having variable spatial and temporal representativity) with an air quality model system. This requires the application of space-time data assimilation algorithms preserving the BLUE (Best Linear Unbiased Estimate) property (Talagrand 1998).

## *1.3 Observations*

Ground-based in situ observations of chemical constituents are the backbone of the observation suite; they are usually provided by regional national or European environmental protection agencies. Typically, ozone, nitrogen dioxide, sulphur, carbon monoxide and particulate matter integrated up to 10 $\mu$m ($PM_{10}$) or 2.5 $\mu$m ($PM_{2.5}$) are measured. While these measurement sites operate on a regular basis (sometimes the data can be provided in near real time), the deployment strategy of site locations is not adapted to data assimilation needs, unlike for meteorological data assimilation (see chapter *Assimilation of Operational Data*, Andersson and Thépaut). In particular, locations are often not spatially representative of model grid sizes larger than 10 km resolution, and the density of measurement sites is biased toward densely populated areas. These facts must be considered when interpreting chemical data assimilation results. As a critical consequence of the different scales and chemical regimes (see above), legacy surface in situ observations are not only sparse, but are also hampered by this spatial representativity problem in populated areas. The available ozone radiosonde network is even much sparser.

On special occasions, there are campaign data available, with a much larger quantity of measurements. However, typically, these data sets are spatially and temporally limited.

Satellite data are a highly valuable complement to in situ data. However, the following has to be considered. Tropospheric satellite data are limited to only very few species and often only given in terms of tropospheric columns. Available data include nitrogen dioxide, elevated levels of sulphur dioxide and formaldehyde, mostly retrieved from GOME, Global Ozone Monitoring Experiment (onboard the ERS-2 platform) or SCIAMACHY (Scanning Imaging Absorption spectrometer for Atmospheric CHartographY), aboard Envisat (e.g., Eskes and Boersma 2003; Heue et al. 2005). Recently, further tropospheric column data has become available through OMI (Ozone Monitoring Instrument) and GOME-2

sensors onboard EOS Aura and METOP, respectively. Carbon monoxide soundings from MOPITT (Measurements Of Pollution in The Troposphere) sensors (Deeter et al. 2003) as well as neural network retrieved ozone profiles are available (Müller et al. 2003). A full overview is presented in chapter *Research Satellites* by Lahoz.

Finally, in situ observations from commercial aircraft, in the framework of the MOZAIC (Measurement of Ozone and Water Vapor by Airbus In-Service Aircraft) activity (Thouret et al. 2000) can be assimilated. The MOZAIC initiative (Marenco et al. 1998) consists of automatic and regular measurements by long range passenger airliners flying all over the world. A special element of this observing system are vertical profiles of observations obtained during the take-off and landing phases. Target species are $O_3$, water vapour, CO and total reactive nitrogen, $NO_y$, measurements. The next generation of aircraft data will be provided by IAGOS (Integration of routine Aircraft measurements into a Global Observing System). A survey on observations for data assimilation is presented in chapter *The Global Observing System* (Thépaut and Andersson).

Section 2 presents a brief review of advanced tropospheric chemistry data assimilation. Section 3 features the theoretical approach of the complexity reduced Kalman filter and 4-dimensional variational data assimilation (4D-Var), and an implementation of a comprehensive system. These two methods are regarded as the most advanced methods of the data assimilation canon. Sample results are provided in Sect. 4.

## 2 Spatio-Temporal Data Assimilation Studies

### *2.1 Tropospheric Gas Phase Data Assimilation*

We present a short and non-exhaustive survey of studies performed in tropospheric data assimilation, emphasizing the algorithms developed. We focus on the methodological differences between tropospheric chemistry data assimilation and meteorological and upper atmosphere chemistry data assimilation, hence partly complementing chapter *Assimilation of Operational Data* (Andersson and Thépaut).

Early attempts to analyse tracer fields were based on monovariate kriging techniques of surface concentrations (e.g., Fedorov 1998). These methods produce chemical state estimates, frequently referred to as analyses. When applying purely spatial techniques, there are very little significant differences between chemistry data assimilation and meteorological data assimilation. This is in contrast to tropospheric chemistry *spatio-temporal* data assimilation, where there are notable differences. Attempting to combine observations made at different times, intermittently applied spatial data assimilation procedures cannot make best use of known physical and chemical laws as constraints. The ability to do so would not only enlarge the observational data base for data assimilation with measurements made over a full time interval but, depending on the model set-up, would also enforce a degree of chemical consistency.

From a theoretical viewpoint, only advanced spatio-temporal data assimilation or inversion techniques are candidates for a solution to this problem. These methods are able to combine model information with data in a consistent way while, at the same time, providing a Best Linear Unbiased Estimate (BLUE). One underlying assumption of the BLUE is the validity of Gaussian probability density functions. Under not very stringent conditions, the BLUE property is satisfied by the 4D-Var technique and the Kalman filter (This is also true for the variants of the Kalman filter commonly applied.).

In the framework of an identical twin set-up, a first implementation of the 4D-variational technique for emission optimization including reactive chemistry is described in Elbern et al. (2000). A first real-world application with a fully fledged CTM, the EURAD (EURopean Air pollution Dispersion) model, is given in Elbern and Schmidt (2001). By including all emitted species at each surface grid point, the typical optimization space of initial values for atmospheric chemical state constituents is replaced by a scaled emission rate space in Elbern et al. (2007). An example on the assimilation of satellite data in tropospheric chemistry models is given by a dedicated section in Lahoz et al. (2007).

A practical application on the microscale has been presented by Quélo et al. (2005) for $NO_x$ emissions and their diurnal profile, using the Polair3D model. Another regional tropospheric 4D-Var assimilation system is STEM-2K1 (Chai et al. 2006). Dedicated flight missions also provide measurements for spatio-temporal data assimilation. The measurements obtained during the NASA Transport and Chemical Evolution over the Pacific (TRACE-P) airborne mission (see, e.g., Talbot et al. 2003) were the first tropospheric airborne data assimilated using the chemical 4D-Var method. Chai et al. (2006) made use of the limited area Sulfur Transport Eulerian Model, version 2K1 (STEM-2K1) and its adjoint. The assimilation set-up simplifies the background error covariance matrix to a diagonal formulation. To assess the added value of the STEM-2K1 model, see Carmichael et al. (2003) for a discussion of TRACE-P related model experiments. Comparison of model simulations of the TRACE-P mission with and without advanced data assimilation, exhibited several interesting features. For example, the authors found that assimilating ozone observations from one of two independent flights improved model prediction of the other flight which used ozone measurements withheld for validation. Specifically, after only assimilating $NO_y$ observations, the adjusted initial fields led to better predictions of NO, $NO_2$, and PAN (peroxyacetyl nitrate), based on a comparison with the withheld measurements. In addition, the model predictions of $NO_y$ improved significantly after assimilating the observations of the aforementioned chemical species, which are independent of the withheld $NO_y$ measurements. Adopting the variational inversion technique at the global scale, Müller and Stavrakou (2005) assimilate tropospheric column retrievals of CO and $NO_2$, to assess emission rates at continental scales.

In the Netherlands, two CTMs have been coupled to sophisticated variants of the complexity reduced Kalman filter. These include the reduced rank square root Kalman filter of the Long Term Ozone Simulation (LOTOS) model (van Loon et al.

2000) and the EUROS model (Hanea et al. 2004). The reduced rank square root approach is set up to factorize covariance matrices using a few principal components (Verlaan and Heemink 1995). Further elaboration of this technique by its combination with an ensemble Kalman filter method resulted in additional skill (Hanea et al. 2004). Optimization parameters include emission rates, photolysis rates, and deposition rates, corrections for which are formally introduced as "noise" parameters in the Kalman filter formulation.

Other very recent applications of the ensemble Kalman filter are due to Constantinescu et al. (2007a, b). In this approach, the covariance matrix is not constructed by the ensemble directly, but by autoregressive modelling. Nevertheless, implementation of special measures has proved necessary to avoid "ensemble collapse" effects with "filter divergence", an effect due to the insufficient spread of the ensemble members and an inadequate model error covariance formulation.

Independently of activities termed "data assimilation", research on the solution of inversion problems to provide source and sink estimates has been well established over the last few decades. In most cases, inversion with respect to quasi-passive tracers has been performed. Newsam and Enting (1988) and Enting and Newsam (1990) addressed the global problem of the distribution of sources and sinks of carbon dioxide by the inversion of a diffusion equation, formally solved using associated Legendre functions.

Following this work, a variety of other studies were made, all based on a very limited number of flask measurements (Bousquet et al. 1999a, b; Enting et al. 1995; Fan et al. 1998; Gloor et al. 1999; Gurney 2002). The variational approach has been also adopted for source and sink estimates, with the aim of providing a better specification of greenhouse gas budgets (Kaminski et al. 1999a, b; Houweling et al. 1999).

In order to optimize model parameters, Kaminski et al. (2002) assimilated 41 $CO_2$ measurement sets into a simplified terrestrial biosphere model using the 4D-Var technique, thereby achieving more realistic flux simulations. To overcome the limitations of $CO_2$ in situ observations, satellite data from the Atmospheric Infrared Sounder (AIRS) have been assimilated into the European Centre for Medium-range Weather Forecasts (ECMWF) model using the 4D-Var technique by Engelen et al. (2004). As results were only satisfactory in the tropical regions, improved global source and sink estimates cannot be expected with current database and assimilation system configurations.

At the mesoscale, Robertson and Langner (1992) used variational data assimilation for source estimation in the framework of ETEX (European Tracer Experiment). Using adjoint modelling ideas, Issartel (2003) applied the concept of retro-plumes for source identification. A different approach was taken by Bocquet (2005a, b), the maximum entropy principle was invoked to estimate the position, time, and strength of emission sources.

All emission source studies cited above have focused on source or sink estimates of a single passive tracer, without modelling reactive chemistry. Only a few attempts

have been made to solve the source inversion problem for reactive chemistry, where precursor sources are estimated using observational data from product pollutants.

## *2.2 Tropospheric Aerosol Data Assimilation*

In recent years, both model based chemistry data assimilation and complex aerosol modelling using fully-fledged air quality models have become increasingly important. As both disciplines are challenging in terms of development and computational demands, the attempt to combine state-of-the-art data assimilation methods with state-of-the-art aerosol modules, in turn combined with advanced satellite retrieval methods, has not yet been made. Instead, all these lines of work have evolved separately. Two examples are given to demonstrate the current state of affairs. In Collins et al. (2001) the authors applied the MATCH model, in which sulphate, black carbon, organic carbon and mineral dust are predicted while sea salt aerosols are diagnosed. Optimal interpolation is applied as the assimilation scheme. The assimilation parameter is the aerosol optical depth (AOD) retrieved from NOAA AVHRR (Advanced Very High Resolution Radiometer) over the oceans. By contrast, van Loon et al. (2000) used a Reduced Rank Square Root Kalman filter (RRSRKF) to assimilate AOD from ATSR-2 (Along Track Scanning Radiometer) into the LOTOS model, which crudely estimates the model AOD by doubling the value resulting from modelled $SO_4^{2-}$, $NO_3^-$, and $NH_4^+$. A variational approach for aerosol dynamics in a box model is presented by Sander et al. (2005), who use an adjoint formulation of the integro-differential equation for coagulation, growth, and nucleation processes. More fundamental studies on the feasibility of variational aerosol data assimilation are, for example, presented in Henze et al. (2004). While in Collins et al. (2001) emphasis is placed on modelling a more sophisticated aerosol optical depth using state-of-the-art modules, the OI assimilation scheme applied satisfies the BLUE property only when used as a purely spatial algorithm.

Data assimilation has also been extended to inverse modelling of biomass burning emissions. Zhang et al. (2005), using a Bayesian inversion technique, found special sensitivity of the results to a priori emissions and to the altitude of the aerosol layer. In a 4D-Var context, Benedetti and Fisher (2007) introduced the NMC (National Meteorological Center) method to assess the background error statistics of global aerosol distributions for earth system monitoring, using both satellite aerosol retrievals and in situ data. A similar operational variational AOD data assimilation system is described by Zhang et al. (2008). Yumimoto et al. (2008) coupled a 4D-Var data assimilation system to the regional dust model RAMS/CFORS-4DVAR to carry out an adjoint inversion of a heavy dust event over eastern Asia; the vertical profiles of the dust extinction coefficients derived from a Lidar network were directly assimilated. The authors demonstrated significant improvements for dust emission inversion.

A comprehensive review of the emerging field of advanced chemical data assimilation can be found in Carmichael et al. (2007).

## 3 Advanced Methods in Tropospheric Chemistry Data Assimilation

The tropospheric chemistry data assimilation (TCDA) problem described above has implications for the data assimilation methodology to be selected. Chemistry-transport models are typically forced by 3-D meteorological analyses and, unless care is taken, chemical or other imbalances can result. Detrimental effects include the generation of spurious relaxations toward a chemical state which is no longer subject to the constraint of an objective quality criterion. To solve this problem, the chemical kinetic equations of the model can, as part of the assimilation algorithm, be used as a constraint to estimate both a balanced and most probable chemical state or, analogously, parameter values. In this way, the system at least potentially satisfies the BLUE property. An advantage of this approach is that the BLUE property allows for hypothesis testing.

There are two families of algorithms satisfying the BLUE property in a spatio-temporal context: the four-dimensional variational data assimilation (4D-Var), and Kalman filtering (KF). In both cases, the methodology involves the extension of the state vector to include sources and their uncertainties.

### *3.1 Kalman Filter Equations*

The Kalman filter method is based on the Kalman filter equations. The forecast equation propagating the model state $\mathbf{x}^f$ from time $i-1$ to $i$ by the model resolvent or linear integration operator $\mathbf{M}(t_i, t_{i-1})$ reads

$$\mathbf{x}^f(t_i) = \mathbf{M}(t_i, t_{i-1})\mathbf{x}^a(t_{i-1}), \tag{1}$$

where superscripts $a$ and $f$ indicate analysis and forecast, respectively. Adopting the standard notation, the forecast error covariance matrix $\mathbf{P}_i^f$ at time step $i$ is estimated by

$$\mathbf{P}_i^f = \mathbf{M}(t_i, t_{i-1})\mathbf{P}_i^a\mathbf{M}^T(t_i, t_{i-1}) + \mathbf{Q}, \tag{2}$$

involving the analysis error covariance matrix $\mathbf{P}_i^a$ and model error covariance matrix $\mathbf{Q}$. Given the vector of observations $\mathbf{y}_i$, and the forecast model equivalent obtained by the linear observation operator $\mathbf{H}$ applied to the forecast state $\mathbf{x}^f(t_i)$, the optimal estimate of the state

$$\mathbf{x}^a(t_i) = \mathbf{x}^f(t_i) + \mathbf{K}_i(\mathbf{y}_i - \mathbf{H}\mathbf{x}^f(t_i)), \tag{3}$$

is computed using the Kalman gain matrix

$$\mathbf{K}_i := \mathbf{P}_i^b\mathbf{H}_i^T(\mathbf{H}_i\mathbf{P}_i^b\mathbf{H}_i^T + \mathbf{R}_i)^{-1} \in \mathcal{R}^{n\times p_i}. \tag{4}$$

The analysis error covariance matrix $\mathbf{P}_i^a$ is given by

$$\mathbf{P}_i^a = (\mathbf{I} - \mathbf{K}_i\mathbf{H})\mathbf{P}_i^b. \tag{5}$$

The Kalman filter equations given above are computationally too costly, in particular (2). Thus, for practical problems, the Ensemble Kalman Filter (EnKF) and the Reduced Rank Square Root Kalman filter (RRSRKF) are used, as they allow for a feasible spatio-temporal data assimilation approach, while approximating the BLUE property. Hanea et al. (2004) successfully implemented both methods for tropospheric chemistry data assimilation, and Constantinescu et al. (2007a, b) the EnKF.

## *3.2 Ensemble Kalman Filter*

The practical realisation of the EnKF is described in Hamill (2006) and is given here for completeness. An ensemble matrix $\mathbf{X}^f$ is composed of ensemble members

$$\mathbf{X}^f := (\mathbf{x}_1^b, \ldots, \mathbf{x}_m^b), \tag{6}$$

with ensemble mean

$$\bar{\mathbf{x}}^f := \frac{1}{m}\sum_{i=1}^{m}\mathbf{x}_i^b. \tag{7}$$

Introducing the perturbation of the *ith* member $\mathbf{x}_i'^b := \mathbf{x}_i^b - \bar{\mathbf{x}}^{\mathbf{f}}$, the matrix of the ensemble perturbations then reads

$$\mathbf{X}'^f := (\mathbf{x}_1'^b, \ldots, \mathbf{x}_m'^b) \tag{8}$$

Let $\tilde{\mathbf{P}}^f$ denote the ensemble estimate of the forecast error covariance matrix $\mathbf{P}^f$. Then, this is calculated by

$$\tilde{\mathbf{P}}^f = \frac{1}{m-1}\mathbf{X}'^f\mathbf{X}'^{bT}. \tag{9}$$

For the stochastic update algorithm

$$\mathbf{x}_i^a = \mathbf{x}_i^b + \mathbf{K}_i(\mathbf{y}_i - \mathcal{H}(\mathbf{x}_i^b)), \tag{10}$$

we generate $m$ sets of "perturbed observations"

$$\mathbf{y}_i = \mathbf{y} + \mathbf{y}_i', \quad i = 1, \ldots, m, \tag{11}$$

where the $\mathbf{y}_i' \propto \mathcal{N}(0, \mathbf{R})$ have a Gaussian error distribution.

With $\mathcal{H}$ being the non-linear observation operator, the elements of the Kalman gain matrix can then be calculated as follows

$$\overline{\mathcal{H}(\mathbf{x}^f)} := \frac{1}{m}\sum_{i=1}^{m}\mathcal{H}(\mathbf{x}_i^b) \tag{12}$$

$$\tilde{\mathbf{P}}^f\mathbf{H}^T := \frac{1}{m-1}\sum_{i=1}^{m}(\mathbf{x}_i^b - \bar{\mathbf{x}}^f)(\mathcal{H}(\mathbf{x}_i^b) - \overline{\mathcal{H}(\mathbf{x}^f)})^T \tag{13}$$

and

$$\mathbf{H}\tilde{\mathbf{P}}^f\mathbf{H}^T := \frac{1}{m-1}\sum_{i=1}^{m}(\mathcal{H}(\mathbf{x}_i^b) - \overline{\mathcal{H}(\mathbf{x}^f)})(\mathcal{H}(\mathbf{x}_i^b) - \overline{\mathcal{H}(\mathbf{x}^f)})^T. \tag{14}$$

We thus obtain for the analysis mean

$$\bar{\mathbf{x}}^a(t_{i+1}) = \bar{\mathbf{x}}^f(t_{i+1}) + \mathbf{K}_{i+1}(\mathbf{y} - \mathcal{H}(\bar{\mathbf{x}}_i^b)), \tag{15}$$

and for the individual perturbations

$$\mathbf{x}'^a_{(t+1)} = \mathbf{x}'^b_{(t+1)} + \mathbf{K}_{t+1}\mathcal{H}(\mathbf{x}_i'^b), \tag{16}$$

from which the analysis error covariance matrix $\mathbf{P}^a$ can be calculated analogously to (9).

### *3.3 Reduced Rank Square Root Kalman Filter*

The approach of Hanea et al. (2004) approximates the $n \times n$ covariance matrices $\mathbf{P}^{f,a}$ by a product of suitably selected $n \times q,\ q \ll n$, low ranked matrix $\mathbf{S}^{f,a}$. The $q$ leading eigenvectors are the basis for the determination of $\mathbf{S}$. With the eigenvector decomposition $\mathbf{P}^{f,a} = \mathbf{V}^{f,a}(\mathbf{D}^{f,a})(\mathbf{V}^{f,a})^T$, $\mathbf{D}$ denoting the diagonal matrix of eigenvalues, matrix $\mathbf{S}^{f,a} = \mathbf{V}^{f,a}(\mathbf{D}^{f,a})^{1/2}$ complies with this requirement. The same procedure is applied to the system noise matrix $\mathbf{Q}$, factorised by $\mathbf{T}$ with $n \times r,\ r \ll n$. Hence, we have

$$\mathbf{P} \approx \mathbf{S}\mathbf{S}^T, \qquad \mathbf{Q} \approx \mathbf{T}\mathbf{T}^T. \tag{17}$$

The forecast step (3) remains unchanged. However, the calculation of the forecast error covariance matrix only uses $2 \times p$ model integrations:

$$\mathbf{S}^f\mathbf{S}^{fT} = \mathbf{M}\mathbf{S}^a\mathbf{S}^{aT}\mathbf{M}^T + \mathbf{T}\mathbf{T}^T. \tag{18}$$

With model errors, we have $\mathbf{S}^f = [\mathbf{M}\mathbf{S}^a, \mathbf{T}]$, and using the definition $\psi := \mathbf{H}\mathbf{S}^f$, the Kalman gain matrix reads

$$\mathbf{K} = \mathbf{S}^f \psi^T (\psi \psi^T + \mathbf{R})^{-1}.$$

The analysis error covariance matrix can then be rewritten

$$\begin{aligned} \mathbf{S}^a \mathbf{S}^{aT} &= (\mathbf{I} - \mathbf{KH}) \mathbf{S}^f \mathbf{S}^{fT} \\ &= \mathbf{S}^f [\mathbf{I} - \psi^T (\psi \psi^T + \mathbf{R})^{-1} \psi] \mathbf{S}^{fT}. \end{aligned} \tag{19}$$

In practice, all calculations can be performed without actually calculating the full matrices **P**, and only the square root representation is needed and supported by the algorithm. Hence, the positive semi-definiteness of the covariance matrices is maintained. This is implemented by the square root form

$$\mathbf{S}^a = \mathbf{S}^f [\mathbf{I} - \psi^T (\psi \psi^T + \mathbf{R})^{-1} \psi]^{1/2}. \tag{20}$$

Several methods exist to carry out the measurement updates. Hanea et al. (2004) adopted a *scalar update* formalism, where each measurement is processed individually. Defining $\gamma := (\Psi^T \Psi + \mathbf{R})^{-1}$ for each time step, the above formula can be rearranged to give

$$\mathbf{S}^a = \mathbf{S}^f - \mathbf{K} \psi \left[ 1 + (\gamma \mathbf{R})^{1/2} \right]^{-1}. \tag{21}$$

### *3.4 4D Variational Data Assimilation*

In the case of 4D-Var, examples of a spatio-temporal BLUE applied in tropospheric chemistry include Elbern et al. (2007), with the EURAD–IM (EURopean Air pollution Dispersion–Inverse Model).

The most notable aspect in this implementation is the additional inversion for emission rate optimization and for non-observed species. Here, deviations of the background chemical state $\mathbf{x}(t_0) - \mathbf{x}_b = \delta \mathbf{x}(t_0)$ and the emission inventory $\mathbf{e}(t_0) - \mathbf{e}_b = \delta \mathbf{e}(t_0)$ may be combined to define an incremental formulation of a cost function, objective function, or distance function $J$ as follows (see for example Elbern et al. 2000 for a more detailed description):

$$\begin{aligned} J(\delta \mathbf{x}(t_0), \delta \mathbf{e}) = &\frac{1}{2} (\delta \mathbf{x})^T \mathbf{B}^{-1} \delta \mathbf{x} + \frac{1}{2} \int_{t_0}^{t_N} (\delta \mathbf{e})^T \mathbf{K}^{-1} \delta \mathbf{e} dt + \\ &\frac{1}{2} \int_{t_0}^{t_N} (\mathbf{d}(t) - \mathbf{H}(t) \delta \mathbf{x}(t))^T \mathbf{R}^{-1} (\mathbf{d}(t) - \mathbf{H}(t) \delta \mathbf{x}(t)) dt \end{aligned} \tag{22}$$

where $J$ is a scalar functional defined on the time interval $t_0 \leq t \leq t_N$, and dependent on the vector valued state variable $\mathbf{x}(t)$. $\mathbf{d}(t) := \mathbf{y}(t) - \mathbf{H}(t) \delta \mathbf{x}_b(t)$ is the observation minus model discrepancy at time $t$, when the first guess initial values and emission inventory values are taken. The error covariance matrices are defined as follows:

for the first guess or background values $\mathbf{B} \in \mathbb{R}^{N\times N}$ with $N$ the number of model variables; for the emission factors $\mathbf{K} \in \mathbb{R}^{E\times E}$ with $E$ the number of emitting grid points times the emitted species. Observation errors are denoted $\mathbf{R} \in \mathbb{R}^{M(t)\times M(t)}$, with $M(t)$ the number of available observations at time $t$. Operator $\mathbf{H}(t)$ calculates the model equivalent for each observation.

We want to determine the gradient of $J$ with respect to the joint chemical state and emission rate variable $\mathbf{z} = (\delta\mathbf{x}, \delta\mathbf{e})^T$, and compute the gradient $\partial J/\partial(\delta\mathbf{x}, \delta\mathbf{e})^T$. The gradient of the cost function $J$ is given by

$$\begin{aligned}\partial J/\partial(\delta\mathbf{x}, \delta\mathbf{e})^T = &- \mathbf{B}^{-1}(\delta\mathbf{x}(t)) - \sum_{t_0}^{t_N} \mathbf{M}^T(t_0, t)\mathbf{H}^T(t)\mathbf{R}^{-1}(\mathbf{d} - \mathbf{H}(t)\delta\mathbf{x}(t)) \\ &- \sum_{t_0}^{t_N} \mathbf{K}^{-1}(\mathbf{e}_b(t) - \mathbf{e}(t)),\end{aligned} \tag{23}$$

where $\mathbf{M}^T(t_0, t)$ denotes the adjoint (= transposed, $T$) model operator, formally integrating from time $t$ backwards in time to the initial time $t_0$. With the square root factorizations $\mathbf{B} = \mathbf{B}^{1/2}(\mathbf{B}^{1/2})^T$ and $\mathbf{K} = \mathbf{K}^{1/2}(\mathbf{K}^{1/2})^T$ we define new variables $v$ and $w$ by (where $u$ is the amplitude of the emissions $\mathbf{e}$)

$$v := \mathbf{B}^{-1/2}\delta x, \qquad w := \mathbf{K}^{-1/2}\delta u, \tag{24}$$

leading to a minimization problem equivalent to Eq. (22). The cost function is then given by

$$\begin{aligned}J(v, w) = &\frac{1}{2}v^T v + \frac{1}{2}w^T w + \\ &\frac{1}{2}\sum_{i=0}^{T}(d_i - \mathbf{H}\delta x_i)^T\, \mathbf{R}^{-1}\, (d_i - \mathbf{H}\delta x_i).\end{aligned} \tag{25}$$

The gradient of $J$ with respect to $(\mathbf{v}, \mathbf{w})^T$ can be shown to be

$$\begin{aligned}\nabla_{(\mathbf{v},\mathbf{w})^T} J = &- \begin{pmatrix}\mathbf{v}\\ \mathbf{w}\end{pmatrix} - \begin{pmatrix}\mathbf{B}^{1/2} & \mathbf{0}\\ \mathbf{0} & \mathbf{K}^{1/2}\end{pmatrix} \times \\ &\sum_{m=0}^{T} \mathbf{M}^T(t_0, t_m)\mathbf{H}^T\mathbf{R}^{-1}(\mathbf{d}(t_m) - \mathbf{H}\boldsymbol{\delta}\mathbf{x}(t_m)),\end{aligned} \tag{26}$$

This optimization problem can be solved by a quasi-Newton minimisation procedure, for example L–BFGS.

### 3.5 Implementation of a Chemical 4D-Var System

In the EURAD–IM, a comprehensive tropospheric Eulerian model operating on continental to local scales, the CTM calculates the transport, diffusion, and gas phase transformations of about 60 chemical species with more than 150 reactions. For a CTM, the differential equation can be written as:

$$\frac{\partial c_i}{\partial t} + \nabla \cdot (\mathbf{v} c_i) - \nabla \cdot \left( \rho \mathbf{G} \nabla \frac{c_i}{\rho} \right) - \sum_{r=1}^{R} \left( k(r)(s_i(r_+) - s_i(r_-)) \prod_{j=1}^{U} c_j^{s_j(r_-)} \right) = E_i + D_i \tag{27}$$

where $c_i$ is the concentration of species $i$, $\mathbf{v}$ is the wind velocity, $s \in \mathrm{N}_0$ is the stoichiometric coefficient, *k(r)* is the reaction rate of reaction $r$, either productive ($r_+$) or destructive ($r_-$) for species $i$, $U$ is the number of species in the mechanism, $E_i$ is the emission rate of species $i$, $D_i$ is the deposition rate of species $i$, the air density is denoted by $\rho$, and $\mathbf{G}$ is the symmetric eddy diffusivity tensor.

After application of the variational calculus, the adjoint formulation of (27) reads

$$-\frac{\partial \delta c_i^*}{\partial t} - \mathbf{v} \nabla \delta c_i^* - \frac{1}{\rho} \nabla \cdot (\rho \mathbf{K} \nabla \delta c_i^*) + \sum\nolimits_{r=1}^{R} \left( k(r) \frac{s_i(r_-)}{c_i} \prod\nolimits_{j=1}^{U} c_j^{s_j(r_-)} \sum\nolimits_{n=1}^{U} (s_n(r_+) - s_n(r_-)) \delta c_n^* \right) = 0 \tag{28}$$

with $\delta c_i^*$ the adjoint variable of $c_i$, while $D_i$ is held fixed.

The variational chemistry data assimilation algorithm has four components:

1. The forward model;
2. The adjoint of its tangent-linear form;
3. The background error covariance matrices;
4. The minimization routine.

## 4 Examples

The 4D-Var technique allows for the assimilation of a wide variety of data types. A careful estimation of the error of (spatial) representativity is however a prerequisite for success. Specifically, model grid resolutions of about 50 km, widely used for continental scale integration domains, admit only a limited number of species to be assimilated using point measurements. For example, quickly oxidizing point and line sources of emitted $\mathrm{NO}_x \in \{\mathrm{NO}, \mathrm{NO}_2\}$ should only be assimilated by observa-

tion sites situated at background locations, which are rarely available. In practice, gaseous constituent assimilation using coarse grid models mostly applies to ozone observations.

## *4.1 Nested Application of 4D-Var*

In order to also exploit measurements of $NO_x$ species, a nesting technique is implemented for adjoint modelling and applied to the 1998 urban plume campaign BERLIOZ (Volz-Thomas et al. 2003) around the metropolitan area of Berlin, Germany. Assimilated species are $O_3$, NO, $NO_2$, CO and $SO_2$ within an assimilation window of 14 h, from 0600 to 2000 UTC on July 20, 1998. The nesting procedure includes a coarse grid simulation with horizontal grid size of 54 km and three recursively nested grids with a nesting ratio of three. Hence, there is a 2 km final resolution. Figure 3 demonstrates the assimilation performance for 2 measurement stations found within the greater Berlin area, as achieved by the 6 km resolution grid (nested level 2) using analyses from a joint emission rate and initial value optimization (see next section). When using an analysis from a nested 4D-Var, a significant improvement in the forecast can be seen beyond the assimilation interval. Thus, it can be concluded that, for the conditions studied, a 6 km horizontal resolution allows for satisfactory exploitation of the suburban $NO_x$ observation sites.

Figure 4 demonstrates the improvement achieved through a longer-running (10 days) 4D-Var nesting application for a suite of observed species, in this case for the VERTIKO campaign in June 2003 (Bernhofer and Köstner 2005). The simulation set-up has three domain levels with a coarse grid resolution of 125 km and a nesting ratio of 5. Improvements are given in cost function values normalized by the control run values at each nested level, and averaged over the whole simulation period.

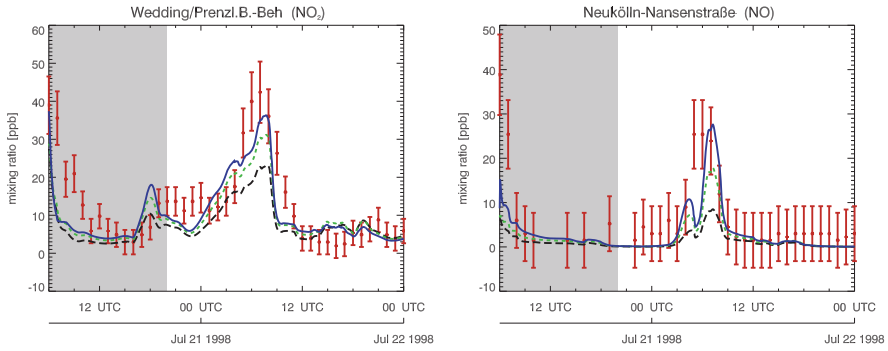



**Fig. 3** Assimilation results for stations in the Berlin area obtained with a grid resolution of 6 km. (*left panel*: Wedding/Prenzl.B-Beh, $NO_2$; *right panel*: Neukölln-Nansenstraße, NO) *Green line*: first guess run, using an analysis obtained on a 18 km grid. *Blue line*: assimilation result based on an analysis on a 6 km grid. *Black line*: results for no data assimilation. Observations are given in red, and their error estimates as *vertical red bars*. Only the *grey shaded* time interval has been used for the assimilation; other observations are only used for quality control

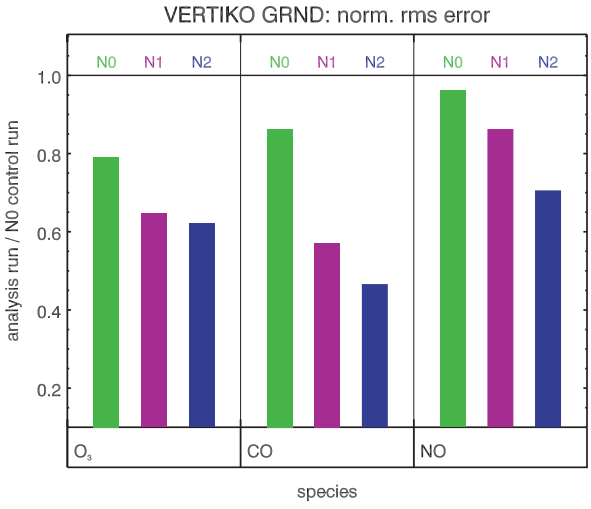



**Fig. 4** Relative reduction of the root-mean-square errors for ozone, carbon monoxide and nitrogen oxide due to nested grids (N0 to N2) with increasing horizontal resolution (from 125 to 5 km). A normalization value of 1 is given for the coarse grid (125 km) simulations without data assimilation

Ozone, carbon and nitrogen oxide exhibit different levels of performance at the different nesting levels. Ozone analyses are already quite reasonable at coarser grids (25 km), reflecting the smoother distribution pattern of a secondary pollutant. In contrast, NO shows improved performance as one goes from a coarser to a finer grid resolution. Even for a grid resolution of 5 km, the source distribution and relatively short lifetime of NO cannot be completely represented (see previous paragraph); this requires additional nesting levels. Carbon monoxide exhibits a behaviour which mixes elements of that of ozone and NO, reflecting a primarily emitted constituent with a relatively longer lifetime in the troposphere (compared to NO).

## *4.2 Emission Rate Estimates*

As a unique feature, the adjoint calculus has the potential to optimize initial values as well as emission rates. The impact of emission rate optimization is demonstrated by Figure 5, which shows $SO_2$, CO, $NO_2$, and xylene optimization factors over the integration domain of the finest BERLIOZ grid (2 km resolution). The inversion process at each grid level passes the result to the next finer grid, allowing for an increasingly better resolved emission estimate, provided the necessary observational density is achieved.

As Berlin is mostly a large urban island within a more rural environment, sulphur emissions are confined to the greater metropolitan area. The upper left panel of Figure 5 clearly indicates a slightly reduced emission rate over the densely populated areas, likely indicating moderately larger success in reduction efforts than estimated by the emission inventory. In the case of CO, similar effects can only be claimed for the area east of Berlin.

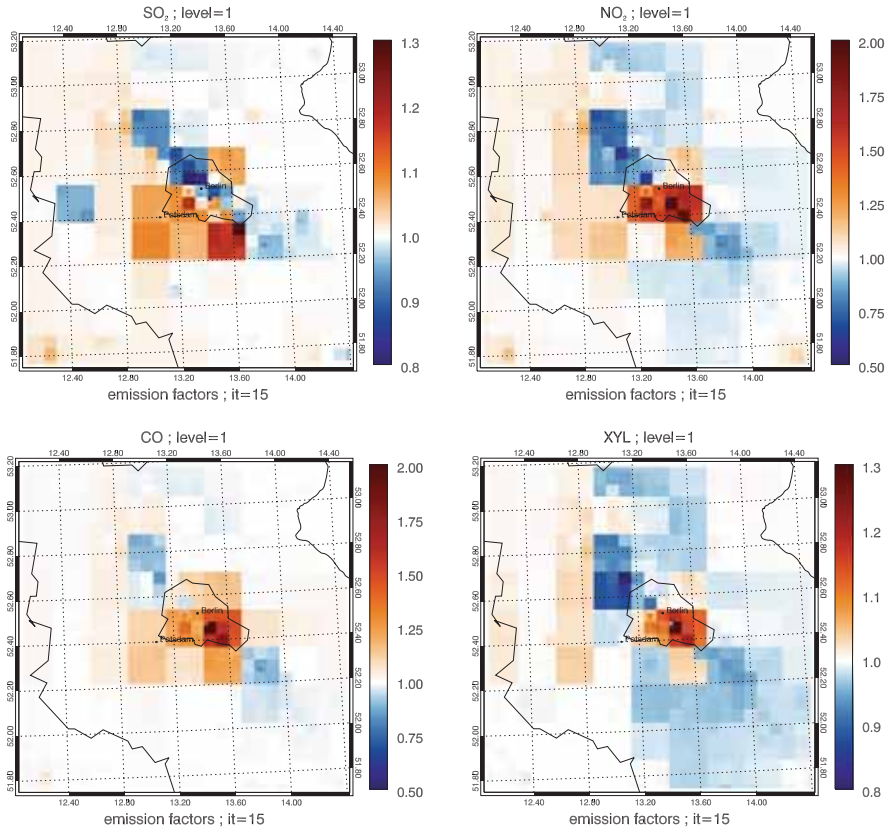



**Fig. 5** $SO_2$ (*upper left panel*), CO (*lower left panel*), $NO_2$ (*upper right panel*), and xylene (*lower right panel*) optimization factors over the integration domain of the finest grid (2 km resolution) over the greater Berlin area. The factors are for the surface layer. *Coloured squares* indicate the impact areas of the individual nested areas, ranging from 54 to 2 km resolution, involving 4 nested levels

While emission rates do not vary much for $NO_2$, xylene appears to be underestimated by the emission inventory, with an amplification factor of about 1.2. In all exhibited cases, the inversion results remain well within the error bars of the inventory. Emission rate optimization of $SO_2$ and CO is mainly based on observations of concentrations of these species. In the case of other emissions, which are rarely observed, inference can only be made based on measurements of other observed species, mainly ozone. This capability to infer information on non-observed species from observed species is another prominent feature of the 4D-Var technique. In this context it should be noted that short campaigns like BERLIOZ may be insufficient to build up reliable quantitative statistics for emission inversion, as underlying error covariance statistics need longer estimation times; these estimation times, in turn, depend on meteorological conditions.

## 4.3 Tropospheric Satellite Data Assimilation

Satellite retrievals from tropospheric height levels are an emerging issue in Earth Observation, although there is a limited number of species like $SO_2$, $NO_2$ and formaldehyde, which are presently amenable for retrieval. Moreover, in these cases, data are presented in terms of tropospheric columns. See also chapter *Constituent Assimilation* (Lahoz and Errera).

The conceptual flexibility of the variational technique must be invoked where data assimilated, like tropospheric columns, have no direct correspondence to a model parameter. In the case of tropospheric columns, data are given in terms of molecules per $cm^2$. The model correspondence (operator **H** in the cost function) can then be calculated, along with its adjoint, and included in the algorithm. Making use of the technique of preconditioning (see Elbern et al. 2007), the optimization

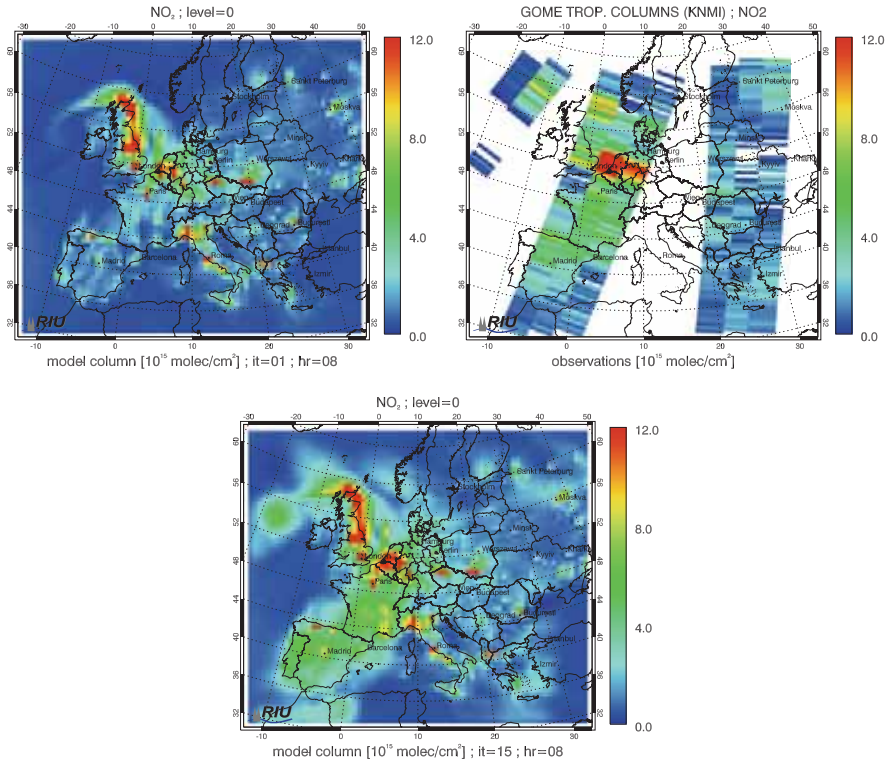



**Fig. 6** $NO_2$ tropospheric column assimilation of GOME using averaging kernels for July 20, 1998. *Upper panel* shows first guess (*left*) and retrieved tropospheric columns by KNMI, the Royal Dutch Meteorological Institute (*right*); *lower panel* gives the model equivalent after 4D-Var assimilation. Both model equivalent fields (*upper left panel* and *lower panel*) are at 0800 UTC; optimization of initial values is at 0600 UTC. Units are given in terms of $10^{15}$ molecules/$cm^2$

procedure of the assimilation can adapt the model column to the retrieval in a manner consistent with the model and the specified error covariances. Figure 6 gives an example of the assimilation of $NO_2$ tropospheric columns obtained from KNMI (Eskes and Boersma 2003). Spain, France, and the Ruhr area are areas with discrepancies between the retrievals and the model, prior to assimilation. The final analysis largely removes these discrepancies, which can be attributed to the model. However, in situ observations of $NO_2$ are included as well, and this explains the elevated levels found in the analysis over the UK.

## *4.4 Aerosol Assimilation*

Key components of the EURAD–IM model system are inorganic and secondary organic aerosol modules. The Modal Aerosol Dynamics model for Europe (MADE) has been developed for the EURAD model as an extension to the EURAD CTM to allow for a detailed treatment of aerosol effects in the model. Due to the complexity of the atmospheric aerosol system, an approach has been chosen which is fast

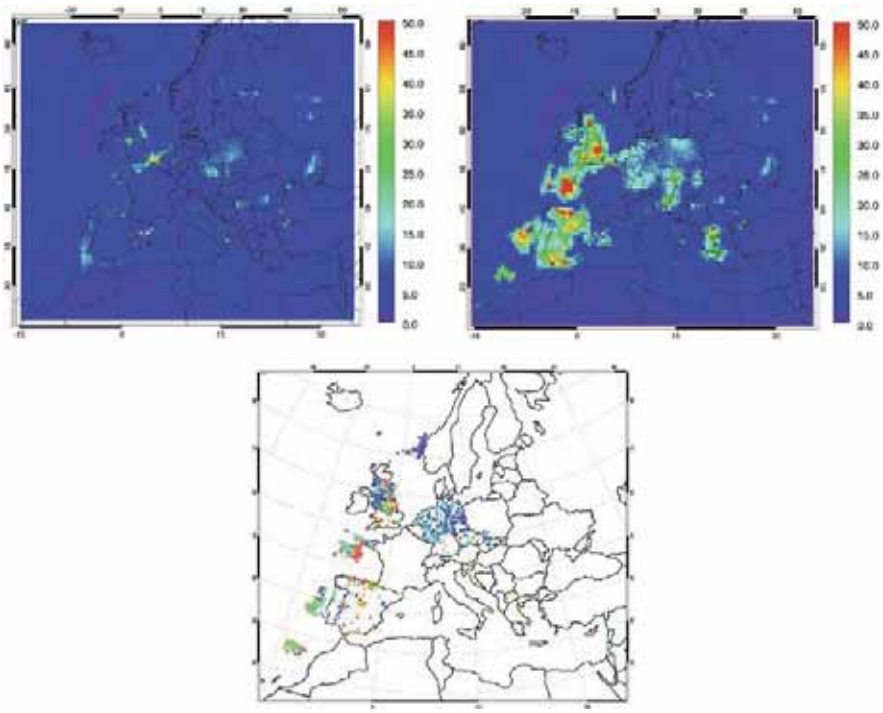


**Fig. 7** $PM_{10}$ concentrations [$\mu g/m^3$] at July 13, 2003, 1100 UTC, as forecast without assimilation (*upper left*). Available observations, both in situ and SYNAER retrievals, are shown in the *lower panel*; assimilation results based on all observations are shown *upper right*

enough for this application and also provides sufficient information on the particle size distribution. In MADE, the particle size distribution of the aerosol is represented by three overlapping lognormal modes: the Aitken mode, the accumulation mode and the coarse mode; all cases are provided with variable mean values. Due to the complexity of the aerosol model, no full adjoint is available yet. Therefore, a 3D-Var algorithm has been implemented, which only involves the adjoint of the observation operator (Nieradzik and Elbern 2006).

The analysis presented makes use of SYNAER retrievals (Holzer-Popp et al. 2002), providing for two ENVISAT sensor based particulate matter retrievals. This allows for freely adjustable integrated aerosol sizes, and provides size distributed aerosol mass information. On July 13, 2003, when a SCIAMACHY footprint covered the United Kingdom, there was a comparably dense aerosol measurement network for comparison. The assimilation procedure has been conducted with available ground based in situ data; only with satellite data; and with both observational sets combined. Figure 7 shows that the model forecast (without data assimilation) gives too low aerosol loads. Both the satellite and situ data enforce higher (and more realistic) values in the analyses.

## 5 Outlook

Both, Kalman filtering and the 4D-Var data assimilation method prove useful for applications in air quality simulations. Looking for analogies with meteorological weather forecasting, the problems for air quality applications resemble the challenges associated with low level humidity assimilation, where the following issues must be addressed: markedly non-isotropic and inhomogeneous correlation lengths, especially in the boundary layer; frequent violation of the assumption of the tangent linear approximation; and significant violation of the perfect model assumption due to deficiencies in the knowledge of near ground highly resolved meteorological parameters. In addition, emission rates are at least as important as initial values, and should be included as an optimization parameter.

For many regions, deposition rates should also be included as an optimization parameter. With an augmented set of optimization parameters, the optimization problem becomes more ill-posed and enhanced precision of estimates of the error covariance matrices is the only possible way to address this problem. The most obvious way to solve this issue would be to perform operational runs of the assimilation system, which would allow compilation of the relevant statistics.

A further fundamental issue are the covariances of the chemical species involved in the models. Only when these are introduced comprehensively to the data assimilation algorithm will non-observed species be able to be analysed from chemically coupled observed species. While Kalman filtering is theoretically able to provide information on non-observed species, observations are too sparse and the diurnal cycle too short for a Kalman filter to spin-up to useful levels of skill. Finally, we note that a priori information is useful to the data assimilation problem, as it is available for updating.

**Acknowledgments** The authors are indebted to the BERLIOZ and VERTIKO project members for measurement data, and to Dr. A. Richter (IFE University of Bremen) and Dr. H. Eskes (KNMI) for satellite retrievals. The work was mainly supported from the German Ministry for Research and Technology in the frame of the AFO2000 project SATEC4D. Computing facilities were granted by ZAM, the Research Centre Jülich, on a Cray T3E and IBM Power 4.

## References

Benedetti, A. and M. Fisher, 2007. Background error statistics for aerosols. *Q. J. R. Meteorol. Soc.*, **133**, 391–405.

Bernhofer, C. and B. Köstner (ed.), 2005. *Vertical Transport of Energy and Trace Gases at Anchor Stations and Their Spatial/Temporal Extrapolation under Complex Natural Conditions (VERTIKO): Project Summary*, Vol 12, Tharandter Klimaprotokolle, Technische Universität Dresden.

Bocquet, M., 2005a. Reconstruction of an atmospheric tracer source using the principle of maximum entropy. I: Theory. *Q. J. R. Meteorol. Soc.*, **131**, 2191–2208.

Bocquet, M., 2005b. Reconstruction of an atmospheric tracer source using the principle of maximum entropy. II: Applications. *Q. J. R. Meteorol. Soc.*, **131**, 2209–2223.

Bousquet, P., P. Ciais, P. Peylin, M. Ramonet and P. Monfray, 1999a. Inverse modeling of annual CO2 sources and sinks, 1. Method and control inversion. *J. Geophys. Res.*, **104**, 26161–26178.

Bousquet, P., P. Ciais, P. Peylin, M. Ramonet and P. Monfray, 1999b. Inverse modeling of annual CO2 sources and sinks, 2. Sensitivity study. *J. Geophys. Res.*, **104**, 26179–26193.

Carmichael, G.R. et al., 2003. Regional-scale chemical transport modeling in support of the analysis of observations obtained during the TRACE-P experiment. *J. Geophys. Res.*, **108**, 8823.

Carmichael, G.R., A. Sandu, T. Chai, D.N. Daescu, E.M. Constantinescu and Y. Tang, 2007. Predicting air quality: Improvements through advanced methods to integrate models and measurements. *J. Comp. Phys.*, **227**, 3540–3571.

Chai, T.F., G.R. Carmichael, A. Sandu, Y.H. Tang and D.N. Daescu, 2006. Chemical data assimilation of transport and chemical evolution over the Pacific (TRACE-P) aircraft measurements. *J. Geophys. Res.*, **111**, doi:10.1029/2005JD005883.

Collins, W., P. Rasch, B. Eaton, B. Khattatov, J.-F. Lamarque, and C. Zender, 2001. Simulating aerosols using a chemical transport model with assimilation of satellite aerosol retrievals: Methodology for INDOEX. *J. Geophys. Res.*, **106**, 7313–7336.

Constantinescu, E.M., A. Sandu, T. Chai and G.R. Carmichael, 2007a. Ensemble based chemical data assimilation. I: General approach. *Q. J. R. Meteorol. Soc.*, **133**, 1229–1243.

Constantinescu, E.M., A. Sandu, T. Chai and G.R. Carmichael, 2007b. Ensemble based chemical data assimilation. II: Covariance localization. *Q. J. R. Meteorol. Soc.*, **133**, 1245–1256.

Deeter, M.N. et al., 2003. Operational carbon monoxide retrieval algorithm and selected results for the MOPITT instrument. *J. Geophys. Res.*, **108**, 4399.

Elbern, H. and H. Schmidt, 2001. Ozone episode analysis by four-dimensional variational chemistry data assimilation. *J. Geophys. Res.*, **D4**, 3569–3590.

Elbern, H., H. Schmidt, O. Talagrand and A. Ebel, 2000. 4D.variational data assimilation with an adjoint air quality model for emission analysis. *Environ. Model Software*, **15**, 539–548.

Elbern, H., A. Strunk, H. Schmidt and O. Talagrand, 2007. Emission rate and chemical state estimation by 4-dimensional variational inversion. *Atmos. Chem. Phys.*, **7**, 1–59.

Engelen, R., E. Andersson, F. Chevallier, A. Hollingsworth, M. Matricardi, A.P. McNally, J.N. Thépaut and P.D. Watts, 2004. Estimating atmospheric CO2 from advanced infrared satellite radiances within an operational 4D-Var data assimilation system: Methodology and first results. *J. Geophys. Res.*, **109**, D19309, doi:10.1029/2004JD004777.

Enting, I.G. and G.N. Newsam, 1990. Inverse problems in atmospheric constituent studies: II. Sources in the free atmosphere. *Inverse Prob.*, **6**, 349–362.

Enting, I.G., C.M. Trudinger and R.J. Francey, 1995. A synthesis inversion of the concentration and d13C of atmospheric CO2. *Tellus*, **47B**, 35–52.

Eskes, H.J. and K.F. Boersma, 2003. Averaging kernels for DOAS total-column satellite retrievals. *Atmos. Chem. Phys.*, **3**, 1285–1291.

Fan, S.M., M. Gloor, J. Mahlman, S. Pacala, J. Sarmiento, T. Takahashi and P. Tans, 1998. A large terrestrial carbon sink in North America implied by atmospheric data and oceanic carbon dioxide data and models. *Science*, **282**, 442–446.

Fedorov, V., 1998. Kriging and other estimators of spatial field characteristics (with special reference to environmental studies). *Atmos. Environ.*, **23**, 174–184.

Gloor, M., S.M. Fan, S. Pacala, J. Sarmiento and M. Ramonet, 1999. A model-based evaluation of inversions of atmospheric transport, using annual mean mixing ratios, as a tool to monitor fluxes of nonreactive trace substance like $CO_2$ on a continental scale. *J. Geophys. Res.*, **104**, 14245–14260.

Gurney, K.R., 2002. Towards robust regional estimates of $CO_2$ sources and sinks using atmospheric transport models. *Nature*, **415**, 626–630.

Hamill, T.M., 2006. Ensemble-based atmospheric data assimilation. In *Predictability of Weather and Climate*, Palmer, T.N. and R. Hagedorn (eds.), Cambridge University Press, Cambridge, pp. 124–156.

Hanea, R.G., G.J.M. Velders and A. Heemink, 2004. Data assimilation of ground-level ozone in Europe with a Kalman filter and chemistry transport. *J. Geophys. Res.*, **109**, D10302, doi:10.1029/2003JD004283.

Hanna, S.R., J.C. Chang and M.E. Fernau, 1998. Monte Carlo estimates of uncertainties in predictions by a photochemical grid model (UAM-IV) due to uncertainties in input variables. *Atmos. Environ.*, **32**, 3619–3628.

Hanna, S.R., Z.G. Lu, H.C. Frey, N. Wheeler, J. Vukovich, S. Arunachalam, M. Fernau and D.A. Hansen, 2001. Uncertainties in predicted ozone concentrations due to input uncertainties for the UAM-V photochemical grid model applied to the July1995 OTAG domain. *Atmos. Environ.*, **35**, 891–903.

Henze, D., J.H. Seinfeld, W. Liao, A. Sandu and G.R. Carmichael, 2004. Inverse modeling of aerosol dynamics: Condensational growth. *J. Geophys. Res.*, **109**, D14201, doi: 10.1029/2004JD004593.

Heue, K.P., A. Richter, M. Bruns, J.P. Burrows, C. von Friedeburg, U. Platt, I. Pundt, P. Wang and T. Wagner, 2005. Validation of SCIAMACHY tropospheric NO2-columns with AMAXDOAS measurements. *Atmos. Chem. Phys.*, **5**, 1039–1051.

Holzer-Popp, T., M. Schroedter and G. Gesell, 2002. Retrieving aerosol optical depth and type in the boundary layer over land and ocean from simultaneous GOME spectrometer and ATSR-2 radiometer measurements, 1, Method description. *J. Geophys. Res.*, **107**, 4578, doi: 10.1029/2001JD002013.

Houweling, S., T. Kaminski, F. Dentener, J. Lelieveld and M. Heimann, 1999. Inverse modeling of methane sources and sinks using the adjoint of a global transport model. *J. Geophys. Res.*, **104**, 26137–26160.

Issartel, J.P., 2003. Rebuilding sources of linear tracers after atmospheric concentration measurements. *Atmos. Chem. Phys.*, **3**, 2111–2125.

Kaminski, T., M. Heimann and R. Giering, 1999a. A coarse grid three-dimensional global inverse model of the atmospheric transport: 1. Adjoint model and Jacobian matrix. *J. Geophys. Res.*, **104**, 18535–18553.

Kaminski, T., M. Heimann and R. Giering, 1999b. A coarse grid three-dimensional global inverse model of the atmospheric transport: 2. Inversion of the transport of CO2 in the 1980s. *J. Geophys. Res.*, **104**, 18555–18581.

Kaminski, T., W. Knorr, P.J. Rayner and M. Heimann, 2002. Assimilating atmospheric data into a terrestrial biosphere model: A case study of the seasonal cycle. *Glob. Biogeochem. Cycles*, **16**, doi:10.1029/2001GB001463.

Lahoz, W.A., A.J. Geer, S. Bekki, N. Bormann, S. Ceccherini, H. Elbern, Q. Errera, H.J. Eskes, D. Fonteyn, D.R.J.B.K.S. Massart, V.-H. Peuch, S. Rharmili, M. Ridolfi, A. Segers, O. Talagrand, H. T. A. F. Vik and T. von Clarmann, 2007. The Assimilation of Envisat data (ASSET) project. *Atmos. Chem. Phys.*, **7**, 1773–1796.

Marenco, A. et al., 1998. Measurement of ozone and water vapor by Airbus in-service aircraft: The MOZAIC airborne program, An overview. *J. Geophys. Res.*, **103**, 25631–25642.

Muller, J.F. and T. Stavrakou, 2005. Inversion of CO and NOx emissions using the adjoint of the IMAGES model. *Atmos. Chem. Phys.*, **5**, 1157–1186.

Müller, M.D., A.K. Kaifel, M. Weber, S. Tellmann, J.P. Burrows and D. Loyola, 2003. Ozone profile retrievel from Global Ozone Monitoring Experiment (GOME) data using a neural network approach (Neural Network Ozone Retrieval System (NNORSY)). *J. Geophys. Res.*, **108**, 4497, doi:10.1029/2002JD002784.

Newsam, G.N. and I.G. Enting, 1988. Inverse problems in atmospheric constituent studies: I. Determination of surface sources under a diffusive transport approximation. *Inverse Prob.*, **4**, 1037–1054.

Nieradzik, L.P. and H. Elbern, 2006. Variational assimilation of combined satellite retrieved and in situ aerosol data in an advanced chemistry transport model. In *Proceedings of the ESA Atmospheric Science Conference*, ESA, ESA-ESRIN, Frascati.

Quélo, D., V. Mallet and B. Sportisse, 2005. Inverse modeling of nox emissions at regional scale over northern france. preliminary investigation of the second-order sensitivity. *J. Geophys. Res.*, **110**, D24310, doi:10.1029/2005JD006151.

Robertson, L. and J. Langner, 1992. Source function estimate by means of variational data assimilation applied to the ETEX-I tracer experiment. *Atmos. Environ.*, **32**, 4219–4225.

Sander, R., A. Kerkweg, P. Jöckel and J. Lelieveld, 2005. Technical note: The new comprehensive atmospheric chemistry module MECCA. *Atmos. Chem. Phys.*, **5**, 445–450.

Schmidt, H. and D. Martin, 2003. Adjoint sensitivity of episodic ozone in the Paris area to emissions on the continental scale. *J. Geophys. Res.*, **108**, 8561, doi:10.1029/2001JD001583.

Talagrand, O., 1998. A posteriori evaluation and veri_cation of analysis and assimilation algorithms. In *Proceedings of the Workshop on Diagnosis of Data Assimilation Systems*, European Centre for Medium-range Weather Forecasts, Reading, England, 2–4 November.

Talbot, R. et al., 2003. Reactive nitrogen in asian continental outflow over the western pacific: Results from the NASA Transport and Chemical Evolution over the Pacific (TRACE-P) airborne mission. *J. Geophys. Res.*, **108**, 8803, doi: 10.1029/2002JD003129.

Thouret, V., J. Cho, R. Newell, A. Marenco and H. Smit, 2000. General characteristics of tropospheric trace constituent layers observed in the mozaic program. *J. Geophys. Res.*, **105**, 17379–17392.

van Loon, M., P.J.H. Builtjes and A.J. Segers, 2000. Data assimilation of ozone in the atmospheric transport chemistry model LOTOS. *Environ. Model Software*, **15**, 603–609.

Verlaan, M. and A.W. Heemink, 1995. Reduced rank square root filters for large scale data assimilation problems. In **2***nd International Symposium on Assimilation of Observations in Meteorology and Oceanography,* Tokyo, Japan.

Volz-Thomas, A., H. Geiss, A. Hofzumahaus and K.H. Becker, 2003. Introduction to special section: Photochemistry experiment in BERLIOZ. *J. Geophys. Res.*, **108**, 8252, doi: 10.1029/2001JD002029.

Yumimoto, K., I. Uno, N. Sugimoto, A. Shimizu, Z. Liu and D. M. Winker, 2008. Adjoint inversion modeling of Asian dust emission using lidar observations. *Atmos. Chem. Phys.*, **8**, 2869–2884.

Zhang, S., J.E. Penner and O. Torres, 2005. Inverse modeling of biomass burning emissions using Total Ozone Mapping Spectrometer aerosol index for 1997. *J. Geophys. Res.*, **110**, D21306, doi: 10.1029/2004JD005738.

Zhang, J., J.S. Reid, D.L. Westphal, N.L. Baker and E.J. Hyer, 2008. A system for operational aerosol optical depth data assimilation over global oceans. *J. Geophys. Res.*, **113**, D10208, doi: 10.1029/2007JD009065.

# Part V
# Wider Applications

# Ocean Data Assimilation

**Keith Haines**

## 1 Introduction to the Ocean Circulation

The oceans form a key component of the Earth's weather and climate system. As well as being important to forecast in their own right to facilitate human activities, such as shipping, fishing, drilling for oil and coastline management and leisure, it is thought that an active ocean model is necessary for all atmospheric predictions on time-scales of a month and longer (Mansfield 1986). Another great challenge for oceanographers is to understand how and where the oceans are absorbing half of all the anthropogenic $CO_2$ being released (Battle et al. 2000), and whether this state of affairs will continue indefinitely. Ocean modelling and ocean data assimilation can play an important role in understanding the changing climate through the reanalysis of historical ocean data. We will return to some of these applications later in the chapter. It is not the intention here to cover all aspects of ocean data assimilation; in particular much of the theory is generic and can be found elsewhere in this book or in many good reviews (e.g. Bennett 1992; Wunsch 1996; Haines 2003a, b, c; see chapters in Part I, *Theory*). Instead we focus on particular applications and problems related to ocean data assimilation and try to give a perspective on some of the current and future challenges.

From a volumetric perspective the oceans are dominated by cold ($\sim$2°C), relatively fresh water ($\sim$34 psu, practical salinity units) whose properties are determined by near surface processes at high latitudes in the Arctic, and around the Antarctic continent. Water from these regions spreads equatorward and fills the deep ocean basins which are typically around 5 km deep. Waters from the Arctic need to pass across a number of straits and sills as they flow into the North Atlantic, and are thus modified by mixing in the process, whereas Antarctic waters are less modified and therefore tend to retain higher densities and to form the deepest "bottom waters" in the world oceans. In middle and lower latitudes a layer of warmer, slightly saltier, waters typically around 1 km deep, sits on top of these cold polar waters. These

K. Haines (✉)
Environmental Systems Science Centre, University of Reading, Reading RG6 6AL, UK
e-mail: kh@mail.nerc-essc.ac.uk

W. Lahoz et al. (eds.), *Data Assimilation*, DOI 10.1007/978-3-540-74703-1_20,

upper waters are called the "thermocline waters" and are characterized by strong temperature and density gradients. The circulation in these thermocline waters tends to be wind driven. Wind stresses encourage these warm waters to pool together into thick layers in some areas while in other areas they are much thinner. This leads to large horizontal temperature and density gradients which in turn lead to large horizontal pressure gradients and complex geostrophic current systems. This variation in thermocline thickness also represents available potential energy, and temporal variations of this potential energy storage and release can lead to low frequency variability in the ocean circulation which would be very useful to forecast. Both El Niño events in the tropical Pacific, and variations in the North Atlantic and North Pacific subtropical gyre strength, may be driven by such processes (Goddard and Philander 2000).

In mid latitudes the warm water pools tend to gather on the western sides of the ocean basins between the latitudes of the strongest westerly winds and the easterly trade wind belts in the tropics, for reasons to do with vorticity conservation (Stommel 1948). Here they form strong anticyclonic subtropical gyres with strong boundary currents flowing poleward along the coastline at their western edge. The Gulf Stream is such a typical boundary current which flows up the eastern coast of the USA. When these currents leave the coast and flow into the ocean interior, the strong front between the warm waters on the equatorward side of the current and the much colder waters poleward of them represents stored available potential energy which can be released through baroclinic instability. The Gulf Stream meanders greatly and forms cut-off rings and eddies via this instability; these may drift in the ocean for periods up to a year or more. Initializing and forecasting the formation and movement of such meanders and rings is one important application of ocean data assimilation. These rings carry water masses for large distances and the vertical motions around them and their associated fronts are important for surface nutrient supply and hence ocean biology and fisheries.

In the tropics warm water pools tend to form in the western side of ocean basins due to the Trade Winds, in particular in the western equatorial Pacific Ocean around Indonesia large volumes of very warm water can build up over a period of years, with a much thinner thermocline layer existing in the eastern Pacific. Every few years (irregularly) the available potential energy represented by this warm water pool is released by El Niño events which see the warm water layer flow eastward along the Equator into the mid and eastern Pacific causing great changes in the ocean and atmospheric circulation. These El Niño events involve coupled ocean and atmospheric processes on a large scale, and it is through initializing the ocean component of coupled models that these events have been successfully forecast on periods out to 6 months ahead. This seasonal forecasting application of ocean data assimilation is now a major focus for a number of operational agencies around the world, e.g., European Centre for Medium-Range Weather Forecasts, ECMWF (Stockdale et al. 1998); Japanese Meteorological Agency, JMA (Ishii et al. 1998); National Centers of Environmental Prediction, NCEP (Barnston et al. 1999).

The various phenomena in the ocean that we wish to model and forecast through data assimilation techniques thus have a wide range of spatial and temporal scales

and this has a big influence on the approaches taken to modelling and to building data assimilation systems. The forecasting of near surface ocean currents and the mesoscale eddy field have much in common with the forecasting of atmospheric storms, with the underlying baroclinic instability process being similar in each. However, when studying climate time-scale phenomena, the properties of water masses, i.e., the volume of water with a particular set of temperature and salinity properties, are of more interest. The large range of important dynamical scales involved in the ocean also represents a major computational modelling challenge.

The structure of the rest of this chapter is as follows. Section 2 will discuss ocean modelling and highlight a few of the challenges that make it different from atmospheric modelling. We look at some of the current issues and model requirements for operational oceanography, seasonal forecasting to medium range climate prediction, and ocean reanalyses. Section 3 will look at some of the ocean observations that are available, both in the past and the current and planned systems which are part of the GOOS (Global Ocean Observing System; http://www.ioc-goos.org/). In particular, we will focus on altimeter sea level anomalies and geoid data from the new gravity satellites, and we will consider the Argo float array (http://www.argo.ucsd.edu) and the in situ water property data it provides. Section 4 is a general introduction to the current issues and applications associated with ocean data assimilation techniques. Section 5 will focus on the important area of altimeter data assimilation; Sect. 6 will discuss in situ temperature and salinity data assimilation. Section 7 will conclude with a forward look on ocean assimilation challenges.

## 2 Ocean Modelling Methods

A key objective in modelling the ocean is to determine the circulation. Ocean state estimation, Wunsch (1996), is first and foremost a modelling method to determine the time mean currents and ocean transports of mass, heat and freshwater. These transports are integrated quantities that cannot be measured directly and yet are critical to understanding how the ocean functions as part of the climate system. Full depth observations of temperature ($T$) and salinity ($S$), at high accuracy, are needed for the calculations and the World Ocean Circulation Experiment (WOCE), 1990–2002, sampled many regions of the deep ocean for the first time. Figure 1 shows the set of WOCE sections from the Atlantic. Ocean inverse theory is the data assimilation method used to combine these data into a self-consistent set of transports across all the available sections.

Ocean state estimation is limited in that it does not model the time-evolving circulation. Following developments in Meteorology (see chapter *The Role of the Model in the Data Assimilation System*, Rood), much of the early work on time-dependent ocean modelling, and certainly much of the work on ocean data assimilation in such models, was carried out with quasi-geostrophic models, seeking to simulate the wind driven upper ocean circulation, typified by the subpolar and subtropical

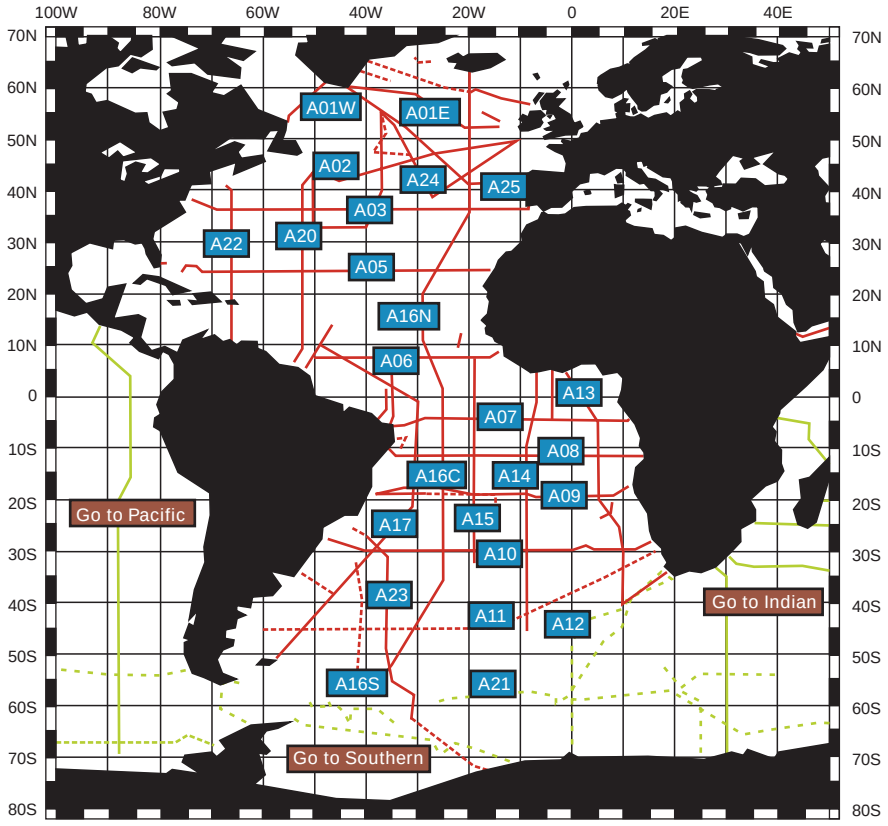



**Fig. 1** Location of Atlantic cruises where high quality *top* to *bottom* hydrographic data were gathered during the World Ocean Circulation Experiment (WOCE), 1990–2002

gyres (e.g. Bryan 1963; Holland 1978; Marshall 1984). While these models successfully capture the geostrophic aspects of the mid latitude upper ocean circulation (typified by small Rossby number flows) the quasi-geostrophic equations assume a uniform background stratification over the whole domain and thus are deficient in representing the large changes in stratification going from the tropical to the polar oceans (In contrast, tropospheric stratification changes much less between the tropics and the poles.). As ocean stratification reduces towards the poles, the Rossby deformation radius also reduces leading to a smaller mesoscale eddy field dominating the energetics and mixing processes in the ocean. In the atmosphere, although mesoscale cyclones do get smaller towards the poles (e.g. Polar Lows), the range of dynamical scales is not as great. This great range of important dynamical scales in the oceans has meant that full global ocean modelling is very expensive computationally. Only very recently have there been serious attempts to model the global oceans while retaining scales necessary to represent the ocean mesoscale (Hurlburt et al. 2008, and references therein). This computational challenge has also

stimulated particular oceanographic interest in advanced modelling techniques, including finite element models, curvilinear coordinate models (e.g. The Princeton Ocean Model, POM; Blumberg and Mellor 1987; Mellor 1996) and unstructured meshes (e.g. The Imperial College Ocean Model, ICOM; Piggott et al. 2008). Recently, atmospheric modellers have also renewed their interest in much higher resolution in order to model the dynamics of organized convective complexes within large-scale general circulation models (GCMs).

Today, modelling aimed at operational ocean forecasting is a relatively new activity which is picking up the challenges that operational meteorological forecasting faced 30 years ago. A strong operational ocean forecasting activity is seen as guaranteeing the continuation of a global ocean observing system (GOOS) and providing an invaluable source of data for improved understanding of many aspects of ocean and climate behaviour. As an example, the European consortium MyOcean (a project for the European Marine Core Service; http://myocean.oceanobs.com/html/about-us_en.html) is building a hierarchy of nested ocean models to allow ocean forecasting to cover a full range of scales from global 1/12° models to coastal models going down to 1–2 km resolution, and providing simulations of ice-ocean interactions in the Nordic and Arctic seas. Typically, these models are used for forecasting from 1 to 2 weeks ahead, with this time-scale likely limited by the accuracy in forecasting the meteorological winds required to drive them.

A common modelling framework, NEMO (Nucleus for European Modelling of the Ocean; http://www.locean-ipsl.upmc.fr/NEMO/) has been adopted in Europe to allow for collaborative and strategic model development of the wide range of important processes. Many of the applications of operational oceanography lie in the shelf seas and near coasts, and it is an ongoing challenge to understand what aspects of the shelf and coastal waters can be forecast, and which observations are most needed to make such forecasts successful. The emphasis has been on forecasting of ocean currents, mainly but not exclusively near surface, and studying the impact of interannual variations in upper ocean circulation on biogeochemistry, in the form of phytoplankton blooms and fish stocks.

In contrast to operational oceanography, modelling for seasonal forecasting has mainly focused on tropical oceans and in particular on the tropical Pacific where coupled models are needed to simulate the onset of El Niño events. Key ocean component requirements appear to be a reasonably high (1/3° or better) meridional resolution near the Equator in order to correctly represent the equatorial waveguide. Challenges here involve initializing the coupled models in such a way that the atmosphere and ocean are in balance so that large shocks to the boundary layer are not present at the beginning of each forecast (Stockdale 1997; Mochizuki et al. 2007). Seasonal forecasts are typically run for periods out to 6 months ahead, over which time the model biases and drift become important. Bias is typically corrected for by subtracting a pre-determined model drift from the forecasts (Stockdale 1997), but there may be more rigorous ways of achieving such corrections using data assimilation, e.g. Dee (2005) (briefly discussed later; see also chapter *Bias Estimation*, Ménard). The atmospheric behaviour out to 6 months ahead has a large chaotic

component so that seasonal forecasts are based on ensembles of coupled model runs which sample uncertainty in the evolving coupled state and seek to forecast only those elements of the forecast ensemble which are robust. Recently, the impact of model biases has also been reduced through the ensemble method by creating multi-model ensembles which use several different models to average out these model bias impacts on the forecasts (Palmer et al. 2004).

New challenges for ocean modelling and assimilation are now emerging in the need to analyse and forecast changes in climate. There are two challenges in particular worth highlighting. The first concerns the capability to extend coupled models from a seasonal to an interannual or medium-range climate forecasting role, and to properly understand the potential to forecast non-ENSO (El Niño Southern Oscillation) related signals. The recent paper by Smith et al. (2007) demonstrates the possibilities for a decadal prediction system. The sensitivity of all aspects of the coupled system to initial conditions in the ocean, land and cryosphere needs to be explored more fully. This will be the only route to making forecasts out to a few years ahead which would provide valuable input to policymakers. The second challenge is to use modelling and data assimilation to reconstruct critical aspects of the ocean circulation which have importance to climate. This could be called ocean synthesis, or ocean reanalysis if done in the context of an operational ocean analysis system. The time mean and time varying circulation and transports in the ocean are key quantities to be obtained by these methods. A good example of a key quantity of interest is the Atlantic meridional overturning circulation (MOC) and the extent to which it can be derived from current and past ocean observations combined with ocean modelling, Balmaseda et al. (2007).

## 3 Observational Ocean Data

If the objective is to use ocean assimilation for decadal hindcast studies, or for ocean reanalysis, then the record of historical ocean observations can be very sparse. Most data consist of vertical profiles of temperature, salinity and, occasionally, other quantities taken from ships at regular or irregular intervals along cruise tracks. The scientific focus of making hydrographic measurements was often on identifying water masses by their temperature and salinity properties. The “Core” method of Wust (1935) was then used to infer the time-mean ocean circulation from the spreading pathways of water. This involves identifying maxima or minima in properties such as salinity or oxygen in vertical profiles and then tracing these extrema back to their intersections with the surface or mixed layer. Figure 2 is the WOCE section A16 from Fig. 1, and it shows a trans-Atlantic section of water property measurements suitable for analysis. More sophisticated methods such as end-member analysis (Tomczak 1981) provide further information on the origin and circulation pathways of water masses. These methods rely on the long lifetimes, and hence Lagrangian tracer properties, of the water masses once they are out of touch with the ocean mixed layer. This contrasts with the much shorter lifetime of meteorological air masses, essentially because the lack of penetrative radiational heating into

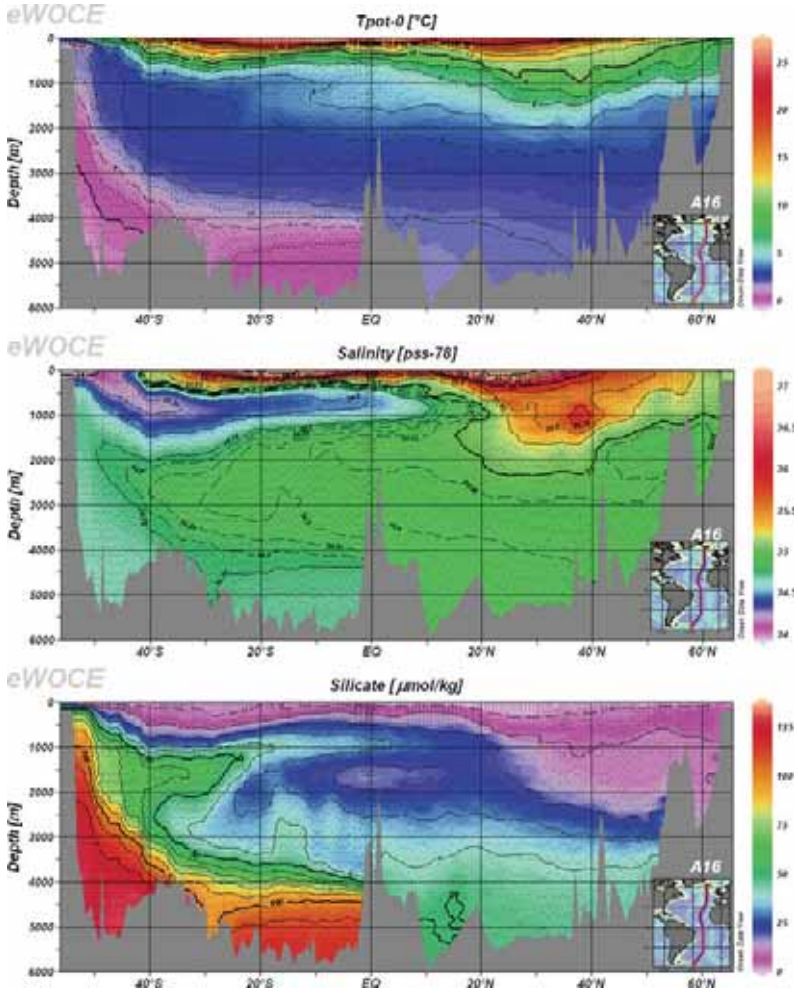



**Fig. 2** Temperature (*top plot*, °C), salinity (*middle plot*, psu) and silicate (*bottom plot*, µmol $kg^{-1}$), measured during WOCE (World Ocean Circulation Experiment) cruise A16 running approximately North-South through the Atlantic (see *inset*). Property values indicate the spreading of waters through the deep ocean. Antarctic Intermediate water (AIW – *dark blue* at ~1,000 m depth, south of the Equator), North Atlantic Deep water (NADW – *bright green* between ~1,000 m and ~4,000 m depth, north of 20°S) and Antarctic Bottom water (ABW – *dark green* at ~5,000 m depth, south of the Equator) are clearly distinguished in the middle plot. The highest temperatures, salinity and silicate values are marked in *red*. The lowest temperatures, salinity and silicate values are marked in *purple*. See Haines (2003a)

the ocean means that the ocean circulation is being driven from the top, compared to the atmosphere being primarily driven from below.

However, most of the ocean observations in the historical databases such as WOA (World Ocean Atlas; Conkwright et al. 2002) are made up of temperature measurements alone, and so do not permit such detailed water property analyses.

Nonetheless, temperature primarily determines the ocean density, and hence the immediate geostrophic circulation, through most of the upper 1 km, except in polar regions where salinity becomes more important (Gill 1982). Ocean temperature also determines the solubility of gases such as $CO_2$. The ocean temperatures and heat content also need to be understood from a global warming perspective and because the thermosteric signal largely determines ocean sea level changes, see chapter 5 of the IPCC 4th Assessment (IPCC 2007). Assimilation of these measurements can therefore potentially help in the quantification of these effects.

Before the 1970s most temperature measurements were made by Mechanical Bathythermographs (MBTs) covering only the top 200–450 m in depth, and with most measurements being made within a few 100 km of coastlines. Expendable Bathythermographs (XBTs), measuring down to 800 m, came into wider use in the 1970s and were deployed to be operated by Voluntary Observing Ships (VOS) along normal trade routes. This widened the coverage, but still produced a heavy North Atlantic and North Pacific bias in distribution. In the late 1980s and 1990s the Tropical Atmosphere Ocean (TAO; http://www.pmel.noaa.gov/tao/) array of bottom moored buoys was put in place in the tropical Pacific to monitor upper ocean temperatures to provide understanding and warning of El Niño events. The success of TAO later lead to the deployment of a similar tropical Atlantic array, PIRATA (formerly the Pilot Research Moored Array in the Tropical Atlantic, now Prediction and Research Moored Array in the Tropical Atlantic; http://www.pmel.noaa.gov/pirata/), recognizing the huge importance of tropical oceans around the globe in influencing atmospheric behaviour (Philander 2002).

However, only with the development of the Argo programme of ~3,000 autonomous profiling (down to 1.5–2 km) floats deployed throughout the world oceans (starting around the year 2000 and reaching completion in 2007), has a truly global and regular set of ocean temperature and salinity (and hence density) measurements become available. An example of the typical Argo coverage of data in a single month is shown in Fig. 3, for June 2008. These measurements provide a new source of data against which ocean models can really be tested in considerable detail. The challenge for data assimilation is to show that the models can successfully interpret all of these data and derive circulation patterns with greater accuracy than hitherto.

Although profile observations such as those discussed above provide the only way of properly sampling the three-dimensional (3-D) structure of the oceans, satellite observations now also play a vital role in understanding the oceans. The accuracy and long lifetime of the TOPEX/Poseidon altimeter satellite (1992–2005) transformed the synoptic measurement of the ocean. With its successor satellites JASON-1 and JASON-2 being successful, it has now provided continuous measurement of ocean dynamic topography (sea level) changes, with an accuracy of a few cm for over 17 years. Horizontal gradients in sea level relative to the Earth's geopotential surface, or geoid, determine the surface geostrophic currents. Unfortunately, we do not know the geoid as well as is needed for this calculation, with Fig. 4 showing a recent estimate of mean sea level variations directly from satellite altimeter and gravity data. This map is fairly smooth because it is only on the larger scales that

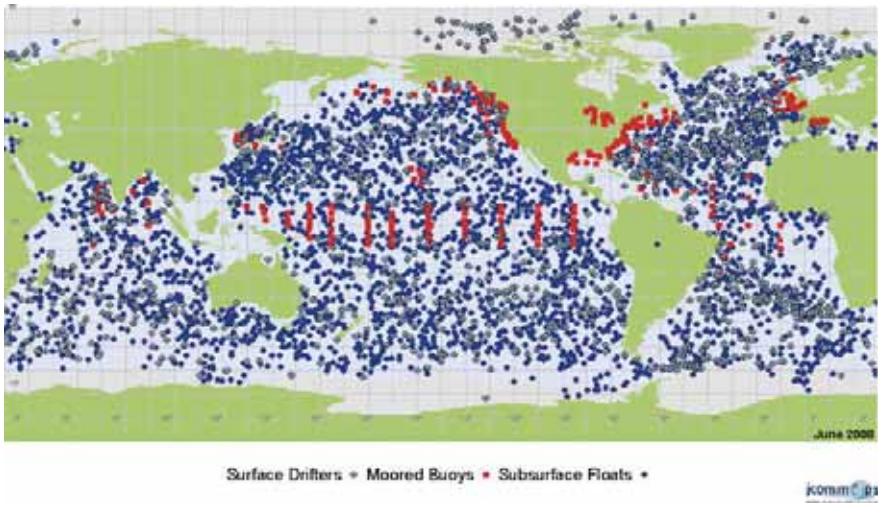



**Fig. 3** Current Argo observations (June 2008). The distribution is dominated by subsurface float data

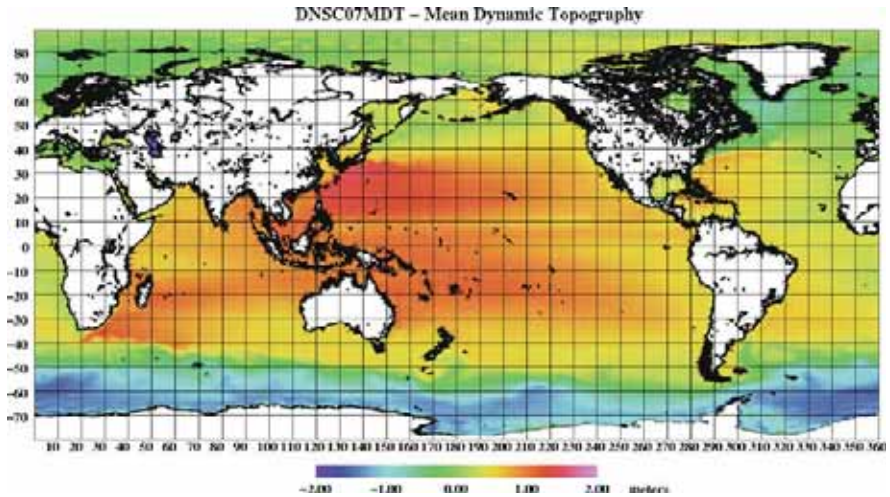



**Fig. 4** Surface mean dynamic topography from the Danish National Space Centre, units of metres. Gradients in this field give the mean surface geostrophic ocean circulation. *Red* indicates positive values; *blue* indicates negative values

the geoid is known with sufficient accuracy (Hughes and Bingham 2008; Bingham et al. 2008). However, the time varying component of the altimeter signal can be used to give the time varying component of surface currents with better accuracy. This continuous record of altimeter data has been the main stimulus to operational oceanography over the last decade and has been the key driver for most ocean data assimilation research. We will return to this later in the chapter.

In the last few years there has been an effort to resolve the missing geoid by the development of new gravity satellites GRACE (GRAvity and Climate Experiment) and GOCE (Gravity and Ocean Circulation Explorer) – see chapter *Research Satellites* (Lahoz). GRACE has already provided a much better global geoid than has been available before, giving a better resolution on ocean gyre scales (Tapley et al. 2005). GOCE (launched in March 2009) should increase the resolution of this geoid down to useful scales of ∼100 km, and help resolve many individual ocean currents (Drinkwater et al. 2007). The ability to assimilate a combination of altimeter and ocean geoid information is a new challenge to the ocean assimilation community (Drecourt et al. 2006; Lea et al. 2008), to which we return later in the chapter.

Finally, satellite measurements of ocean surface temperature can also provide important information about ocean current pathways, and should also provide information about air-sea interactions. The record of infra-red (IR) measurements from AVHRR (Advanced Very High Resolution Radiometer) goes back more than two decades. These products are combined with in situ data to produce regular SST (sea surface temperature) products such as that from Reynolds and Smith (1994), which are then commonly used in ocean and seasonal assimilation work at several operational centres (e.g. ECMWF). The accuracy of these IR measurements was greatly increased (to 0.1 K) by the ATSR (Along Track Scanning Radiometer) instrument series (starting in 1993 on ERS-1, the first European Remote Sensing satellite) which use a dual path measurement to provide atmospheric corrections. Skin temperatures (i.e., changes in temperature within microns of the sea surface) still cause some problems for IR measurements, as does the inability to measure through cloud (Merchant et al. 2008). Recently, microwave measurements have permitted global SST maps to be produced continuously, as microwaves can penetrate clouds. Microwave SST accuracies are influenced by the sea state (roughness on cm scales corresponding to the microwave wavelengths); however, combined products using ATSR accurate IR information along with the global microwave coverage has made big advances possible. For example, the Global Ocean Data Assimilation Experiment (GODAE) High Resolution SST project (GHRSST; see http://www.ghrsst-pp.org/index.htm), has recently started to generate accurate global SST products in an operational system, e.g., OSTIA (Operational Sea Surface Temperature and Sea Ice Analysis; http://ghrsst-pp.metoffice.com/pages/latest_analysis/ostia.html). It is expected that the assimilation of these products into both atmospheric and ocean forecasting systems will have important impacts on forecast skill.

The above discussion covers the major ocean observational datasets used in physical ocean modelling and data assimilation work at the present time. A clear omission, and an area that is becoming increasingly important, is the application of data assimilation to marine biological modelling. Here, the key data sets are in situ measurements from maintained mooring sites which can measure various nutrient concentrations, as well as sampling biological parameters such as phytoplankton and zooplankton concentrations. On a more global scale, ocean colour satellites, in particular SeaWIFS (Sea-viewing Wide Field-of-view Sensor;

http://oceancolor.gsfc.nasa.gov/SeaWiFS/), have provided a continuous record since 1999, and these data are a main target for modelling and data assimilation studies.

We finish this section by providing a representative list of past, current and future data sources for ocean data assimilation.

(i) In situ measurements:

- Temperature profiles: Mostly from Expendable bathythermographs (XBTs) from Voluntary Observing Ships. From 1950 to 1970 only the top 200–450 m were normally sampled, but more recently this has increased to 800 m. A typical annual global distribution (1993) is shown in Fig. 5 (which includes Tropical Atmosphere-Ocean, TAO, buoy array data from the tropical Pacific);
- Conductivity Temperature Depth (CTD) instruments measure temperature and salinity (and hence density), sometimes to full ocean depth, with high accuracy from research vessels. Historically there are a much lower number of observations. Fixed moorings, also measuring temperature and salinity, are becoming more widespread for special observing or monitoring campaigns, e.g. in the Arctic straits (ASOF 2008), across the Atlantic at 26°N (the Atlantic Meridional Overturning Circulation, AMOC array; Cunningham et al. 2007), or in the straits of the Indonesian archipelago. Eventually, all these data should become available for model validation and assimilation;
- The Argo neutrally buoyant float programme now provides temperature ($T$) and salinity ($S$) profiles in the top 2,000 m, surfacing every 2 weeks and transmitting data via satellite. They have a nominal 5-year deployment lifetime. Real time Argo data are available from www.ifremer.com.fr/coriolis. In

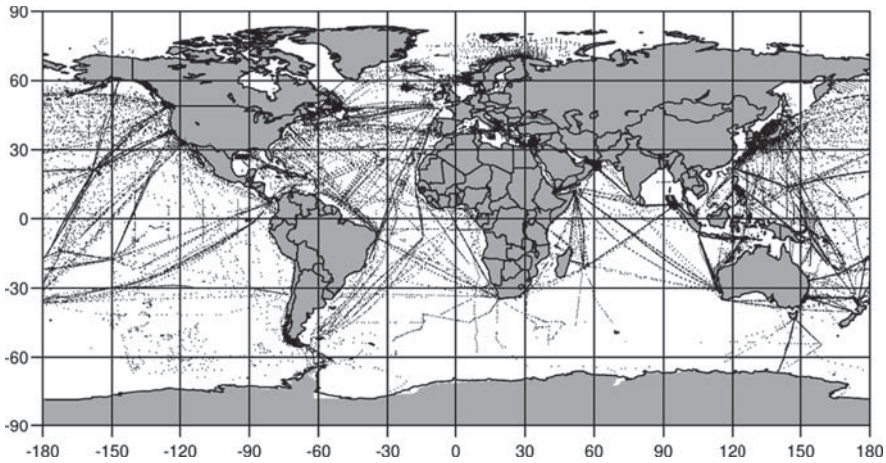



**Fig. 5** Locations of all the temperature profile data available for 1993. A total of 69,980 profiles were used. See Haines (2003b)

future, gliders will provide steerable automated in situ observing platforms, although it is not clear whether they could be developed cheaply enough to replace Argo technology.

(ii) Satellite Measurements (see chapter *Research Satellites*, Lahoz):

- Satellite Sea Surface Temperatures (SSTs): These are measured by passive IR (AVHRR, ATSR) or microwave (SSM/I, Special Sensor Microwave Imager; TRMM, Tropical Rainfall Measuring Mission) instruments. The data have been little used for assimilation so far due to problems with rapid diurnal variations in surface temperature, and skin temperature effects which are not well represented in ocean models; however, with the new combined products such as OSTIA, there will be a high priority to use these data at the operational centres;
- New satellites measuring the Earth's gravity field, GRACE and GOCE missions, will greatly improve knowledge of the Earth's geoid. These data will be very important for inverse calculation of the ocean circulation, as well as assimilation into time dependent models. More information can be found at http://www.csr.utexas.edu/grace/ and http://www.esa.int/export/esaLP/goce.html;
- Satellite altimeters have provided continuous global coverage since 1992. The altimeter instrument is a microwave radar measuring sea level relative to the satellite with an accuracy of 2–3 cm. Corrections for atmospheric signal delays, and inverse barometer and tidal sea level variability are required. Altimeter sea level slopes relative to the geoid indicate surface geostrophic currents. Data are available along tracks or as maps, usually every 10 days; these data can be assimilated to provide mesoscale upper ocean current variations.

## 4 Ocean Data Assimilation: Applications and Current Issues

In this section we briefly list some of the issues that arise in the application of data assimilation techniques to the modelling of the large scale ocean. Later sections focus on two particular open ocean data assimilation problems: assimilation of altimeter data (Sect. 5); and in situ temperature and salinity data assimilation (Sect. 6). First, two important ocean assimilation applications, associated with surface waves and with coastal sea level, are mentioned very briefly for completeness. These applications have the following characteristics:

- *Ocean wave forecasting:* This is run as an operational service for ships and oil platforms and relies very strongly on meteorological conditions. Models of surface wave spectra and propagation directions contain representations of "Wind Sea" and "Swell". Satellite altimeters (see Sect. 6) can give wave height measurements for assimilation but the spatial/temporal coverage is relatively poor

on meteorological timescales. A multi-altimeter mission might improve this. A good although now dated overview of surface wave modelling and assimilation is given in Komen et al. (1994);

- *Tidal/Storm surge forecasting:* This is operational for coastlines, estuaries, lagoons and tidal rivers. There is a strong meteorological dependency with wind driven Ekman and Inverse Barometer (–1 hPa atmospheric pressure change = +1 cm sea level change) effects building up the sea level during storms. Data from tide gauges along coasts can be assimilated (surges propagate anti-clockwise around basins in the Northern Hemisphere). Altimeter data assimilation for storm surge modelling is still being researched. Coverage is limited at the short time-scales involved and there are difficulties relating mean sea level with the height system reference for the tide gauges; this may be overcome with improved geoid data in future. Examples of forecasting systems include the Adriatic (Venice Lagoon), and the North Sea (the Thames Flood barrier).

Both of the above examples are mature areas, with well-understood dynamical models that are essentially two-dimensional, so that optimal methods of error treatment, such as the Kalman filter, are tractable (see chapter *Mathematical Concepts of Data Assimilation*, Nichols). However as for many environmental forecasting systems, the real challenges lie in predicting extreme (i.e., dangerous) events, which is much more difficult. The above two areas are intimately bound up with the problems of meteorological forecasting (see chapter *Numerical Weather Prediction*, Swinbank).

The open ocean data assimilation sections to follow focus on the interaction between assimilation and physical processes in ocean or coupled models, rather than on the formulation of the error characteristics, the main features of which are covered elsewhere in the book (see chapter *Error Statistics in Data Assimilation: Estimation and Modelling*, Buehner). Note that one particular physical difference between the ocean and atmospheric assimilation problems is that the longer timescales of thermodynamic processes in the ocean (already alluded to in Sect. 2) mean that careful treatment of conservative or Lagrangian properties of the water masses during the assimilation cycle has proved a valuable constraint which can be exploited to constrain covariances between physical quantities and improve ocean assimilation results.

Issues in ocean data assimilation include:

(i) Operational oceanography applications (see also Sect. 5):

- Combined assimilation of altimeter sea level anomalies and mean dynamic topography (MDT) data from geoid measurements (Rio and Hernandez 2004);
- Assimilation accounting for biases (observational and/or forecast; e.g. Dee and da Silva 1998; Dee 2005, for atmospheric applications);
- Assimilation of sea level data at higher latitudes;
- Differences between sequential approaches (e.g. Kalman filtering) and variational approaches (e.g. 4-D variational, 4D-Var) – see Lea et al. (2008).

(ii) Seasonal forecasting applications:

- Ensemble assimilation and forecasting methods (e.g. Ensemble Kalman filter, EnKF);
- Balancing increments for assimilation of in situ data (e.g. Weaver et al. 2005);
- Multi-model skill versus single-model skill;
- Extension to systems encompassing interannual to decadal forecasting (e.g. the Hadley Centre Decadal Prediction System, DePreSys), and which include assimilation of observed ocean and atmospheric anomalies (Smith et al. 2007).
- Benefits of Earth system assimilation, i.e., direct assimilation into the coupled system;
- 4D-Var applications to the coupled system (e.g. Awaji et al. 2002).

(iii) Assimilation for climate reanalyses:

- Comparison of the long-window 4D-Var ECCO (Estimating the Circulation and Climate of the Ocean; http://www.ecco.ucsd.edu) consortium approach against the 3D-Var and sequential assimilation (optimal interpolation, OI; Kalman filter) approaches used in operational applications;
- The problem of chaos in the long-window 4D-Var approach (Lea et al. 2000);
- Balancing increments and the construction of water mass variability;
- Attribution of ocean reanalysis changes to changes in surface forcing, e.g. Stammer et al. (2004).
- Assimilation of sea surface salinity (SSS), see, e.g., Durand et al. (2002) (see also Sect. 7).

(iv) Biological ocean assimilation:

- The need for assimilation of high resolution information, i.e., mesoscale eddies (e.g. Oschlies and Garcon 1998);
- Needs of operational biological modelling versus carbon cycle modelling;
- The use of semi-prognostic methods for bias correction and shock reduction (e.g. when mixing different water masses; Eden and Oschlies 2006);
- Interactions between biogeochemical cycles and marine food webs (e.g. the IMBER, Integrated Marine Biogeochemistry and Ecosystem Research, project; http://www.imber.info);
- Developing fast and efficient climate-carbon models that incorporate the biogeochemistry of the oceanic carbon cycle (see, e.g., Palmer and Totterdell 2001).

(v) General Comments:

- There is a need to demonstrate the added value of ocean data assimilation (beyond simply statistical interpolation) in the recently data rich era (3,000 Argo floats and up to 4 altimeters flying);

- There is a need to demonstrate the value of surface ocean prediction for medium range weather forecasting;
- The potential of Observing System Simulation Experiments (OSSEs; see chapter *Observing System Simulation Experiments*, Masutani et al.) to plan the GOOS needs further exploration – this is also an aspiration for activities beyond GODAE (see below);
- Operational analyses from the international GODAE programme can now be found through a number of websites (www.godae.org; www.mersea.eu.org). Data portal technology, e.g. www.reading.ac.uk/mersea, is also allowing easy browsing of these products. Assimilated data comes from altimeters, from Voluntary Observing Ships (VOS) and from Argo profiling floats (see Haines 2003b). Uses for GODAE products include pollution monitoring, fishing and tourism – see Koblinsky and Smith (2001);
- GODAE ended in 2008 but the community activities are continuing under GODAE OceanView. Aspirations for activities include: (i) transition to operational systems, from the demonstration of valuable services to the provision of operational services with high availability and reliability; (ii) continuous improvements of systems through scientific and technological advances, including extending existing capabilities to the coastal zone and ecosystems, and studying the coupled ocean-atmosphere system and climate; (iii) development of operational services and further links with applications; (iv) ensuring a suitable sustained ocean observing system through the demonstration of the value of observations (e.g. via OSSEs and Observing System Experiments, OSEs). A GODAE journal special issue has been published in *Oceanography* (2009), Vol 22(3) (see http://www.godae.org/GODAE-Special-Issue.html).

## 5 Altimeter Data Assimilation

This section and the next describe in more detail, from an oceanographic perspective, how the two most important sources of ocean data may be assimilated.

### *5.1 General Considerations*

If sea level variations can be measured relative to the surface of constant geopotential called the geoid, which we will identify as $z = 0$, then they are equivalent to pressure variations which are related geostrophically to the surface flow**.** Interestingly, Hughes and Bingham (2008) have suggested recently that the geopotential could instead be measured at the position of the sea level, thus avoiding some of the complications involved in defining the geoid. Mean sea level is used in the meteorological community as if it were a geopotential surface, however it is not. The true geopotential surface deviates from mean sea level by up to 1 m (irrelevant in meteorology but very important for determining ocean currents). Figure 6

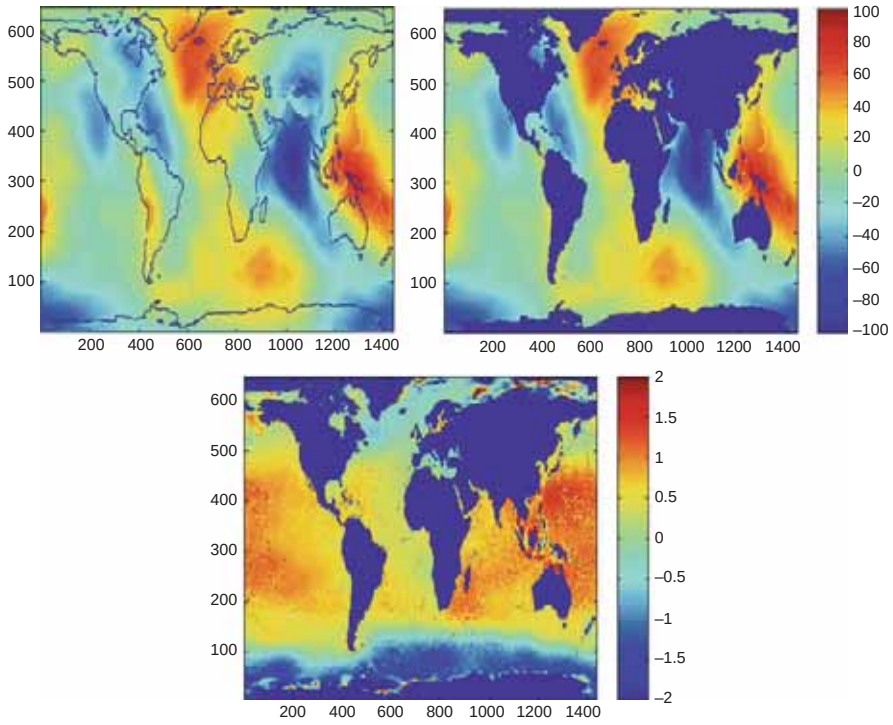



**Fig. 6** The *top left panel* shows the EGM96 geoid height relative to a reference ellipsoidal Earth. The *top right panel* shows the mean sea level determined by altimeter data relative to the same ellipsoid as EGM96. This mean sea level was produced by Hernandez and Shaeffer (2000). The *lower panel* shows the difference (sea level minus geoid). All units are in m. In the *top panels*, *red* indicates values larger than the reference ellipsoid; *blue* indicates values lower than the reference ellipsoid. In the *bottom panel*, *red* indicates positive differences; *blue* indicates negative differences. See Haines (2003b)

(left panel) shows one of the better global geoids, EGM96 (although products from GRACE have now superceded this), as variations from a reference ellipsoidal Earth. Figure 6 (middle panel) shows the mean sea level determined by altimeter data relative to the same ellipsoidal Earth (Hernandez and Shaeffer 2000). Figure 6 (right panel) shows the difference, which should be a stream function for the mean surface geostrophic flow (Fig. 4 was calculated in a similar manner). Although the large-scale features are reasonably consistent with this, the small mesoscale features are completely unrealistic due to inaccuracies in the geoid data at these scales.

Although the true geoid is not known accurately enough at the mesoscale, the time varying component of the altimeter signal, or the sea level anomaly, can still be assimilated with mesoscale accuracy. This must be compared with an equivalent "anomaly" sea level from the model, and to define this from the full sea level we need a separate definition of mean sea level. Several ways have been used to define this mean sea level:

1. A previous run of the ocean model without data assimilation is often used to determine a mean sea level over some period (this period should really be the same as that used to define the altimeter sea level anomalies). A disadvantage is that it is often known that this model mean sea level is biased in some areas and this bias will be preserved;
2. A previous run of the ocean model can be performed without altimeter data assimilation but with assimilation of hydrographic data. The resulting sea level can be used for a subsequent assimilation run over the same period with both hydrographic and altimeter data assimilated (Fox and Haines 2003);
3. An independent sea level anomaly can be determined from hydrographic and other data. Fox et al. (2000) and Killworth et al. (2001) used only climatological hydrography with a dynamic height calculation. More recently, Rio and Hernandez (2004) have combined hydrography with surface drifter information and satellite geoid data to define a mean dynamic topography. Niiler et al. (2003) have defined mean dynamic topography from surface drifter data alone;
4. It is possible to calculate independent local geoids using geodetic data alone with the necessary small-scale accuracy, e.g. Hunegnaw et al. (2009), but these have not been used up to now in ocean assimilation studies.

Of the methods introduced above, (2) or (3) are perhaps the most reliable, although small scales may still be absent or incorrectly represented where very few in situ data are available (e.g. the Southern Ocean). We now look at the problem of recovering subsurface information from altimeter surface topography data by considering the physical relationships between variables of altimeter data.

## *5.2 Physical Relationships Between Variables*

Knowledge of the covariance relationships between sea level anomalies and anomalies in other quantities (e.g. temperature, salinity, currents at depth) is needed in some form for any altimeter assimilation method. What we emphasize here is that there are different ways of representing this information, and some ways are more succinct and easier to verify empirically than others.

Sea level or pressure variations on the geoid $p(z = 0)$ can be broken down into pressure variations at the sea floor ($z = -H$) and hydrographic variations from the water column density (assuming hydrostatic balance):

$$p(0) = p(-H) - g \int_{-H}^{0} \rho(z) dz. \tag{1}$$

Large-scale rapid variations in $p(0)$ tend to be barotropic and have variations at $p(-H)$ similar in magnitude and well correlated with $p(0)$ variations. Smaller scale

more persistent variations in $p(0)$ tend to be strongly baroclinic with only weak correlations with $p(-H)$. It is necessary to be able to make this distinction if observations of sea level $p(0)$ variations are to be correctly assimilated into a model. In what follows it will also be useful to define:

$$D(x,y) = -g \int_{-z_0}^{0} \rho(x,y,z)dz, \tag{2}$$

the dynamic height at the sea surface relative to some level $-z_0$, which is determined entirely from hydrographic data. Provided that all horizontal pressure variations at level $-z_0$ are negligible (i.e., the level of no motion assumption) then $p(0) = D(x, y)$.

Methods for obtaining subsurface quantities by projection of sea level anomalies below the surface can be broadly broken into two classes, empirical projection and dynamical projection. Much of the discussion below is based on that given in Haines (1994).

*Empirical projection.* This should be based on concurrent observations of local hydrographic or current meter data and sea level over a long period of time. However, usually this criterion cannot be met. A way of developing relationships is with Empirical Orthogonal Functions (EOFs). To illustrate this method we look at some early results from De Mey and Robinson (1987). They used 1 year of data from the POLYMODE current meter array in the North-West Atlantic to develop EOFs of the vertical pressure variability, shown in Fig. 7. The first two modes represent 81.5 and 16.7% of the pressure variance, respectively. They reasoned that if only sea surface height data are available it makes sense to project them onto the surface enhanced first EOF mode and thereby to recover the pressure variations at depth. De Mey and Robinson (1987) used this method to assimilate the surface

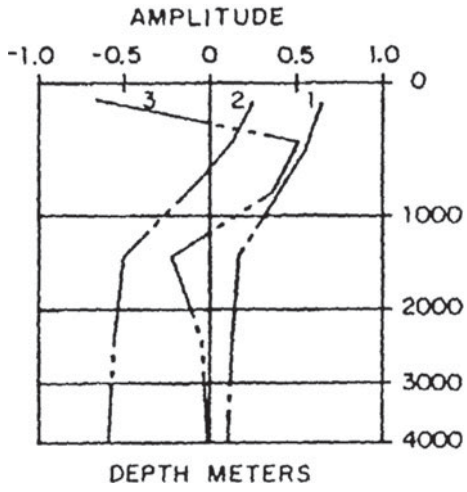



**Fig. 7** First three *vertical* Empirical Orthogonal functions (labelled 1, 2 and 3) of pressure variations in the POLYMODE experiment from De Mey and Robinson (1987). See also Haines (2003b)

data alone from POLYMODE and managed to partly recover deeper pressures and currents.

The problem with this method is that the vertical modes can be very variable spatially and possibly also in time, depending on the vertical thermocline structure. Hurlburt et al. (1990) developed a much wider set of correlation functions to relate sea level variability at one location with three-dimensional pressure variations. The problem here is that the only way to develop these full covariances is by using model output data, which may well be strongly biased.

Variations on the EOF theme can be found where sea level is correlated directly with hydrographic water properties, temperature and salinity (Mellor and Ezer 1991; Ezer and Mellor 1994). Figure 8 shows correlations of sea surface height variations and density variations at several depths within Mellor and Ezer's limited area high resolution model of the Gulf Stream along the US East coast. The correlations are high even down 1,000 m below the surface, making it feasible to use these correlations for assimilation of sea level data. One big advantage is that the temperature,

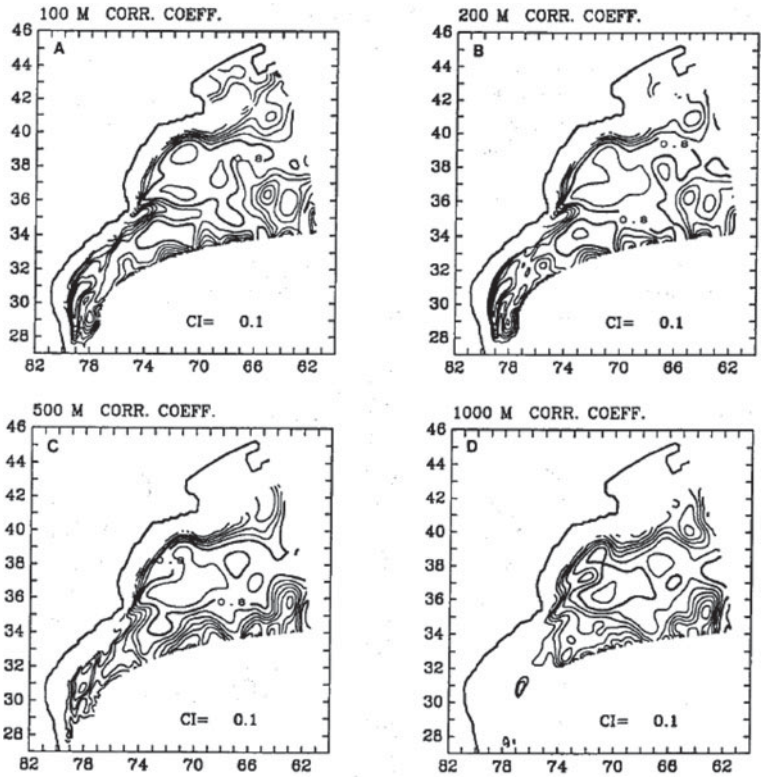



**Fig. 8** Correlations between surface height variance and subsurface density variance at different levels below the surface. Contour intervals are 0.1 apart and the 0.8 contour is in *bold*. From Mellor and Ezer (1991). See also Haines (2003b)

salinity and density are the appropriate state variables for full primitive equation ocean models, while the pressure field correlations are really only suitable for quasi-geostrophic models. However, the correlations still have to be derived from a model and are not, therefore, necessarily realistic.

Oschlies and Willebrand (1996) used vertical current correlations obtained from a primitive equation model of the North West Atlantic. This is rather similar to calculating pressure correlations; however, they also calculated consistent changes in density (via the thermal wind relation) with the temperature-salinity characteristics preserved in the model. This makes the method similar in some respects to the methods discussed below under dynamical projection. There does not obviously seem to be much to distinguish between these empirical methods, however in the next section we show that there are other ways to project altimeter data.

*Dynamical projection.* These ideas were originally developed around quasi-geostrophic theory and it is still useful to review some results in this framework. The key result is that a change of state vector will lead to new coordinates in which knowledge about the ocean variability at depth can be more succinctly expressed in terms of the sea level.

Haines (1991) used an idealized 4-layer quasi-geostrophic ocean gyre model to illustrate the assimilation of sea level data. Figure 9 shows the stream function $\psi$ and the potential vorticity $q$ in each layer at some instant. The flow in the model broadly represents subtropical and subpolar gyres with a strong current between them penetrating an ocean basin and going unstable (cf. the Gulf Stream). Altimeter sea level data is equivalent to observations of $\psi_1$, the top layer stream function. The

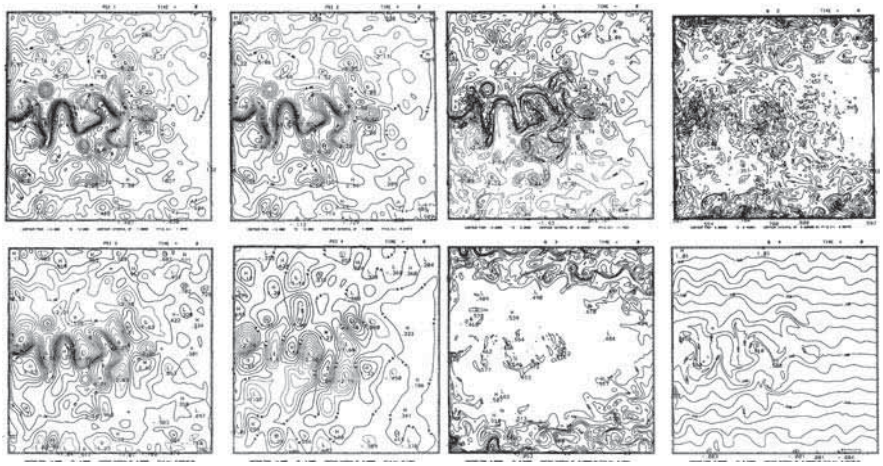


**Fig. 9** Stream function (*left four panels*) and potential vorticity (*right four panels*) fields from a 4-layer quasi-geostrophic ocean box model. The surface layer field is in the *top left* of each group, second layer *top right*, third layer *bottom left* and the fourth layer in the *bottom right*. Note the very different covariance relations between anomalies in the different fields. From Haines (1991). See also Haines (2003b)

fields are related by:

$$q_1 = \nabla^2\psi_1 + \beta y - \gamma_{1,2}^2(\psi_1 - \psi_2), \tag{3a}$$

$$q_2 = \nabla^2\psi_2 + \beta y - \gamma_{2,1}^2(\psi_2 - \psi_1) - \gamma_{2,3}^2(\psi_2 - \psi_3), \tag{3b}$$

$$q_3 = \nabla^2\psi_3 + \beta y - \gamma_{3,2}^2(\psi_3 - \psi_2) - \gamma_{3,4}^2(\psi_3 - \psi_4), \tag{3c}$$

$$q_4 = \nabla^2\psi_4 + \beta y - \gamma_{4,3}^2(\psi_4 - \psi_3), \tag{3d}$$

where $\beta$ is the northward Coriolis gradient and $\gamma$ are the Rossby deformation radii between layers. The point of showing these fields is that $q_i$ ($i = 1$–$4$) provides an alternative state vector for describing the system and yet the correlations in the vertical are completely different for $\psi$ and $q$. In particular, $q$ is virtually uncorrelated vertically, due largely to its Lagrangian properties and the fact that $q$ gradients have been mixed away below the surface. Haines (1991) suggested a mixed state vector representation for data assimilation purposes using $\psi_1$ from observations and $q_2$, $q_3$, $q_4$ from the a priori model. This mixed description is complete in the sense that all the other fields may be found, and it removes the need to use vertical correlations in $\psi$.

Figure 10 illustrates the convergence of an identical twin assimilation experiment (see chapter *Observing System Simulation Experiments*, Masutani et al.). Note particularly that, although $q_2$, $q_3$, $q_4$ are not changed at all during each assimilation, their errors decrease over time as the model runs forward, so that deep property

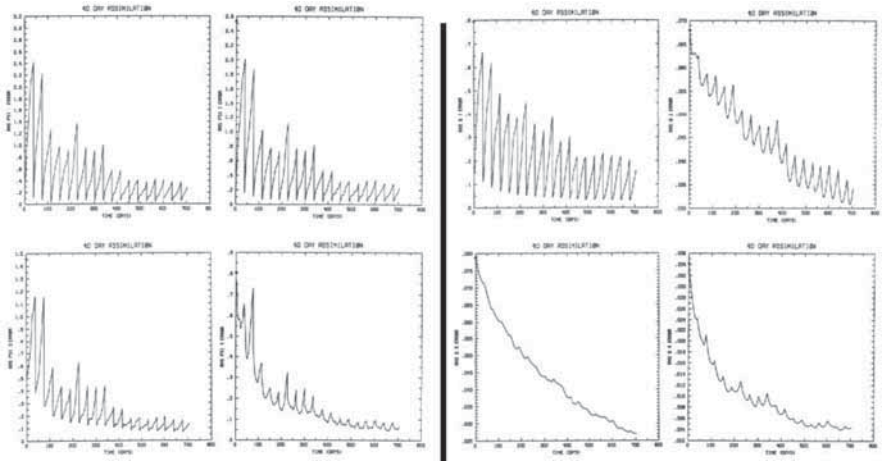


**Fig. 10** Convergence of the stream function (*left four panels*) and potential vorticity (*right four panels*) root-mean-square (RMS) errors during a data assimilation twin experiment in which surface stream function data are assimilated (representing surface altimeter data) every 40 days. The layers are as described in Fig. 9. From Haines (1991). See also Haines (2003b)

fields ($q$ in this case) are recovered over time despite the lack of correlation with the surface stream function. This is a powerful and attractive idea when considering the importance of deep tracer fields (see Haines 2003a).

Cooper and Haines (1996) extended this idea to a primitive equation framework of the oceans, in which potential vorticity is given by:

$$q = \frac{f}{\rho_0}\frac{\partial \rho}{\partial z}. \tag{4}$$

Now, using the a priori model $q$ for the deep oceans means keeping the stratification $\delta\rho/\delta z$ constant *as a function of* $\rho$. The only way to do this, while still changing the density field, is to vertically displace the water column; see Fig. 11. To close the problem it was assumed that the sea level anomalies are essentially baroclinic, so a constraint of no change to the deep pressure field was imposed:

$$\Delta p(0) = g\int_{-H}^{0} \Delta\rho(z)dz, \text{ where } \Delta\rho(0) = \frac{\partial \rho}{\partial z}\Delta h, \tag{5}$$

for small vertical displacements $\Delta h$. It should be noted that the constraint of no change to the deep pressure is different from the solution found in Haines (1991) in the quasi-geostrophic framework. In the quasi-geostrophic framework this constraint is equivalent to taking $\psi_1$ from observations, and $q_2$, $q_3$ and $\psi_4$ from the model a priori, and it would be an interesting exercise to carry out such a study in the idealized framework.

The dynamical approach to the vertical projection of altimeter data, described above, makes a virtue out of not changing certain quantities in the model at all and thus dispensing with multivariate covariance functions completely. The idea can be compared to the constraints imposed during ocean state estimation where minimal changes to water properties within regions are sought (Wunsch 1996). However, the necessary recasting of the state vector can also be used to develop a covariance

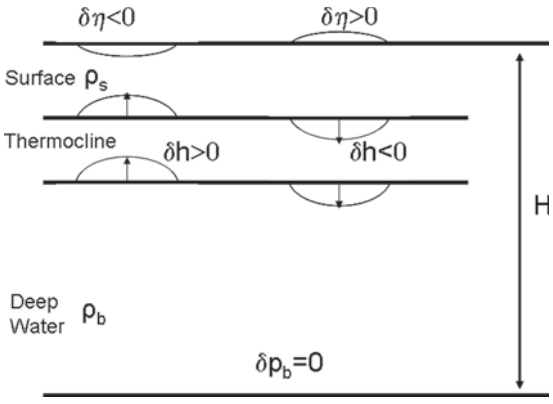



**Fig. 11** Schematic altimeter assimilation by *vertical* displacement of the thermocline. The *vertical* displacements $\delta h$ are calculated to cancel the sea level change leading to no change in pressure at the ocean floor. From Cooper and Haines (1996). See also Haines (2003b)

approach. Gavart and De Mey (1997) studied the empirical covariances of sea level and the depth of isopycnals. Since the depth of all density (or temperature) surfaces contains precisely the same information as the density (or temperature) as a function of depth, the two descriptions are obviously equivalent. However the covariance information looks quite different for the two descriptions. Empirical orthogonal functions were calculated from hydrographic profiles measured around the Azores current (SEMAPHORE project; Eymard et al. 1996) using different descriptions of the data (see Fig. 12). Note, particularly, that the first EOF of vertical displacement is very uniform with depth, providing support for the Cooper and Haines (1996) method. When the hydrographic data were projected only onto the first mode $z$-level EOF, and dynamic heights calculated from the result, then 8.4 $cm^2$ of sea level variance is left as a residual. In contrast when a projection onto the first mode isopycnal EOF is made, only 3.6 $cm^2$ of residual sea level variance remains. Thus, the isopycnal coordinate system provides a more compact representation of the hydrographic variance, which can be used in the altimeter data assimilation process.

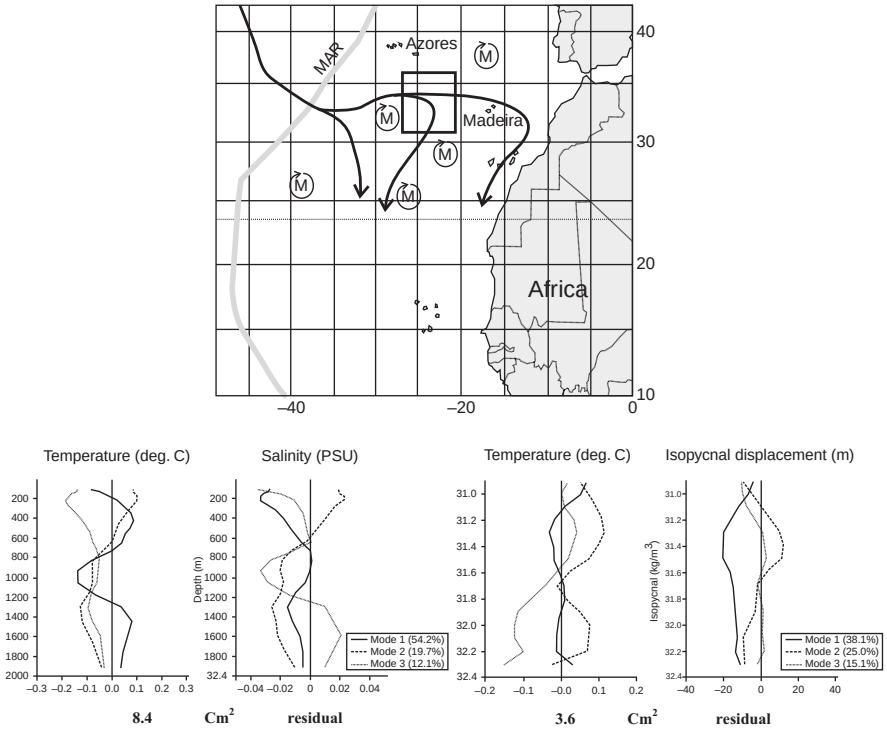



**Fig. 12** Empirical Orthogonal functions (EOF) describing the *vertical* variance of temperature (°C) and salinity (psu) in the Azores box (*top panel*). The *bottom left* two panels show the first 3 EOFs as a function of depth. The *bottom right two panels* show the first 3 EOFs of temperature on an isopycnal and depth of an isopycnal. The residuals below (8.4 $cm^2$; 3.6 $cm^2$) show the dynamic height (sea level) variance not explained by the 3 EOFs above. From Gavart and De Mey (1997). See also Haines (2003b)

This brief description of the vertical projection issue for assimilating sea level information is still a topic of considerable interest in the operational ocean forecasting community, see for example Isern-Fontanet et al. (2008). It may be with the recent advent of Argo float data across the world oceans that a better understanding of the connection between sea level variability and the subsurface fields can be achieved which would help to improve further the approaches to assimilating altimeter data. We look at issues associated with assimilating in situ hydrographic data in the next section.

## 6 In Situ Temperature and Salinity Assimilation

In this section we look at the assimilation of subsurface temperature and salinity profile data. Particular attention is paid to covariances between temperature and salinity. In the past, most subsurface data has consisted of temperature ($T$) profiles (as a function of depth, $z$) only without coincident salinity; however, the Argo float programme now provides regular salinity measurements. As discussed in Haines (2003b), the vast majority of $T$ profile data from Expendable Bathythermographs (XBTs) or from moorings, tend to be of limited depth and focused on the upper ocean heat content. The TAO mooring data in particular are the main resource for ocean assimilation for seasonal forecasting activities and we shall illustrate the methods used by reference to results from the ECMWF seasonal forecasting system.

The starting point for $T$ profile assimilation at ECMWF is an optimal interpolation (OI) method. Observed $T_o(z)$ profiles are compared with model $T_m(z)$ profiles. The misfits $(T_o–T_m)(z)$ are then spread out over some influence radius, with some weighting, and the calculated innovations are added to the model fields, slowly over a period of days to reduce the assimilation shocks. The details are not important but it is important to note that early schemes (a) made no change to the temperature below the deepest observed $T_o(z)$ (only 450 m for Tropical Atmosphere-Ocean, TAO data) and (b) made no updates to the salinity field. The consequences of these omissions have been shown to be quite severe. Of course, if sufficient observations existed, covariance matrices could perhaps be used to update the salinity and deeper temperature fields, but these data are not available.

Troccoli and Haines (1999) offered an alternative method of updating the deeper temperatures in a model, as well as a way of updating the salinity, that has now been adopted at ECMWF. Figure 13 illustrates the problem clearly. Unstable density profiles can easily be created at the base of an observed $T$ profile, or even within the range of the observed $T$ profile, by a standard univariate $T$ assimilation method. The solution suggested is to vertically displace the model water column to ensure a temperature match at the deepest level of the temperature analysis. In addition, within the upper ocean the salinity is modified to preserve the $T/S$ relationships present in the model water column. These two constraints ensure that the final analysed water column is continuous in $T$ and $S$ at all levels and is also guaranteed to be statically stable (see Fig. 13). This scheme was incorporated into the ECMWF seasonal forecasting model and run to produce ocean analyses over a 10-year period. The impacts

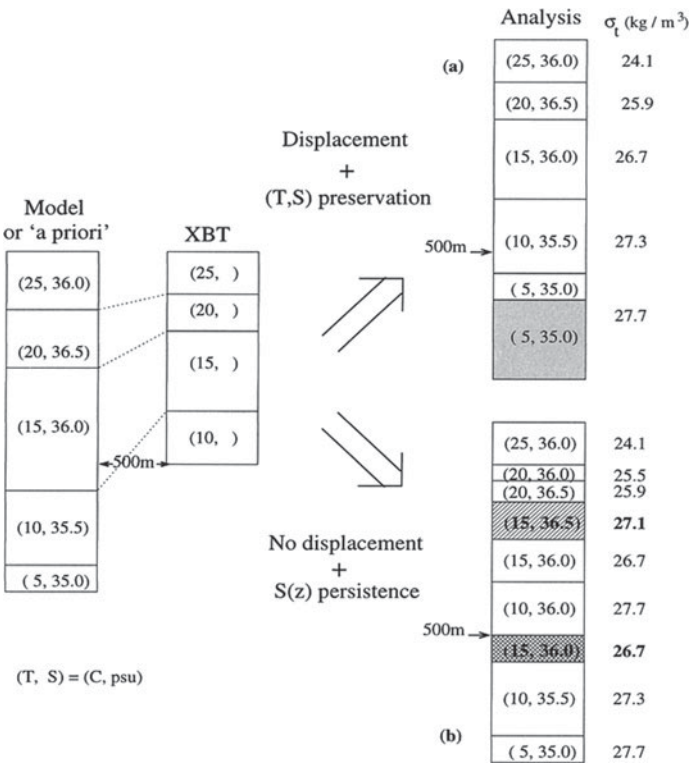



**Fig. 13** Schematic of assimilation of an observed temperature profile into an a priori model, shown on the *left*. *Bracketed terms* show (*T*, *S*) of the waters. *Lower right* shows the simplest assimilation, which directly imports the *T* data. The values to the *right* of the water columns show potential density. *Upper right* shows the analysis of the water-mass conserving scheme of Troccoli and Haines (1999). See also Haines (2003c)

on the *T* and *S* fields in the tropical Pacific are illustrated in Fig. 2 from Troccoli et al. (2002). The new method appears to reduce the strong mixing of the temperature and salinity fields that otherwise occur around the Equator. There is also some evidence that this allows some improvements in the ability of the coupled model to make El Niño forecasts (Segschneider et al. 2001).

The relationship between salinity and temperature can, of course, conventionally be represented in terms of a *T*/*S* covariance function which itself would be a function of depth. However the method described above can also be seen as another example of the use of alternative state vectors to represent the variability in observations. Troccoli and Haines (1999) illustrated this in a study of variances from an intensive conductivity-temperature-depth (CTD) campaign undertaken in a small region of the tropical Pacific. Figure 14 shows data from 104 *T* and *S* profiles taken over a period of 10 days. Clearly, there is considerable variability in both the *T* and *S* profiles as a function of depth, *z*. However the *S*(*T*) relationship shows less variance and if the salinity variance is measured as a function of temperature and projected

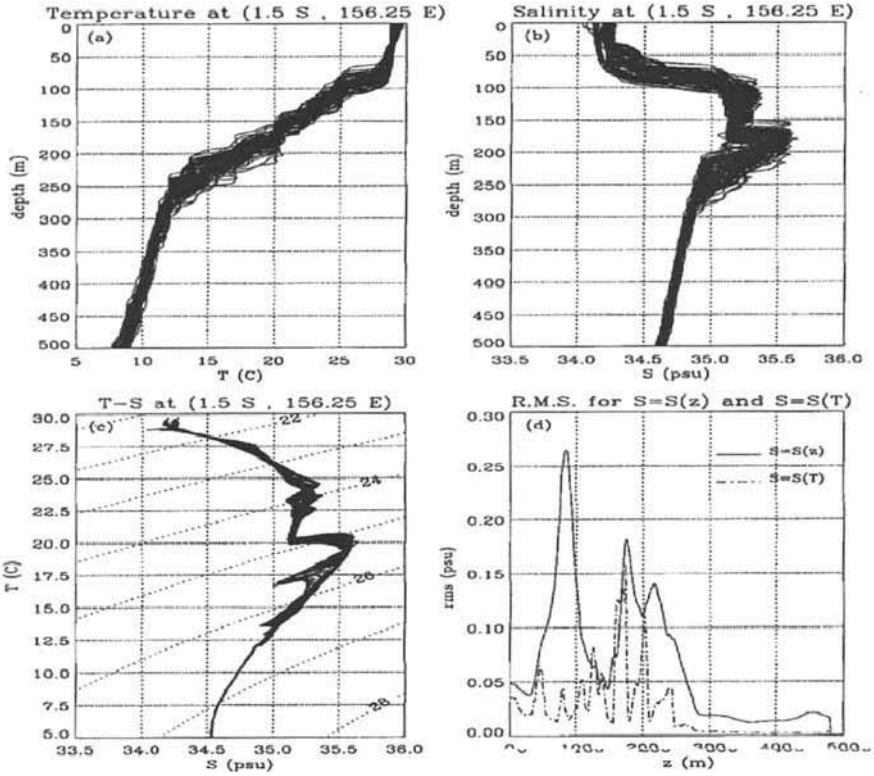



**Fig. 14** Comparison of 104 CTD (conductivity-temperature-depth) profiles from a region of the western tropical Pacific taken over 10 days in 1992. *Top left*: temperature profiles; *top right*: salinity profiles; *bottom left*: *T*-*S* diagram; *bottom right*: $S(z)$ and $S(T)$ root-mean-square (RMS) variance comparison for all profiles. Note that to plot $S(T)$ variance as a function of $z$ in *bottom right*, the mean $T$ profile from *top left* has been used. From Troccoli and Haines (1999). See also Haines (2003c)

back to depth levels using the average $T(z)$ profile then the effective reduction of variance is very clearly seen.

The result of the above assimilation method is to preserve the $T/S$ relation during assimilation of $T$ profiles. This can be regarded as having a similar justification to the preservation of the deeper potential vorticity fields during altimeter assimilation discussed in Sect. 5. Both $T$ and $S$ are Lagrangian conserved properties of the water and are therefore likely to vary together whenever the cause of variation is associated with advection. Most wave processes in the ocean, whether internal waves or Rossby waves, produce variability of this type. The $T/S$ properties are therefore preserved over long time-scales and only change in the ocean interior due to mixing processes. This also applies to an ocean model in which assimilation is performed using these constraints in the assimilation schemes.

These adjustments to the salinity field during temperature profile assimilation were incorporated into a more general assimilation framework by Ricci et al. (2005),

and the methods can also be seen as a particular oceanographic example of the use of balanced and unbalanced variables within the assimilation procedure (Derber and Bouttier 1999; Weaver et al. 2005). Recently, the Troccoli and Haines (1999) scheme was extended to include an improved method for assimilating salinity observations (see Haines et al. 2006; Smith and Haines 2009). Whereas salinity balancing increments ensure that the $T/S$ water mass properties are not altered when only temperature data are available, the Haines et al. (2006) method advocates assimilation of salinity observations *along isotherms*, i.e., directly assimilating innovations to $S(T)$ between model and observations. This ensures that the new salinity increments are orthogonal, and additive, to the balancing salinity increments. Haines et al. (2006) also argue that larger space/time covariance scales can be used to give a greater impact on the ocean state analysis.

A further issue of relevance to hydrographic data assimilation is the treatment of model bias. We will not treat this in detail here but important discussions and methodology can be found in Bell et al. (2004), Balmaseda et al. (2007) and Chepurin et al. (2005). These papers deal with hydrography bias issues, mainly in the tropical Pacific, in the context of operational oceanography, seasonal forecasting, and ocean reanalysis, respectively.

## 7 Future Prospects for Ocean Data Assimilation

These are exciting times for ocean data assimilation. The global observing network for the ocean is in place now; models and computers are becoming more fit for purpose. The challenges for ocean data assimilation include: the development of operational oceanography; pushing the limits of medium-range forecasting, seasonal to decadal forecasting and ocean reanalyses; learning to improve ocean models on the basis of data assimilation (there is a better chance of this in the oceans because of the longer time-scales of thermodynamic processes). Data assimilation will be crucial for linking observations that are separated spatially and temporally over long periods. The oceans will also play an increasing role in longer range weather and climate forecasting based on initial conditions. This will require efforts to improve assimilation into coupled atmosphere-ocean models and efforts to maintain a good observing system. Finally, ocean data assimilation will play an important role in monitoring climate change (see, e.g., the Rapid Climate change, RAPID, programme; http://www.noc.soton.ac.uk/rapid/rapid.php).

## References

ASOF, Arctic-Subarctic Ocean Fluxes, 2008. Dickson, R.R., J. Meincke and P. Rhines (eds.), Springer, 736pp.

Awaji, T., N. Sugiura, T. Mochizuki, S. Masuda, T. Miyama, H. Igarashi, Y. Ishikawa, K. Horiuchi, Y. Sasaki, Y. Hiyoshi and N. Komori, 2002. Research development of 4-dimensional data assimilation system using a coupled climate model and construction of reanalysis datasets for initialization. In *Chapter 4 Research Revolution 2002: Research Project for Sustainable Coexistence of Human, Nature, and the Earth*. Annual Report of the Earth Simulator Center April 2006–March 2007.

Balmaseda, M.A., D. Dee, A. Vidard and D.L.T. Anderson, 2007. A multivariate treatment of bias for sequential data assimilation: Application to the tropical oceans. *Q. J. R. Meteorol. Soc.*, **133**, 167–179.

Balmaseda, M.A., G.C. Smith, K. Haines, D. Anderson, T.N. Palmer and A. Vidard, 2007. Historical reconstruction of the Atlantic meridional overturning circulation from ECMWF operational ocean reanalysis. *Geophys. Res. Lett.*, **34**, L23615, doi:10.1029/2007GL031645.

Barnston, A.G., Y. He and M.H. Glantz, 1999. Predictive skill of statistical and dynamical climate models in SST forecasts during the 1997–1998 El Niño episode and the 1998 La Niña onset. *Bull. Amer. Meteorol. Soc*., **80**, 217–243.

Battle, M., M.L. Bender, P.P. Tans, J.W.C. White, J.T. Ellis, T. Conway and R.J. Francey, 2000. Global carbon sinks and their variability inferred from atmospheric $O_2$ and $\delta^{13}C$. *Science*, **287**, 2467–2470.

Bell, M.J., M.J. Martin and N.K. Nichols, 2004. Assimilation of data into an ocean model with systematic errors near the equator. *Q. J. R. Meteorol. Soc.*, **130**, 873–893.

Bennett, A., 1992. *Inverse Methods in Physical Oceanography*. Cambridge Monographs on Mechanics and Applied Mathematics, Cambridge University Press, Cambridge, UK, 346pp.

Bingham R.J., K. Haines and C.W. Hughes, 2008. Calculating the ocean's mean dynamic topography from a mean sea surface and a Geoid. *J. Atmos. Ocean Tech.*, **25**, 1808–1872.

Blumberg, A.F. and G.L. Mellor, 1987. A description of a three-dimensional coastal ocean circulation model. In *Three-Dimensional Coastal Ocean Models*, Heaps, N. (ed.), American Geophysical Union, Washington, DC, 208pp.

Bryan, K., 1963. A numerical investigation of a nonlinear model of the wind-driven ocean. *J. Atmos. Sci.*, **20**, 594–606.

Chepurin G.A., J.A. Carton and D. Dee, 2005. Forecast model bias correction in ocean data assimilation. *Mon. Weather Rev.*, **133**, 1328–1342.

Conkwright, M.E., R.A. Locarnini, H.E. Garcia, T.D. O'Brien, T.P. Boyer, C. Stephens and J.I. Antonov, 2002. *World Ocean Atlas 2001: Objective Analyses, Data Statistics and Figures, CD-Rom Documentation*, National Oceanographic Data Center, Silver Springs, MD, 17pp.

Cooper, M.C. and K. Haines, 1996. Data assimilation with water property conservation. *J. Geophys. Res*., **101**, 1059–1077.

Cunningham, S.A., T. Kanzow, D. Rayner, et al., 2007. Temporal variability of the Atlantic meridional overturning circulation at 26.5 N. *Science*, **317**, 935–938.

Dee, D.P., 2005. Bias and data assimilation. *Q. J. R. Meteorol. Soc.*, **613**, 3323–3343.

Dee, D. and A. da Silva, 1998. Data assimilation in the presence of forecast bias. *Q. J. R. Meteorol. Soc*., **124**, 269–295.

De Mey, P. and A. Robinson, 1987. Assimilation of altimeter eddy fields in a limited area quasi-geostrophic model. *J. Phys. Oceanogr.*, **17**, 2280–2293.

Derber, J. and F. Bouttier, 1999. A reformulation of the background error covariance in the ECMWF global data assimilation system. *Tellus*, **51A**, 195–221.

Drecourt, J., K. Haines and M. Martin, 2006. Influence of systematic error correction on the temporal behavior of an ocean model. *J. Geophys. Res.*, **111**, C11020, doi:10.1029/2006JC003513.

Drinkwater M.R., R. Haagmans, D. Muzi, A. Popescu, R. Floberghagen, M. Kern and M. Fehringer, 2007. *Proceedings of 3rd International GOCE User Workshop*, 6–8 November, 2006, Frascati, Italy, ESA SP-627.

Durand, F., L. Gourdeau, T. Delcroix and J. Verron, 2002. Assimilation of sea surface salinity in a tropical Oceanic General Circulation Model (OGCM): A twin experiment approach. *J. Geophys. Res.*, **107**, 8004, doi: 10.1029/2001JC000849.

Eden, C. and Oschlies, A., 2006. Adiabatic reduction of circulation-related $CO_2$ air-sea flux biases in a North Atlantic carbon-cycle model. *Global Biogeochem. Cycles*, **20**, GB2008, doi: 10.1029/2005GB002521.

Eymard, L., S. Planton, P. Durand, et al., 1996. Study of the air-sea interactions at the meso-scale: The SEMAPHORE experiment. *Ann. Geophys*., **14**, 968–1015.

Ezer, T. and G.L. Mellor, 1994. Continuous assimilation of GEOSAT altimeter data into a three-dimensional primitive equation Gulf Stream model. *J. Phys. Oceanogr*., **24**, 832–847.

Fox, A.D. and K. Haines, 2003. Interpretation of water mass transformations diagnosed from data assimilation. *J. Phys. Oceanogr*., **33**, 485–498.

Fox, A.D., K. Haines, B. De Cuevas and D.J. Webb, 2000. Altimeter assimilation in the OCCAM global model, Part II: TOPEX/POSEIDON and ERS1 data. *J. Marine Sys*., **26**, 323–347.

Gavart, M. and P. De Mey, 1997. Isopycnal EOFs in the Azores current region: A statistical tool for dynamical analysis and data assimilation. *J. Phys. Oceanogr*., **27**, 2146–2157.

Gill, A.E., 1982. *Atmosphere-Ocean Dynamics,* Academic Press, New York, 662pp.

Goddard, L. and S.G.H. Philander, 2000. The energetics of El Niño and La Niña. *J. Climate*, **13**,1496–1516.

Haines, K., 1991. A direct method for assimilating sea surface height data into ocean models with adjustments to the deep circulation. *J. Phys. Oceanogr.,* **21**, 843–868.

Haines, K., 1994. Dynamics and data assimilation in oceanography. *NATO Series I,* **19**, 1–32.

Haines, K., 2003a. Uses of ocean data assimilation and ocean state estimation. In *Data Assimilation for the Earth System*, NATO Science Series: IV. Earth and Environmental Sciences 26, Swinbank, R., V. Shutyaev and W.A. Lahoz (eds.), Kluwer Academic Publishers, Dordrecht, The Netherlands, pp 289–296, 378pp.

Haines, K., 2003b. Altimeter covariances and errors treatment. In *Data Assimilation for the Earth System*, NATO Science Series: IV. Earth and Environmental Sciences 26, Swinbank, R., V. Shutyaev and W.A. Lahoz (eds.), Kluwer Academic Publishers, Dordrecht, The Netherlands, pp 297–308, 378pp.

Haines, K., 2003c. Assimilation of hydrographic data and analysis of model bias. In *Data Assimilation for the Earth System.* NATO Science Series: IV. Earth and Environmental Sciences 26, Swinbank, R., V. Shutyaev and W.A. Lahoz, Kluwer Academic Publishers, Dordecht, The Netherlands, pp 309–320, 378pp.

Haines, K., J. Blower, J-P. Drecourt, C. Liu, A. Vidard, I. Astin and X. Zhou., 2006. Salinity assimilation using S(T): Covariance relationships. *Mon. Weather Rev.*, **134**, 759–771.

Hernandez, F. and P. Shaeffer, 2000. Altimetric mean sea surfaces and gravity anomaly maps intercomparisons. AVI-NT-011-5242-CLS 48pp, CLS, Ramonville St. Agnes.

Holland, W.R., 1978. The role of mesoscale eddies in the general circulation of the ocean – numerical experiments using a wind driven quasi-geostrophic model. *J. Phys. Oceanogr*., **8**, 363–392.

Hughes, C. and R. Bingham, 2008. An oceanographer′s guide to GOCE and the Geoid. *Ocean Sci*., **4**, 15–29.

Hunegnaw, A., F. Siegismund, R. Hipkin and K.A. Mork (2009), Absolute flow field estimation for the Nordic seas from combined gravimetric, altimetric, and in situ data. *J. Geophys. Res*., **114**, C02022, doi: 10.1029/2008JC004797.

Hurlburt H.E., E.P. Chassignet, J.A. Cummings, A.B. Kara, E.J. Metzger, J.F. Shriver, O.M. Smedstad, A.J. Wallcraft and C.N. Barron, 2008. Eddy-resolving global ocean prediction. In *Eddy-Resolving Ocean Modeling,* AGU Monograph Series, Hecht, M. and H. Hasumi (eds.), American Geophysical Union, Washington, DC, pp 353–382.

Hurlburt, H.E., D.N. Fox and E.J. Metzger, 1990. Statistical inference of weakly correlated subthermocline fields from satellite altimeter data. *J. Geophys. Res*., **95**, 11375–11409.

IPCC, 2007. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change, 2007. Solomon, S., D. Qin, M. Manning, Z. Chen, M. Marquis, K.B. Averyt, M. Tignor and H.L. Miller (eds.), Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA. Available from http://www.ipcc.ch.

Isern-Fontanet, J., G. Lapeyre, P. Klein, B. Chapron and M.W. Hecht (2008), Three-dimensional reconstruction of oceanic mesoscale currents from surface information. *J. Geophys. Res*., **113**, C09005, doi: 10.1029/2007JC004692.

Ishii, M., N. Hasegawa, S. Sugimoto, I. Ishikawa, I. Yoshikawa and M. Kimoto, 1998. An El Niño prediction experiment with a JMA ocean-atmosphere coupled model, "Kookai". *Proceedings of*

*WMO International Workshop on Dynamical Extended Range Forecasting*, Toulouse, France, 17–21 November 1997, WMO/TD-No. 881, pp 105–108.

Killworth, P.D., D.E. Dietrich, Ch. Le Provost, A. Oschlies and J. Willebrand, 2001. Assimilation of altimetric data into an eddy-permitting model of the North Atlantic. *Prog. Oceanogr*., **48**, 313–335.

Koblinsky, C.J. and N.R. Smith (eds.), 2001. *Observing the Oceans in the 21st Century*, GODAE Project Office, Bureau of Meteorology, Melbourne, Australia, 285–306. ISBN-0642-70618-2.

Komen, G.J., L. Cavaleri, M. Donelan, K. Hasselmann and P.A.E.M. Janssen, 1994. *Dynamics and Modelling of Ocean Waves,* Cambridge University Press, Cambridge, UK, 532pp.

Lea, D.J., M.R. Allen and T.W.N. Haine, 2000. Sensitivity analysis of the climate of a chaotic system. *Tellus,* **52A**, 523–532.

Lea, D.J., J-P. Drecourt, K. Haines and M. Martin, 2008. Ocean altimeter assimilation with observational and model bias correction. *Q. J. R. Meteorol. Soc*., **134**, 1761–1774.

Mansfield, D.A., 1986. The skill of dynamical long-range forecasts, including the effect of SST anomalies. *Q. J. R. Meteorol. Soc*., **112**, 1145–1176.

Marshall, J.C., 1984. Eddy-mean flow interaction in a barotropic ocean model. *Q. J. R. Meteorol. Soc*., **110**, 573–590.

Mellor, G.L., 1996. *Users Guide for a Three-Dimensional, Primitive Equation, Numerical Ocean Model*, Program in Atmospheric and Oceanic Sciences, Princeton University, Princeton, NJ, 38pp.

Mellor G.L. and T. Ezer, 1991. A Gulf stream model and an altimetry assimilation scheme. *J. Geophys. Res*., **96**, 8779–8795.

Merchant C.J., D. Llewellyn-Jones, R.W. Saunders, N.A. Rayner, E.C. Kent, C.P. Old, D. Berry, A.R. Birks, T. Blackmore, G.K. Corlett, O. Embury, V.L. Jay, J. Kennedy, C.T. Mutlow, T.J. Nightingale, A.G. O'Carroll, M.J. Pritchard, J.J. Remedios and S. Tett, 2008. Deriving a sea surface temperature record suitable for climate change research from the along-track scanning radiometers. *Adv. Space Res*., **41**, 1–11, doi:10.1016/j.asr.2007.07.041.

Mochizuki, T., H. Igarashi, N. Sugiura, S. Masuda, N. Ishida and T. Awaji, 2007. Improved coupled GCM climatologies for summer monsoon onset studies over Southeast Asia. *Geophys. Res. Lett*., **34**, L01706, doi:10.1029/2006GL027861.

Niiler P.P., N.A. Maximenko and J.C. McWilliams, 2003. Dynamically balanced absolute sea level of the global ocean derived from near-surface velocity observations. *Geophys. Res. Lett*., **30**, 2164–2167.

Oschlies, A. and V. Garcon, 1998. Eddy-induced enhancement of primary production in a model of the North Atlantic Ocean. *Nature*, **394**, 266–269.

Oschlies, A. and J. Willebrand, 1996. Assimilation of Geosat altimeter data into an eddy-resolving primitive equation model of the North Atlantic Ocean. *J. Geophys. Res*., **101**, 14175–14190.

Palmer, J.R. and I.J. Totterdell, 2001. Production and export in a global ecosystem model. *Deep Sea Res. I*, **48**, 1169–1198.

Palmer, T.N., A. Alessandri, U. Andersen, P. Cantelaube, M. Davey, P. Délécluse, M. Déqué, E. Díez, F.J. Doblas-Reyes, H. Feddersen, R. Graham, S. Gualdi, J.-F. Guérémy, R. Hagedorn, M. Hoshen, N. Keenlyside, M. Latif, A. Lazar, E. Maisonnave, V. Marletto, A.P. Morse, B. Orfila, P. Rogel, J.-M. Terres and M.C. Thomson, 2004. Development of a European multi-model ensemble system for seasonal to inter-annual prediction (DEMETER). *Bull. Amer. Meteorol. Soc*., **85**, 853–872.

Philander, S.G., 2002. A review of tropical ocean-atmosphere interactions. *Tellus A*, **51**, 71–90.

Piggott M.D., C.C. Pain, G.J. Gorman, D.P. Marshall and P.D. Killworth, 2008. Unstructured adaptive meshes for ocean modeling. In *Eddy-Resolving Ocean Modeling*, Hecht, M. and H. Hasumi (eds.), American Geophysical Union, Washington, DC, pp 383–408.

Reynolds, R.W. and T.M. Smith, 1994. Improved global sea surface temperature analyses using optimum interpolation. *J. Climate*, **7**, 929–948.

Ricci, S., A.T. Weaver, J. Vialard and P. Rogel, 2005. Incorporating temperature-salinity constraints in the background error covariance of variational ocean data assimilation. *Mon. Weather Rev*., **133**, 317–338.

Rio, M.-H. and F. Hernandez, 2004. High-frequency response of wind-driven currents measured by drifting buoys and altimetry over the world ocean. *J. Geophys. Res*., **108**, 3283, doi:10.1029/2002JC001655.

Segschneider J., D.L.T. Anderson, J. Vialard, M. Balmaseda, T.N. Stockdale, A. Troccoli and K. Haines, 2001. Initialization of seasonal forecasts assimilating sea level and temperature observations. *J. Climate*, **14**, 4292–4307.

Smith, D.M., S. Cusack, A.W. Colman, C.K. Folland, G.R. Harris and J.M. Murphy, 2007. Improved surface temperature prediction for the coming decade from a global climate model. *Science*, **317**, 796–799.

Smith, G. and K. Haines, 2009. Evaluation of the S(T) assimilation method with the Argo dataset. *Q. J. R. Meteorol. Soc*., **135**, 739–756.

Stammer, D., K. Ueyoshi, A. Köhl, W.G. Large, S.A. Josey and C. Wunsch, 2004. Estimating air-sea fluxes of heat, freshwater, and momentum through global ocean data assimilation, *J. Geophys. Res*., **109**, C05023, doi:10.1029/2003JC002082.

Stockdale, T., 1997. Coupled ocean-atmosphere forecasts in the presence of climate drift. *Mon. Weather Rev*., **125**, 809–818.

Stockdale, T.N., D.L.T. Anderson, J.O.S. Alves and M.A. Balmaseda, 1998. Global seasonal rainfall forecasts using a coupled ocean atmosphere model. *Nature*, **392**, 370–373.

Stommel, H., 1948. The westward intensification of wind driven ocean currents. *Trans. Amer. Geophys. Union*., **29**, 202–206.

Tapley, B., J. Ries, S. Bettadpur, D. Chambers, M. Cheng, F. Condi, B. Gunter, Z. Kang, P. Nagel, R. Pastor, T. Pekker, S. Poole and F. Wang, 2005. GGM02 – An improved Earth gravity model from GRACE. *J. Geodesy*, **79**, 467–478, doi:10.1007/s00190-005-0480-z.

Tomczak, M., 1981. A multi-parameter extension of temperature/salinity diagram techniques for the analysis of non-isopycnal mixing. *Prog. Oceanogr*., **10**, 147–171.

Troccoli A., M. Balmaseda, J. Segschneider, J. Vialard, D.L.T. Anderson, K. Haines, T. Stockdale, F. Vitart and A.D. Fox, 2002. Salinity adjustments in the presence of temperature data assimilation. *Mon. Weather Rev*., **130**, 89–102.

Troccoli, A. and K. Haines, 1999. Use of the temperature-salinity relation in a data assimilation context. *J. Atmos. Ocean Tech*., **16**, 2011–2025.

Weaver, A.T., C. Deltel, E. Machu, S. Ricci and N. Daget, 2005. A multivariate balance operator for variational ocean data assimilation. *Q. J. R. Meteorol. Soc*., **131**, 3605–3626.

Wunsch C., 1996. *The Ocean Circulation Inverse Problem,* Cambridge University Press, Cambridge, UK, 442pp.

Wust, G., 1935. Die Stratosphare des Atlantischen Ozeans. Deutsche Atlantische Exped. Meteor, 1925–1927. *Wiss. Erg., Bd*., **VI**, 1. Teil, 2. Lief., 288pp.

# Land Surface Data Assimilation

**Paul R. Houser, Gabriëlle J.M. De Lannoy, and Jeffrey P. Walker**

## 1 Introduction

Accurate knowledge of spatial and temporal land surface storages and fluxes are essential for addressing a wide range of important, socially relevant science, education, application and management issues. Improved estimates of land surface conditions are directly applicable to agriculture, ecology, civil engineering, water resources management, rainfall-runoff prediction, atmospheric process studies, climate and weather prediction, and disaster management (Houser et al. 2004).

While in situ observational networks are improving, the only practical way to observe the land surface on continental to global scales is via satellite remote sensing. Though remote sensing can make spatially comprehensive measurements of various components of the land surface system, it cannot provide information on the entire system (e.g. deep moisture stores), and the measurements represent only a snapshot in time. Land surface process models may be used to continuously predict the temporal and spatial land system variations, but these predictions are often poor, due to model initialization, parameter and forcing errors, and inadequate model physics and/or resolution.

Thus, satellite observations provide an incomplete snapshot of land surface conditions, while models provide a continuous estimate of land surface conditions subject to the model's simplifications. Therefore, an attractive prospect is to combine the strengths of land surface models and observations (and minimize the weaknesses) to provide a superior land surface state estimate. This is the goal of land surface data assimilation.

Data assimilation is the application of recursive Bayesian estimation to combine current and past data in an explicit dynamical model, using the model's prognostic equations to provide time continuity and dynamic coupling amongst the fields (see chapters in Part I, *Theory*). Land surface data assimilation aims to utilize both our

P.R. Houser (✉)
George Mason University, Fairfax, VA, USA
e-mail: phouser@gmu.edu

W. Lahoz et al. (eds.), *Data Assimilation*, DOI 10.1007/978-3-540-74703-1_21,

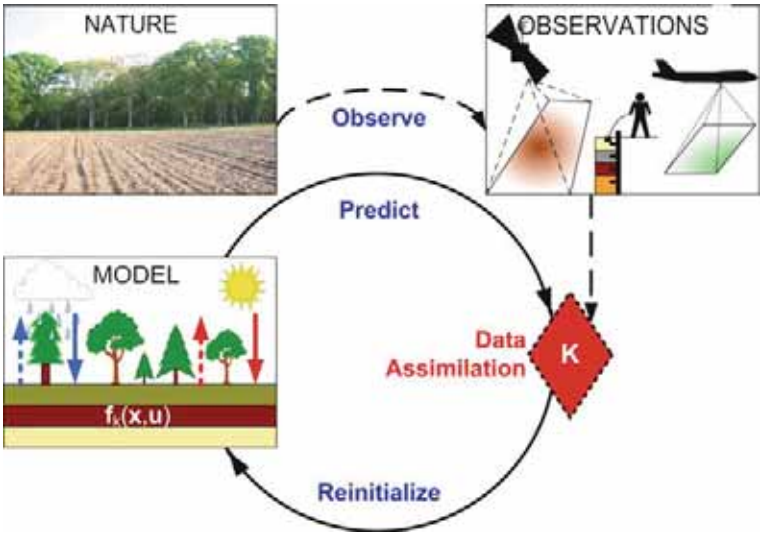



**Fig. 1** Schematic description of the land surface data assimilation process

knowledge of land surface processes as embodied in a land surface model, and information that can be gained from observations, to produce an improved, continuous land surface state estimate in space and time.

Figure 1 illustrates the land surface data assimilation challenge to optimally merge the spatially comprehensive but limited remote sensing observations with the complete but typically poor predictions of a land surface model to yield the best possible hydrological system state estimation. Limited point measurements are often used to calibrate the model(s) and validate the assimilation results (Walker and Houser 2005).

## 2 Background: Land Surface Observations

Earth observing satellites have revolutionized our understanding and prediction of the Earth system over the last 3 decades, particularly in the meteorological and oceanographic sciences. However, historically, remote sensing data have not been widely used in land surface modelling and prediction. This can be attributed to: (i) a lack of dedicated land surface state (water and energy) remote sensing instruments; (ii) inadequate retrieval algorithms for deriving global land surface information from remote sensing observations; (iii) a lack of suitable distributed land surface models for digesting remote sensing information; and (iv) an absence of techniques to objectively improve and constrain land surface model predictions using remote sensing data. Four ways that remote sensing observations have been used in distributed land surface models are: (i) as parametric input data, including soil and land cover properties; (ii) as forcing input data, mainly precipitation, (iii) as initial condition data,

such as initial snow water storage; and (iv) as time-varying land state data, such as soil moisture content, to constrain model predictions.

The historic lack of hydrological missions and observations has been the result of an emphasis on meteorological and oceanographic missions and applications, due to the large scientific and operational communities that drive those fields. However, significant progress has been made over the past decade on defining hydrologically-relevant remote sensing observations through focused ground and airborne field studies. Gradually, satellite-based hydrological data are becoming increasingly available, although little progress has been made in understanding their observational errors. Land surface skin temperature and snow cover data have been available for many years, and satellite precipitation data are becoming available at increasing space and time resolutions. In addition, land cover and land use maps, vegetation parameters (albedo, leaf area index, and greenness), and snow water equivalent data of increasing sophistication are becoming available from a number of sensors. Novel observations such as saturated fraction and changes in soil moisture, evapotranspiration, water level and velocity (i.e., runoff), and changes in total terrestrial water storage are also under development. Furthermore, near-surface soil moisture, a parameter shown to play a critical role in weather, climate, agriculture, flood, and drought processes, is currently available from non-ideal sensor configuration observations. Moreover, two missions targeted at measuring near-surface soil moisture with ideal sensor configuration are expected before the end of the decade (SMOS, SMAP; see Table 1).

**Table 1** Characteristics of hydrological observations potentially available within the next decade (see *Appendix* for details of sensor acronyms)

| Hydrological quantity | Remote sensing technique | Time scale | Spatial scale | Accuracy considerations | Examples of sensors |
|---|---|---|---|---|---|
| Precipitation | Thermal infrared | Hourly<br>1 day<br>15 days | 4 km<br>1 km<br>60 m | Tropical convective clouds only | GOES, MODIS, AVHRR, Landsat, ASTER |
| | Passive microwave | 3 h | 10 km | Land calibration problems | TRMM, SSMI, AMSR-E, GPM |
| | Active microwave | Daily | 10 m | Land calibration problems | TRMM, GPM |
| Surface soil moisture | Passive microwave | 1–3 days | 25–50 km | Limited to sparse vegetation, low topographic relief | AMSR-E, SMOS, SMAP |

**Table 1** (continued)

| Hydrological quantity | Remote sensing technique | Time scale | Spatial scale | Accuracy considerations | Examples of sensors |
|---|---|---|---|---|---|
| | Active microwave | 3 days<br>30 days | 3 km<br>10 m | Significant noise from vegetation and roughness | ERS, JERS, RadarSat |
| Surface skin temperature | Thermal infrared | 1 h<br>1 day<br>15 days | 4 km<br>1 km<br>60 m | Soil/vegetation average, cloud contamination | GOES, MODIS, AVHRR, Landsat, ASTER |
| Snow cover | Visible/thermal infrared | 1 h<br>1 day<br>15 days | 4 km<br>500 m–1 km<br>30–60 m | Cloud contamination, vegetation masking, bright soil problems | GOES, MODIS, AVHRR, Landsat, ASTER |
| Snow water equivalent | Passive microwave | 1–3 days | 10 km | Limited depth penetration | AMSR-E |
| | Active microwave | 30 days | 100 m | Limited spatial coverage | SnoSat, SCLP, Cryosat-2 |
| Water level/velocity | Laser | 10 days | 100 m | Cloud penetration problems | ICESAT, ICESAT2, SWOT, DESDynI |
| | Radar | 30 days | 1 km | Limited to large rivers | TOPEX/ POSEIDON |
| Total water storage changes | Gravity changes | 30 days | 1,000 km | Bulk water storage change | GRACE, GOCS, GRACEII |
| Evaporation | Thermal infrared | 1 h<br>1 day<br>15 days | 4 km<br>1 km<br>60 m | Significant assumptions | GOES, MODIS, AVHRR, Landsat, ASTER |

## 3 Background: Land Surface Modelling

Our knowledge about land surface processes is embedded in land surface models. Models are built upon the analysis of signals entering and leaving the system; they simulate relationships between physical variables in a natural system as a solution of mathematical structures, like simple algebraic equations or more complex systems of partial differential equations (PDEs). Land surface processes are part of the total of global processes controlling the earth, which are typically represented in global general circulation models (GCMs). The land component in these

models is represented in (largely physically-based) land surface models (LSMs), which simulate the water and energy balance over land. The major state variables of these models include the water content and temperature of soil moisture, snow and vegetation. These variables are referred to as prognostic state variables. Changes in these state variables account for fluxes, e.g., evapotranspiration, which are referred to as diagnostic. Most LSMs are soil-vegetation-atmosphere transfer (SVAT) models, where the vegetation is not a truly dynamic component. Recently, coupling of hydrological or SVAT models with vegetation models has received some attention, to serve more specific ecological, biochemical or agricultural purposes.

Most LSMs used in GCMs view the soil column as the fundamental hydrological unit, ignoring the role of, e.g., topography on spatially variable processes (Stieglitz et al. 1997) to limit the complexity and computations for these coupled models. During the last decades, LSMs were built with a higher degree of complexity in order to better represent land surface atmosphere interactions within GCMs or to meet the need for knowledge of the local state and processes in, for example, environmental or agricultural management studies. This includes, e.g., the treatment of more physiological processes, the improvement of the representation of subgrid heterogeneity and the development of distributed models. Ideally, an improved process representation (system model structure) should result in parameters that are easier to measure or estimate. However, a more complex process representation results in more parameters to be estimated and several authors (Beven 1989; Duan et al. 1992) have stated that LSMs are over-parametrized given the data typically available for calibration.

Land surface models need to be tuned to the specific circumstances under study, mainly to limit systematic prediction errors. Model calibration or parameter estimation relies on observed data and can be defined as a specific type of data assimilation. For large scale land surface modelling, full calibration is nearly impossible. For example, the National Aeronautics and Space Administration (NASA) Land Information System (LIS) allows large scale simulation of land processes with a number of land surface models, which are typically fully parametrized and forced with observation-based datasets. Some examples of widely used LSMs are the Community Land Model (CLM), the Variable Infiltration Capacity Model (VIC), the NOAH Model, the Catchment LSM, and the TOPLATS (TOPMODEL-based Land Atmosphere Transfer Scheme) model.

## 4 History of Land Surface Data Assimilation

In earth sciences, Charney et al. (1969) first suggested combining current and past data in an explicit dynamical model, using the model's prognostic equations to provide time continuity and dynamic coupling amongst the fields. This concept has evolved into a family of techniques known as data assimilation (see chapter *Mathematical Concepts of Data Assimilation*, Nichols). In essence, land surface

data assimilation aims to utilize both our hydrological process knowledge as embodied in a land surface model, and information that can be gained from observations. Both model predictions and observations are imperfect and we wish to use both synergistically to obtain a more accurate result. Moreover, both contain different kinds of information, that when used together, provide an accuracy level that cannot be obtained when used individually.

For example, a hydrological model provides both spatial and temporal near-surface and root zone soil moisture information at the model resolution, including errors resulting from inadequate model physics, parameters and forcing data. On the other hand, remote sensing observations contain near-surface soil moisture information at an instant in time, but do not give the temporal variation or the root zone moisture content. While the remote sensing observations can be used as initialization input for models or as independent evaluation, providing we use a hydrological model that has been adapted to use remote sensing data as input, we can use the hydrological model predictions and remote sensing observations together to keep the simulation on track through data assimilation (Kostov and Jackson 1993). Moreover, large errors in near-surface soil moisture content prediction are unavoidable because of its highly dynamic nature. Thus, when measured soil moisture data are available, their use to constrain the simulated data should improve the overall estimation of the soil moisture profile. However, this expectation is based on the assumption that an update in the upper layer is well propagated to deeper layers. This requires that the model correctly defines the relationship between the upper layer soil moisture and the deeper profile soil moisture (Arya et al. 1983) and that the error correlations between the soil moisture predictions in the upper layer and those in deeper layers are well captured.

Data assimilation techniques were pioneered by meteorologists (Daley 1991) and have been used very successfully to improve operational weather forecasts for decades (see chapter *Assimilation of Operational Data*, Andersson and Thépaut). Data assimilation has also been widely used in oceanography (Bennett 1992) for improving ocean dynamics prediction (see chapter *Ocean Data Assimilation*, Haines). However, hydrological data assimilation has just a small number of case studies demonstrating its utility and has very distinct features, when compared to the more chaotic atmospheric or oceanographic assimilation studies. Fortunately, we have been able to develop hydrological data assimilation by building on knowledge derived from the meteorological and oceanographic data assimilation experience, with significant advancements being made over the past decade and an increased interaction between the different earth science branches.

Progress in land surface data assimilation has been primarily limited by a lack of suitable large-domain observations. With the advent of new satellite sensors and technical advances, land surface data assimilation research directions are changing (Margulis et al. 2006). Walker et al. (2003) gave a brief history of hydrological data assimilation, focusing on the use and availability of remote sensing data, and stated that this research field in hydrology is still in its “infancy”. Walker and Houser (2005) gave an overview of hydrological data assimilation, discussing

different data assimilation methods and several case studies in hydrology. van Loon and Troch (2001) gave a review of data assimilation applications in hydrology and added a discussion on the challenges facing future hydrological applications. McLaughlin (1995) reviewed some developments in hydrological data assimilation and McLaughlin (2002) transferred the options of interpolation, smoothing and filtering for state estimation from the engineering sciences to hydrological research.

Soil moisture and soil temperature have been the most studied variables for estimation in land surface models, because of their well-known impact on weather forecasts (Zhang and Frederiksen 2003; Koster et al. 2004) and climate predictions (Dirmeyer 2000). Besides these variables, also snow mass and vegetation properties have received attention. The land surface state variables are highly variable in all three space dimensions. A complete and detailed assessment of these variables is, consequently, a difficult task. Therefore, most studies have focused on data assimilation in one or two dimensions (e.g. soil moisture profiles or single layer fields) and/or relatively simple models.

## *4.1 Early Land Surface State Estimation Studies*

The study by Jackson et al. (1981) was among the first to directly update soil moisture predictions using near-surface soil moisture observations. In this application, the soil moisture values in both layers of the United States Department of Agriculture Hydrograph Laboratory model were substituted with observed near-surface soil moisture observations as they became available. The model's performance improvement was evaluated by annual runoff values. Ottlé and Vidal-Madjar (1994) used a similar approach but with the assimilation of thermal infrared derived near-surface soil moisture content.

Another early study based on the direct insertion assimilation method was that of Bernard et al. (1981). Here, synthetic observations of near-surface soil moisture content were used to specify the surface boundary condition of a classical one-dimensional soil water diffusion model, in order to estimate the surface flux. They found that large soil moisture content variations resulting from rainy periods required special handling of the upper boundary condition. Prevot et al. (1984) repeated this study with real observations and a similar approach was used by Bruckler and Witono (1989). A more popular approach for the improved estimation of land surface fluxes has been the assimilation of screen-level measurements of relative humidity and temperature (Bouttier et al. 1993; Viterbo and Beljaars 1995).

The first known studies to use an "optimal" assimilation approach were those of Milly (1986) and Milly and Kabala (1986). In the first study, a Kalman filter (a statistical assimilation approach) was used to update a simple linear reservoir model with near-surface soil moisture observations. In the second study, an integration of models and remote sensing temperature data using an Extended Kalman filter (EKF) was

proposed. It was not until Entekhabi et al. (1994) that this approach was extended, when synthetically-derived vertical and horizontal polarized passive microwave and thermal infrared observations were assimilated into a one-dimensional soil moisture and temperature diffusion model using the Kalman filter. This synthetic study was further extended by Walker et al. (2001a, b). Since then, there has been a plethora of one-dimensional Kalman filter and variational assimilation studies.

The use of the Kalman filter for larger scale and multi-dimensional applications was early explored by Georgakakos and Baumer (1996), who used it to update a hydrological basin model with two layers of soil moisture with near-surface basin-integrated soil moisture measurements. Results showed that even when the surface observations carried substantial measurement errors, estimation of soil moisture profiles and total soil moisture storage was possible with an error that was smaller than that achieved without the use of remotely sensed data. Houser et al. (1998) was the first detailed study of several alternative assimilation approaches in a distributed model set-up, including direct insertion, statistical correction, Newtonian nudging and optimal interpolation. Both the Newtonian nudging and optimal interpolation approaches, pathological cases of the Kalman filter, showed the greatest improvement. Walker et al. (2002a) were among the first to use a three-dimensional Kalman filter based assimilation in a small catchment distributed hydrological model, while Reichle and McLaughlin (2001) were at the cutting edge with an advanced four-dimensional "optimal" variational assimilation algorithm, which included a radiative transfer model to directly include remotely sensed brightness temperature.

## *4.2 Data Assimilation Beyond State Estimation*

So far in our discussion, data assimilation was meant for state estimation, but we stress that this term can be used for any use or assimilation of observational information for model updating (WMO 1992). Basically, there are four methods for "model updating", depending on what factor is considered to be responsible for the discrepancy between observed and modelled variables:

- *Input updating*: if model input is erroneous or incorrectly defined, then corrections (e.g. through reanalysis) of the input can improve the model accuracy (improvement of the input forcing);
- *State updating*: if the model suffers from deficiencies because of a bad state initialization then one could alter the state of the model so that it comes closer to the observations (state estimation, data assimilation in the narrow sense);
- *Parameter updating*: if the model suffers from deficiencies because of an inefficient parameter choice, one could change the parameters to better adjust the models to the current information (parameter estimation, calibration);
- *Error correction*: sometimes, the model output should be corrected by an integrated error term in order to approach the observations (e.g. bias correction).

State updating can be justified by lack of knowledge about the initial conditions for a model, but with unconstrained state updating, the logic of models is foregone, while this is exactly the main strength of dynamic assimilation and modelling. If an intensive update of the state is needed for good results, the model may simply not be able to produce correct state or flux values. In such cases, assimilation for parameter estimation is better advised. The static parameters obtained through off-line calibration, prior to the actual forecast simulations, may not always result in a proper model definition, because of the state and time dependency of parameters or problems in the model structure or input. Often the model validation residuals show the presence of bias, variation in error (heteroscedasticity) and a correlation structure. Several papers reported the use of filtering techniques for parameter estimation (e.g. Katul et al. 1993; Chen and Zhang 2006). Likely, a combined state and parameter estimation (Thiemann et al. 2001) opens most perspectives for good model simulations. Two options can be considered for such an approach: (i) joint estimation of state and parameters, where the state vector is augmented with a parameter vector (Bras and Rodriguez-Iturbe 1985; Evensen 2003), or the objective function for parameter optimization is extended for state estimation (De Lannoy et al. 2006); and (ii) dual estimation, using two interactive filters or optimization procedures (Hebson and Wood 1985; Moradkhani et al. 2005; Gove and Hollinger 2006; Vrugt et al. 2006). The chapter *Inverse Modelling and Combined State-Source Estimation for Chemical Weather* (Elbern et al.) discusses these ideas in the context of chemical data assimilation.

Another option is to estimate the forecast bias, as an integrated value for all errors in the parameters, the forcings and the model structure along with the state estimation, as originally presented by Friedland (1969) and Dee and da Silva (1998). Among the first studies on forecast bias estimation in land surface models were Bosilovich et al. (2007) for skin temperature assimilation and De Lannoy et al. (2007a, b) for soil moisture data assimilation.

In the remainder of this chapter, we mainly limit data assimilation to state estimation.

## 5 General Concept of Land Surface Data Assimilation

The data assimilation challenge is: given a (noisy) model of the system dynamics, find the best estimates of system states $\hat{\mathbf{x}}$ from (noisy) observations $\mathbf{y}$. Most current approaches to this problem are derived from either the direct observer (i.e., sequential filter) or dynamic observer (i.e., variational through time) techniques. Figure 2 illustrates schematically the key differences between these two approaches to data assimilation. To help the reader through the large amount of jargon typically associated with data assimilation, a list of terminology has been provided (Table 2). We adopt the convention of lowercase bold symbols for vectors and uppercase bold symbols for matrices. Non-linear operators are in bold Kunstler script; their linearization is represented as for a matrix. This section complements material in the chapters in Part I, *Theory*.

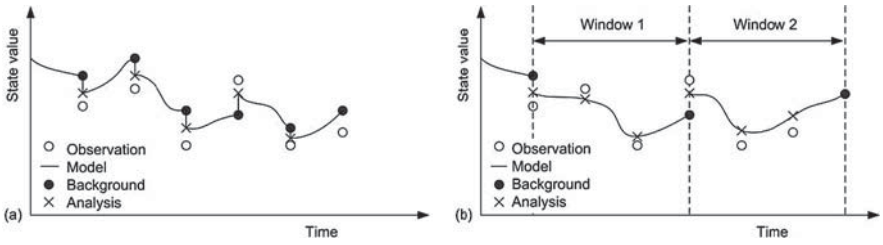



**Fig. 2** Schematic of the (**a**) direct observer and (**b**) dynamic observer assimilation approaches

**Table 2** Commonly used data assimilation terminology

| | |
|---|---|
| State | Condition of a physical system, e.g. soil moisture |
| State error | Deviation of the estimated state from the truth |
| Prognostic | A model state required to propagate the model forward in time |
| Diagnostic | A model state/flux diagnosed from the prognostic states – not required to propagate the model |
| Observation | Measurement of a model diagnostic or prognostic |
| Covariance matrix | Describes the uncertainty in terms of standard deviations and correlations |
| Prediction | Model estimate of states |
| Update | Correction to a model prediction using observations |
| Background | Forecast, prediction or state estimate prior to an update |
| Analysis | State estimate after an update |
| Innovation | Observation-minus-prediction, a priori residual |
| Gain matrix | Correction factor applied to the innovation |
| Tangent linear model | Linearized (using Taylor's series expansion) version of a non-linear model |
| Adjoint | Operator allowing the model to be run backwards in time |

## *5.1 Direct Observer Assimilation*

The direct observer techniques sequentially update the model forecast $\hat{\mathbf{x}}_k^b$ (a priori simulation result), using the difference between observation $\mathbf{y}_k$ and model predicted observation $\hat{\mathbf{y}}_k$, known as the "innovation", whenever observations are available. The predicted observation is calculated from the model predicted or "background" states, indicated by the superscript *b*. The correction, or analysis increment, added to the background state vector is the innovation multiplied by a weighting factor or gain **K**. The resulting estimate of the state vector is known as the "analysis", as indicated by the superscript *a*.

$$\hat{\mathbf{x}}_k^a = \hat{\mathbf{x}}_k^b + \mathbf{K}_k \left( \mathbf{y}_k - \hat{\mathbf{y}}_k \right) \tag{1}$$

The subscript *k* refers to the time of the update. For particular assimilation techniques, like the Kalman filter, the gain represents the relative uncertainty in the

observation and model variances, and is a number between 0 and 1 in the scalar case. If the uncertainty of the predicted observation (as calculated from the background states and their uncertainty) is large relative to the uncertainty of the actual observation, then the analysis state vector takes on values that will closely yield the actual observation. Conversely, if the uncertainty of the predicted observation is small relative to the uncertainty of the actual observation, then the analysis state vector is unchanged from the original background value. The commonly used direct observer methods are: (i) direct insertion; (ii) statistical sorrection; (iii) successive correction; (iv) analysis correction; (v) nudging; (vi) optimal interpolation/statistical interpolation; (vii) 3-D variational, 3D-Var; and (viii) Kalman filter and variants.

While approaches like direct insertion, nudging and optimal interpolation are computationally efficient and easy to implement, the updates do not account for observation uncertainty or utilize system dynamics in estimating model background state uncertainty, and information on estimation uncertainty is limited. The Kalman filter, while computationally demanding in its pure form, can be adapted for near-real-time application and provides information on estimation uncertainty. However, it has only limited capability to deal with different types of model errors, and necessary linearization approximations can lead to unstable solutions. The Ensemble Kalman filter (EnKF), while it can be computationally demanding (depending on the size of the ensemble) is well suited for near-real-time applications without any need for linearization, is robust, very flexible and easy to use, and is able to accommodate a wide range of model error descriptions.

## *5.2 Dynamic Observer Assimilation*

The dynamic observer techniques find the best fit between the forecast model state and the observations, subject to the initial state vector uncertainty $\mathbf{P}_0^b$ and observation uncertainty $\mathbf{R}$, by minimizing over space and time an objective or penalty function $J$, including a background and observation penalty term, such as

$$J(\mathbf{x}_0) = {}^1\!/_2 \left(\mathbf{x}_0 - \hat{\mathbf{x}}_0^b\right)^T \mathbf{P}_0^{b^{-1}} \left(\mathbf{x}_0 - \hat{\mathbf{x}}_0^b\right) + {}^1\!/_2 \sum_{k=0}^{N-1} \left(\mathbf{y}_k - \mathbf{y}_k^0\right)^T \mathbf{R}_k^{-1} \left(\mathbf{y}_k - \mathbf{y}_k^0\right), \quad (2)$$

where the superscript $b$ refers to the initial or "background" estimate of the state vector, the subscript $k$ refers to time, $N$ is the number of time steps, and $T$ denotes the transpose. The term $\mathbf{y}_k^0$ in the observation penalty is based on the result of propagating the state guess $\mathbf{x}_0$ to future time steps: for a particular estimated state realization $\hat{\mathbf{x}}_0^a$, $\mathbf{y}_k^0$ becomes $\hat{\mathbf{y}}_k$. To minimize the objective function over time, an assimilation time "window" is defined and an "adjoint" model is typically used to find the derivatives of the objective function with respect to the initial model state vector $\mathbf{x}_0$. The adjoint is a mathematical operator that allows one to determine the sensitivity of the objective function to changes in the solution of the state equations by a single forward and backward pass over the assimilation window. While

an adjoint is not strictly required (i.e., a number of forward passes can be used to numerically approximate the objective function derivatives with respect to each state), it makes the problem computationally tractable. The dynamic observer techniques can be considered simply as an optimization or calibration problem, where the state vector – not the model parameters – at the beginning of each assimilation window is "calibrated" to the observations over that time period.

The dynamic observer techniques can be formulated with: (i) strong constraint (variational); (ii) weak constraint (dual variational or representer methods). Strong constraint is where the model is assumed perfect, as in Eq. (2), while weak constraint is where errors in the model formulation are taken into account as process noise. This is achieved by including an additional term in Eq. (2) so that

$$J(\mathbf{x}_0) = {}^1\!/_2 \left(\mathbf{x}_0 - \hat{\mathbf{x}}_0^b\right)^T \mathbf{P}_0^{b^{-1}} \left(\mathbf{x}_0 - \hat{\mathbf{x}}_0^b\right) + {}^1\!/_2 \sum_{k=0}^{N-1} \left(\mathbf{y}_k - \mathbf{y}_k^0\right)^T \mathbf{R}_k^{-1} \left(\mathbf{y}_k - \mathbf{y}_k^0\right) + {}^1\!/_2 \sum_{k=0}^{N-1} \boldsymbol{\eta}_k^T \mathbf{Q}_k^{-1} \boldsymbol{\eta}_k \quad , \tag{3}$$

where $\boldsymbol{\eta}$ is the model error vector and $\mathbf{Q}$ is the model error covariance matrix.

Dynamic observer methods are well suited for smoothing problems, but provide information on estimation accuracy only at considerable computational cost. Moreover, adjoints are not available for many existing hydrological models, and the development of robust adjoint models is difficult due to the non-linear nature of hydrological processes.

## *5.3 Features of Data Assimilation*

The potential benefit of data assimilation for hydrological science is tremendous and can be summarized as follows (adapted from Rood et al. 1994). Data assimilation:

- *Organizes the data* by objectively interpolating information from the observation space to the model space. The raw observations are organized and given dynamical consistency with the model equations, thereby enhancing their usefulness;
- *Supplements the data* by constraining the model's physical equations with parsimonious observations, which can be used to estimate unobserved quantities. This allows the progress of research that would be impossible without assimilation, because it allows for a more complete understanding of the true state of a hydrological system (see Fig. 3a);
- *Complements the data* by propagating information into regions of sparse observations using either observed spatial and temporal correlations, or the physical relationships included in the model (see Fig. 3b);
- *Quality controls the data* through comparison of observations with previous forecasts to identify and eliminate spurious data. By performing this comparison

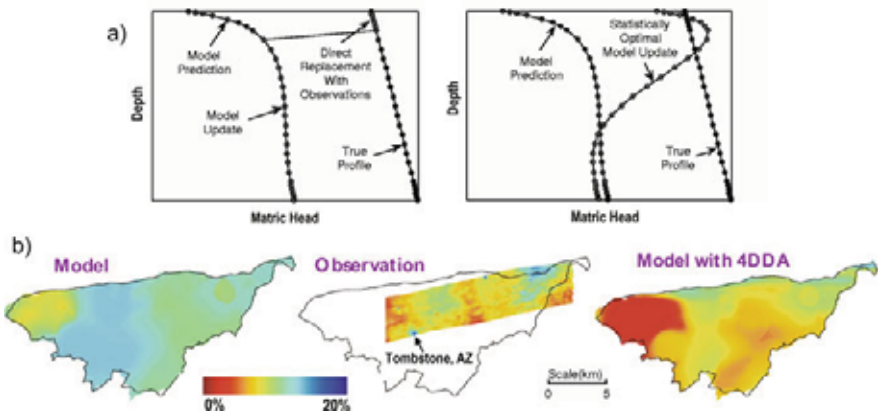



**Fig. 3** Example of how data assimilation supplements data and complements observations: (**a**) Numerical experiment results demonstrating how near-surface soil moisture measurements are used to retrieve the unobserved root zone soil moisture state using (*left panel*) direct insertion and (*right panel*) a statistical assimilation approach (Walker et al. 2001a); (**b**) Six Push Broom Microwave Radiometer (PBMR) images gathered over the USDA-ARS Walnut Gulch Experimental Watershed in Arizona were assimilated into the TOPLATS hydrological model using several alternative assimilation procedures (Houser et al. 1998). The observations were found to contain horizontal correlations with length scales of several tens of km, thus allowing soil moisture information to be advected beyond the area of the observations

repeatedly, it is possible to calibrate observing systems and identify biases or changes in observation system performance;
- *Validates and improves the hydrological models* by continuous model confrontation with real data. This helps to identify model weaknesses, such as systematic errors, and correct them.

## *5.4 Quality Control for Data Assimilation*

One of the major components of any data assimilation system is quality control of the input data stream. Quality control is a pre-assimilation rejection or correction of questionable or bad observations, which begins where the remote sensing product quality control activities end. The observational data from remote sensing systems contain errors that can be classified into two types:

- *Natural error* (including instrument and representativeness error);
- *Gross error* (including improperly calibrated instruments, incorrect registration or coding of observations, and telecommunication error).

These errors can be either random or spatially and/or temporally correlated with each other; inversion techniques and instrument biases can be correlated in time and

space, and calibrations of remote sensing instruments can drift. To address these problems a number of quality control operations are performed.

The quality control process consists of a set of algorithms which examine each data item, individually or jointly, in the context of additional information. Their primary purpose is to determine which of the data are likely to contain unknown (incorrigible) gross errors, and which are not. Quality control proceeds in a three step process: (i) test for potential problem observations; (ii) attempt to correct the problem observation; and (iii) decide the fate of the observation (data rejection). The quality control algorithms can be categorized as follows:

- *Quality control flags* are used to check the data for inconsistencies noted during the measurement, transmission, pre/post processing and archiving stages;
- *Consistency or sanity checks* see if the observation absolute value or time rate of change is physically realistic. This check filters such things as observations outside the expected range, unit conversion problems, etc.;
- *Buddy checks* compare the observation with comparable nearby (space and time) observations of the same type and reject the questioned observation if it exceeds a predefined level of difference;
- *Background checks* examine if the observation is changing similarly to the model prediction. If it is not, and the user has some reasonable confidence in the model, the observation may be questioned.

## *5.5 Validation Using Data Assimilation*

The continuous confrontation of model predictions with observations in a data assimilation system presents a rich opportunity to better understand physical processes and observational quality in a structured, iterative, and open-ended learning process. Inconsistencies between observations and predictions are easily identified in a data assimilation system, providing a basis for observational quality control and validation. Systematic differences between observations and model predictions can identify systematic errors. This methodology clearly illustrates the importance of a good quality forecast and an analysis that is reasonably faithful to the observations. If the hydrological model makes reasonably good predictions, then the analysis must only make small changes to an accurate background field.

The validation of observations in a data assimilation system is centred on: (i) comparisons of new observations with the model forecast and the data assimilation analysis; and (ii) interpretation of the forecast error covariances. The data assimilation validation algorithms can be categorized as follows:

- *Innovation evaluation* compares the observation with the model prediction as either a single point in time or change over time; large or obvious deviations from the model prediction are likely wrong. Means, standard deviations, and time evolution of observed minus predicted fields are examined with the goal of detecting abrupt changes. This allows the estimation of forecast and/or observation bias;

- *Analysis residual evaluation* compares the observation with the data assimilation analysis. Examination of the means, standard deviations, and time evolution of observed minus predicted fields will help to diagnose systematic or abrupt observation system changes. This technique is useful to diagnose the performance of the analysis, and test if the observations are being used effectively (Hollingsworth and Lönnberg 1989). Filter optimization can be achieved through adaptive filtering, using residual information;
- *Observation withholding* is a stringent method for validation in an assimilation system where some of the observational data are withheld from the analysis procedure in data-dense regions. This allows the analysis to be validated against the withheld observations;
- *Error propagation* is undertaken and changes in the regional distribution or absolute value of these errors could indicate observational problems;
- *Model and observation bias* is generally assumed to be zero and uncorrelated in space. These assumptions work reasonably well for in situ observations, but satellite observations are usually biased by inaccurate algorithms, and their errors are usually horizontally correlated because the same sensor is making all the observations. With recent work by Dee and Todling (2000) the bias of the model and observations can be continuously estimated and corrected for. Evaluation of these bias estimates in space and time may lead to additional insights on the observational characteristics.

## 6 Land Surface Data Assimilation Techniques

The text in Sect. 6 complements that in several chapters in Part I, *Theory*.

### *6.1 Land Surface System*

Land surface hydrology process models are typically non-linear, and can be considered to forecast the system state vector $\mathbf{x}$ at time $k+1$ as a function of the system state vector estimate at the previous time step $k$ and a forcing vector $\mathbf{u}$. The model state forecast is subject to a model error vector $\boldsymbol{\eta}$, which represents errors in the model forcing data, initial conditions, parameters and physics. As a result the state is a random variable and cannot be calculated as a classical solution of the deterministic system equations. The state propagation equation is given by

$$\mathbf{x}_{k+1} = \mathcal{M}_k\left(\mathbf{x}_k, \mathbf{u}_k\right) + \boldsymbol{\eta}_k, \tag{4}$$

Where $\mathcal{M}$ is a non-linear operator and $\boldsymbol{\eta}$ is assumed additive for simplicity. This equation can be linearized to obtain the "tangent linear model" as

$$\mathbf{x}_{k+1} = \mathbf{M}_k\mathbf{x}_k + \mathbf{B}_k\mathbf{u}_k + \boldsymbol{\eta}_k. \tag{5}$$

with **M** and **B** the linear state transition matrix and the linear matrix relating the input to the state. The state space equation is subject to the initial state vector

$$\hat{\mathbf{x}}_0 = \mathbf{x}_0 + \boldsymbol{\delta}_0, \tag{6}$$

which is an approximation of the truth $\mathbf{x}$ and an error vector $\boldsymbol{\delta}$ at time step $k = 0$. All subsequent forecasts $\hat{\mathbf{x}}_{k+1}^b$ (predictions, background information for data assimilation) are estimated through the model propagation by:

$$\hat{\mathbf{x}}_{k+1}^b = \mathcal{M}_k\left(\hat{\mathbf{x}}_k^a, \mathbf{u}_k\right), \tag{7}$$

with $\hat{\mathbf{x}}_k^a$ the analysis state obtained through data assimilation at the previous time step, or, if the analysis is unavailable, then $\hat{\mathbf{x}}_k^a$ is replaced by the best a priori estimate (prediction) $\hat{\mathbf{x}}_k^b$ at the previous time step.

Often the state variables are not measured directly, but some other related output from the system is observed. The observation equation is given by

$$\mathbf{y}_k = \mathcal{H}_k\left(\mathbf{x}_k\right) + \boldsymbol{\varepsilon}_k, \tag{8}$$

where $\mathcal{H}$ is a non-linear operator which relates the system state to the output observation, $\mathbf{y}$ is the actual observation and $\boldsymbol{\varepsilon}$ is an error vector (assumed additive for simplicity). This equation can also be linearized as

$$\mathbf{y}_k = \mathbf{H}_k\mathbf{x}_k + \boldsymbol{\varepsilon}_k. \tag{9}$$

The observation predictions $\hat{\mathbf{y}}_k$ are a transformation of the model forecasts to the observation space:

$$\hat{\mathbf{y}}_k = \mathcal{H}_k(\hat{\mathbf{x}}_k^b). \tag{10}$$

A typical observation system in hydrological applications is the transformation $\mathcal{H}_k$ of land surface model state variables ($\hat{\mathbf{x}}_k^b$, e.g., soil moisture) to the actual values measured by satellites ($\hat{\mathbf{y}}_k$, e.g., brightness temperature), based on radiative transfer theory.

Data assimilation aims at using the difference between the observation predictions $\hat{\mathbf{y}}_k$ and the actual observations $\mathbf{y}_k$ to update the model state. Several assimilation techniques explicitly take into account information on the error characterization. The key assumptions of the linear optimal assimilation approaches that will be discussed in this chapter, are that the error terms $\boldsymbol{\eta}$, $\boldsymbol{\delta}_0$ and $\boldsymbol{\varepsilon}$ are uncorrelated (white) through time and have zero mean Gaussian distributions as represented by their covariance matrices **Q**, $\mathbf{P}_0{}^b$ and **R**, respectively. That is

$$\begin{aligned} &\mathcal{E}\left(\boldsymbol{\eta}_k\right) = 0 \quad \mathcal{E}\left(\boldsymbol{\eta}_k\boldsymbol{\eta}_k^T\right) = \mathbf{Q}_k \\ &\mathcal{E}\left(\boldsymbol{\delta}_0\right) = 0 \quad \mathcal{E}\left(\boldsymbol{\delta}_0\boldsymbol{\delta}_0^T\right) = \mathbf{P}_0^b \\ &\mathcal{E}\left(\boldsymbol{\varepsilon}_k\right) = 0 \quad \mathcal{E}\left(\boldsymbol{\varepsilon}_k\boldsymbol{\varepsilon}_k^T\right) = \mathbf{R}_k \end{aligned} \tag{11}$$

where $\mathcal{E}(.)$ is the expectation operator. The assumption that observational and model errors are unbiased relative to each other and the "truth" is the most restrictive assumption, the most commonly violated assumption, and the most detrimental assumption in terms of predictive performance.

## *6.2 Direct Observer Data Assimilation*

One key question in the direct observer data assimilation technique, and the fundamental difference between the various methods, is the choice of the gain matrix **K** in equation

$$\hat{\mathbf{x}}_k^a = \hat{\mathbf{x}}_k^b + \mathbf{K}_k \left(\mathbf{y}_k - \hat{\mathbf{y}}_k\right) \tag{12}$$

Ultimately $\mathbf{K}_k$ should be chosen such that $\hat{\mathbf{x}}_k^a$ approaches the expectation of $\mathbf{x}_k$, as $k$ approaches infinity (as an approximation of the theoretical ensemble mean of the stochastic process). Under the assumption of perfect knowledge of the error characteristics and for linear systems, this can be achieved by choosing **K** as the optimal least squares estimator or Best Linear Unbiased Estimate (*BLUE*) analysis as used for the Kalman gain in the linear Kalman filter (see below). The optimal gain can be shown analytically to be (Jazwinski 1970; Maybeck 1979)

$$\mathbf{K}_k = \mathbf{P}_k^b \mathbf{H}_k^T \left(\mathbf{H}_k \mathbf{P}_k^b \mathbf{H}_k^T + \mathbf{R}_k\right)^{-1}, \tag{13}$$

where $\mathbf{H}_k \mathbf{P}_k^b \mathbf{H}_k^T = \hat{\mathbf{R}}_k$ is the error covariance matrix of the predicted observation $\hat{\mathbf{y}}_k$. However, approximations to the optimal filter equations and/or alternative methodologies of solving the key equations have been sought to limit some difficulties in the practical numerical approximation of this optimal solution.

*Direct insertion*: One of the earliest and most simplistic approaches to data assimilation is direct insertion. As the name suggests, the forecast model states are directly replaced with the observations by assuming that $\mathbf{K} = \mathbf{I}$, the unity matrix. This approach makes the explicit assumption that the model is wrong (has no useful information) and that the observations are right, which both disregards important information provided by the model and preserves observational errors. The risk of this approach is that unbalanced state estimates may result, which causes model shocks: the model will attempt to restore the dynamic balance that would have existed without insertion. A further key disadvantage of this approach is that model physics are solely relied upon to propagate the information to unobserved parts of the system (Houser et al. 1998; Walker et al. 2001a).

*Statistical correction*: A derivative of the direct insertion approach is the statistical correction approach, which adjusts the mean and variance of the model states to match those of the observations. This approach assumes the model pattern is correct but contains a non-uniform bias. First, the predicted observations are scaled by

the ratio of observational field standard deviation to predicted field standard deviation. Second, the scaled predicted observational field is given a block shift by the difference between the means of the predicted observational field and the observational field (Houser et al. 1998). This approach also relies upon the model physics to propagate the information to unobserved parts of the system.

*Successive correction*: The successive corrections method (SCM) was developed by Bergthorsson and Döös (1955) and Cressman (1959), and is also known as observation nudging. The scheme begins with an a priori state estimate (background field) for an individual (scalar) variable $\hat{x}_k^b \in \hat{\mathbf{x}}_k^b$, which is successively adjusted by nearby observations in a series of scans (iterations, $n$) through the data. The analysis at time step $k$ is found by passing through the following sequence of updates:

$$\begin{cases} \hat{x}_k^{a,0} = \hat{x}_k^b \\ \hat{x}_k^{a,1} = \hat{x}_k^{a,0} \quad + \mathbf{k}_k^{1^T}(\mathbf{y}_i - \mathcal{H}_k(\hat{\mathbf{x}}_k^{a,0})) \\ \quad \vdots \\ \hat{x}_k^{a,n} = \hat{x}_k^{a,n-1} + \mathbf{k}_k^{n^T}(\mathbf{y}_i - \mathcal{H}_k(\hat{\mathbf{x}}_k^{a,n-1})) \end{cases} \tag{14}$$

with $\mathcal{H}_k(\hat{\mathbf{x}}_k^{a,n})$ the value of the state estimate at the $n$th iteration, evaluated at all observation points ($\mathcal{H}_k$ is the non-linear interpolation operator), $\mathbf{y}_i$ the vector of all observations within a predefined influence radius $R_k^n$ and $\mathbf{k}_k^n$ is a vector of weights for all observations within the predefined radius of influence. The elements $k_{j,\,k}^n \in \mathbf{k}_k^n$ ($j = 1,\ldots, m$ for all observations) are given by:

$$k_{j,\,k}^n = \frac{c_{j,k}^n}{q^2 + \sum_{j=1}^{m} c_{j,k}^n} \tag{15}$$

with $q$ an estimate of the ratio of the observation error to the background error covariance, $c_{j,k}^n$ any sort of weights. Different weighting functions could be proposed, but for the Cressman scheme, the observations are assumed to be perfect ($q^2 = 0$) and the weights are given by:

$$c_{j,k}^n = \begin{cases} \frac{R_k^{n^2} - d_j^2}{R_k^{n^2} + d_j^2} & d_j < R_k^n \\ 0 & d_j \geq R_k^n \end{cases} \tag{16}$$

with $R_k^n$ the radius of influence, which is mostly shrinking for successive iterations $n$, so that the field is corrected to larger scale features during the first iterations, and conforms to smaller scale features during later iterations; $d_j$ is the distance between the $j$th observation point and the grid point for the analysis.

For the estimation of the complete state vector $\hat{\mathbf{x}}_k^a$ (i.e., multiple grid points), the equation would be as follows for each iteration $n$:

$$\hat{\mathbf{x}}_k^{a,n} = \hat{\mathbf{x}}_k^{a,n-1} + \mathbf{K}_k^{n-1}\left(\mathbf{y}_k - \hat{\mathbf{y}}_k\right) = \hat{\mathbf{x}}_k^{a.n-1} + \mathbf{K}_k^{n-1}\left(\mathbf{y}_k - \mathcal{H}_k(\hat{\mathbf{x}}_k^{a,n-1})\right), \quad (17)$$

with **K** a matrix containing an empirically derived weighting, that takes into account the spatial distribution of observations.

The advantage of this method lies in its simplicity. However, in case of observational error or different sources (and accuracies) of observations, this scheme is not a good option for assimilation, since information on the observational accuracy is not accounted for. Mostly, this approach assumes that the observations are more accurate than model forecasts, with the observations fitted as closely as is consistent. Furthermore, the radii of influence are user-defined and should be determined by trial and error or more sophisticated methods that reduce the advantage of its simplicity. The weighting functions are empirically chosen and are not derived based on physical or statistical properties. Obviously, this method is not effective in data sparse regions. Some practical examples are discussed by Bratseth (1986) and Daley (1991).

*Analysis correction*: This is a modification to the successive correction approach that is applied consecutively to each observation $s$ from 1 to $s_f$ as in Lorenc et al. (1991). In practice, the observation update is mostly neglected and further assumptions make the update equation equivalent to that for optimal interpolation (Nichols 2001).

*Nudging*: Nudging or Newtonian relaxation consists of adding a term to the prognostic model equations that causes the solution to be gradually relaxed towards the observations. Nudging is very similar to the successive corrections technique and only differs in the fact that through the numerical model the time dimension is included. Two distinct approaches have been developed (Stauffer and Seaman 1990). In analysis nudging, the nudging term for a given variable is proportional to the difference between the model simulation at a given grid point and an "analysis" of observations (i.e., processed observations) calculated at the corresponding grid point. For observation nudging, the difference between the model simulation and the observed state is calculated at the observation locations.

The nudging approach approximates the gain matrix by the empirical function

$$\mathbf{K} \approx G\left(\mathbf{W}^T \boldsymbol{\Theta} \mathbf{W}\right)(\mathbf{W}\mathbf{I})^{-1}, \quad (18)$$

where $G$ is a nudging factor that gives the magnitude of the nudging term and has a value from 0 to 1, $\Theta$ is an observational quality factor with a value from 0 to 1, **I** is the identity matrix and **W** is a temporal and spatial weighting function, also with a value from 0 to 1. The function **W** is given by $w_{xy}w_z w_t$, where $w_{xy}$ is a horizontal weighting function (i.e., Cressman), $w_z$ is a similar vertical weighting function, and $w_t$ is a temporal weighting function. Each of these temporal/spatial weighting functions has a value from 0 to 1.

*Optimal interpolation*: The optimal interpolation (OI) approach, sometimes referred to as statistical interpolation, approximates the "optimal" solution from Eq. (12) by choosing

$$\mathbf{K}_k = \mathbf{P}'^b \mathbf{H}_k^T \left( \mathbf{H}_k \mathbf{P}'^b \mathbf{H}_k^T + \mathbf{R}_k \right)^{-1} \tag{19}$$

where $\mathbf{P}'^b$ is an approximated background covariance matrix, often with a "fixed" structure for all time steps, given by prescribed variances and a correlation function determined only by distance (Lorenc 1981). Sometimes, the variances are allowed to evolve in time, while keeping the correlation structure time-invariant.

*3-D Var*: 3D-variational assimilation directly solves the iterative minimization problem given by Eqs. (2) or (3) for $N = 1$ (Parrish and Derber 1992). The same approximation for the background covariance matrix as in the optimal interpolation approach is typically used.

*Kalman filter*: The optimal analysis state estimate $\hat{\mathbf{x}}_k^a$ for linear or linearized systems (Kalman or Extended Kalman filter, EKF) can be found through a linear update equation with a Kalman gain that aims at minimizing the analysis error (co)variance of the analysis state estimate (Kalman 1960). As indicated earlier, the optimal gain can be shown analytically to be

$$\mathbf{K}_k = \mathbf{P}_k^b \mathbf{H}_k^T \left( \mathbf{H}_k \mathbf{P}_k^b \mathbf{H}_k^T + \mathbf{R}_k \right)^{-1}, \tag{20}$$

The updated (analysis) state uncertainty (analysis error covariance) is given by:

$$\mathbf{P}_k^a = \left( \mathbf{I} - \mathbf{K}_k \mathbf{H}_k \right) \mathbf{P}^b \left( \mathbf{I} - \mathbf{K}_k \mathbf{H}_k \right)^T + \mathbf{K}_k \mathbf{R}_k \mathbf{K}_k^T, \tag{21}$$

which reduces to

$$\mathbf{P}_k^a = \left( \mathbf{I} - \mathbf{K}_k \mathbf{H}_k \right) \mathbf{P}_k^b, \tag{22}$$

if, and only if, the optimal Kalman gain is used.

The essential feature which distinguishes the family of Kalman filter approaches from more static techniques, like optimal interpolation, is the dynamic updating of the forecast (background) error covariance through time. In the traditional Kalman filter (KF) approach this is achieved by application of standard error propagation theory, using the (tangent) linear model in Eq. (5). (The only difference between the Kalman filter and the Extended Kalman filter is that the forecast model is linearized using a Taylor series expansion in the latter; the same forecast and update equations are used for each approach.). The state covariance forecast equation is

$$\mathbf{P}_{k+1}^b = \mathbf{M}_k \mathbf{P}_k^b \mathbf{M}_k^T + \mathbf{Q}_k, \tag{23}$$

where $\mathbf{M}_k$ is the linear operator from Eq. (5) and $\mathbf{Q}$ is the model error covariance matrix given in Eq. (11). Thus, the (Extended) Kalman filter requires propagation of the state covariances along with the states, which might be computationally expensive and approximative, because of the system linearization. While the approach gives an optimal analysis for the assumed statistics – see Eq. (11), the initial state

error covariance matrix $\mathbf{P}_0$ and, more seriously, the model error covariance matrix $\mathbf{Q}$ are difficult to define, and often assumed ad hoc.

Equations (1), (7), (10), (12), and (21) form the basis of the Kalman filter approach (Kalman 1960) to data assimilation. On the assimilation time interval $k \in [0, N]$, the analysis $\hat{\mathbf{x}}_k^a$ given by the Kalman filter should be equal to the converged solution obtained by the variational adjoint method at time $k = N$.

The standard Extended Kalman filter update and state covariance forecast equations can be applied directly with a non-linear state forecast model after linearization. This is achieved by numerically approximating the Jacobians $\mathbf{M}$ and $\mathbf{H}$ at each time step $k$ as required by

$$\mathbf{M}_k = \left[\begin{array}{ccc} \frac{\partial \mathcal{M}_1}{\partial x_1} & \cdots & \frac{\partial \mathcal{M}_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial \mathcal{M}_n}{\partial x_1} & \cdots & \frac{\partial \mathcal{M}_n}{\partial x_n} \end{array}\right]_{(\hat{\mathbf{x}}_k^b, \mathbf{u}_k, \mathbf{w}_k = \mathbf{0})} \approx \frac{\partial \hat{\mathbf{x}}_{k+1}^b}{\partial \hat{\mathbf{x}}_k^b} \quad \text{and} \tag{24a}$$

$$\mathbf{H}_k = \left[\begin{array}{ccc} \frac{\partial \mathcal{H}_1}{\partial x_1} & \cdots & \frac{\partial \mathcal{H}_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial \mathcal{H}_n}{\partial x_1} & \cdots & \frac{\partial \mathcal{H}_n}{\partial x_n} \end{array}\right]_{(\hat{\mathbf{x}}_k^b, \mathbf{v}_k = \mathbf{0})} \approx \frac{\partial \hat{\mathbf{y}}_k}{\partial \hat{\mathbf{x}}_k^b}. \tag{24b}$$

However, the cost of doing this is $n$+1 times the standard model run time, where $n$ is the number of state variables to be updated by the assimilation. Note that only states with significant correlation to the observation need be included in the state covariance forecast and update (Walker and Houser 2001). Walker et al. (2001b) avoided the Taylor expansion linearization by adopting a Crank-Nicholson scheme to represent the state propagation.

A further approach to estimating the state covariance matrix is the Ensemble Kalman filter (EnKF). As the name suggests, the covariances are calculated from an ensemble of state forecasts using the Monte Carlo approach rather than a single discrete forecast of covariances. In this case, $N$ ensemble members of model predicted states $\hat{\mathbf{x}}_k^b$ (each containing $n$ state variables) are stored as $\hat{\mathbf{X}}_k^b$ using different initial conditions and forcing (Turner et al. 2007), different parameters and/or models, different model error (e.g. additive/multiplicative), etc., in order to get a representative spread of state forecasts amongst the ensemble members. While this is quite straightforward, the question of what model error $\boldsymbol{\eta}$ to apply, and how, is still a major unknown. Moreover, special care is required when the number of ensembles $N$ is less than the number of observations $m$.

Using this approach, the background state covariance matrix is calculated as

$$\mathbf{P}_k^b = \frac{\left(\hat{\mathbf{X}}_k^b - \bar{\hat{\mathbf{X}}}_k^b\right)\left(\hat{\mathbf{X}}_k^b - \bar{\hat{\mathbf{X}}}_k^b\right)^T}{N-1}. \tag{25}$$

where $\bar{\hat{\mathbf{X}}}_k^b$ is a matrix with all identical columns of ensemble mean state estimates. This could then be used in Eq. (12) directly, except some mathematical techniques are typically used so only matrices of size ($n \times m$) are required (Evensen 1994; Houtekamer and Mitchell 1998). Thus, $\mathbf{P}^b$ is never calculated explicitly. Here the analysis equation for each member $j$ is presented as

$$\hat{\mathbf{x}}_{j,k}^a = \hat{\mathbf{x}}_{j,k}^b + \mathbf{B}_k^T \mathbf{b}_{j,k}, \tag{26}$$

where

$$\mathbf{B}_k^T = \mathbf{P}_k^b \mathbf{H}_k^T \tag{27a}$$

$$\mathbf{b}_{j,k} = \left(\mathbf{H}_k \mathbf{P}_k^b \mathbf{H}_k^T + \mathbf{R}_k\right)^{-1} \left(\mathbf{y}_k - \hat{\mathbf{y}}_{j,k}\right). \tag{27b}$$

By rearranging Eq. (27a) and introducing a zero mean random observation error term $\boldsymbol{\varepsilon}_k$ with covariance matrix $\mathbf{R}$ for the solution of each ensemble member $j$ (to assure sufficient spread), $\mathbf{b}$ is solved for each ensemble from

$$\left(\mathbf{H}_k \mathbf{P}_k^b \mathbf{H}_k^T + \mathbf{R}_k\right) \mathbf{b}_{j,k} = \left(\mathbf{y}_k + \boldsymbol{\varepsilon}_{j,k} - \hat{\mathbf{y}}_{j,k}\right), \tag{28}$$

where

$$\mathbf{H}_k \mathbf{P}_k^b \mathbf{H}_k^T = \frac{\mathbf{q}_k \mathbf{q}_k^T}{N-1} \tag{29}$$

and

$$\mathbf{q}_k = \mathbf{H}_k \left(\hat{\mathbf{X}}_k^b - \bar{\hat{\mathbf{X}}}_k^b\right) = \left(\hat{\mathbf{Y}}_k - \bar{\hat{\mathbf{Y}}}_k\right). \tag{30}$$

The matrix $\hat{\mathbf{Y}}_k$ contains the predicted observation vector for each of the respective ensemble members. In this case it is not necessary to solve for $\mathbf{H}$ either, and the updates are made individually to each of the ensemble members. Finally, $\mathbf{B}$ can be estimated from

$$\mathbf{B}_k^T = \frac{\left(\hat{\mathbf{X}}_k^b - \bar{\hat{\mathbf{X}}}_k^b\right)}{N-1} \mathbf{q}_k^T. \tag{31}$$

Reichle et al. (2002b) applied the Ensemble Kalman filter to the soil moisture estimation problem and found it to perform as well as the numerical Jacobian approximation approach to the Extended Kalman filter, with the distinct advantage that the error covariance propagation is better behaved in the presence of large model non-linearities. This was the case even when using only the same number of ensembles as required by the numerical approach to the Extended Kalman filter, i.e., $n$+1.

## 6.3 Dynamic Observer Assimilation Methods

*4D-Var*: In its pure form, the 4-D (3-D in space, 1-D in time) "variational" (otherwise known as Gauss-Markov) dynamic observer assimilation methods use an adjoint to efficiently compute the derivatives of the objective function $J$ with respect to each of the initial state vector values $x_0$ (see the chapter *Variational Assimilation*, Talagrand). This adjoint approach is derived by defining the Lagrangian functional $\mathcal{L}$ as the *adjoining* of the model to the cost function $J$ – Eq. (3), using Lagrange multipliers $\boldsymbol{\lambda}$

$$\mathcal{L} = J + \sum_{k=0}^{N-1} \boldsymbol{\lambda}_{k+1}^{T} \left[ \mathbf{x}_{k+1} - \mathcal{M}_k \left( \mathbf{x}_k, \mathbf{u}_k \right) \right], \tag{32}$$

where ideally the second term is zero; this term guides the state estimates within the range specified by the model dynamics. Thus the Lagrange multiplier is chosen such that $\nabla \mathcal{L} = 0$ and $\boldsymbol{\lambda}_N = 0$, yielding (i.e., backward pass)

$$\boldsymbol{\lambda}_k = \mathbf{M}_k^T \boldsymbol{\lambda}_{k+1} - \mathbf{H}_k^T \mathbf{R}_k^{-1} \left( \mathbf{y}_k - \hat{\mathbf{y}}_k \right). \tag{33}$$

The derivative of the objective function is given from the Lagrange multiplier at time zero by $-\boldsymbol{\lambda}_0^T$ (Castelli et al. 1999; Reichle and McLaughlin 2001; Reichle et al. 2001). Note that $\mathbf{M}^T$, the adjoint operator, is derived from the tangent linear model in Eq. (5), and effectively needs to be saved during the forward pass (Bouttier and Courtier 1999). Solution to the variational problem is then achieved by minimization and iteration. In practical applications the number of iterations is usually constrained to a small number. While "adjoint compilers" are available (see http://www.autodiff.com/tamc/) for automatic conversion of the non-linear forecast model into a tangent linear model, application of these is not straightforward. It is best to derive the adjoint at the same time as the model is developed.

Given a model integration with finite time interval, and assuming a perfect model, 4D-Var and the Kalman filter yield the same result at the end of the assimilation time interval. Inside the time interval, 4D-Var is more optimal, because it uses all observations at once (before and after the time step of analysis), i.e., it is a smoother. A disadvantage of sequential methods is the discontinuity in the corrections, which causes model shocks. Through variational methods, there is a larger potential for dynamically based balanced analyses, which will always be situated within the model climatology. Operational 4D-Var assumes a perfect model: no model error can be included. With the inclusion of model error, coupled equations are to be solved for minimization. Through Kalman filtering it is in general simpler to account for model error.

Both the Kalman filter and 3D/4D-Var rely on the validity of the linearity assumption. Adjoints depend on this assumption and incremental 4D-Var is even more sensitive to linearity. Uncertainty estimates via the Hessian are critically dependent on a valid linearization. Furthermore, with variational assimilation it is

more difficult to obtain an estimate of the quality of the analysis or of the state's uncertainty after updating.

In the framework of estimation theory, the goal of variational assimilation is the estimation of the conditional mode (maximum a posteriori probability) estimate, while for the Kalman filter the conditional mean (minimum variance) estimate is sought.

Hybrid assimilation methods have been explored in which a sequential method is used to produce the a priori state error or background error covariance for variational assimilation.

### *6.4 Challenges in Land Surface Data Assimilation*

In order for the "optimal" assimilation techniques to be truly optimal, the error characterization should be almost flawless. Therefore, recent studies have focused on the first and second order error characterization in land surface modelling. Typically, either model predictions or observations are biased. Studies by Reichle and Koster (2004), Bosilovich et al. (2007) and De Lannoy et al. (2007a, b) scratch the surface of how to deal with these biases in land surface modelling. The second order error characterization is of major importance to optimize the analysis result and for the propagation of information through the system. Tuning of the error covariance matrices has, therefore, gained attention with the exploration of adaptive filters in land surface modelling (Reichle et al. 2008; De Lannoy et al. 2009).

Furthermore, it is important to understand that land surface data assimilation applications are dealing with non-closure or imbalance problems, caused by external data assimilation for state estimation. In a first attempt to attack this problem, Pan and Wood (2006) developed a constrained Ensemble Kalman filter which optimally redistributes any imbalance after conventional filtering. They applied this technique over a 75,000 $km^2$ domain in the US, using the terrestrial water balance as constraint.

## 7 Assimilation of Land Surface Observations

Estimation of the land surface state has mainly been focused on the soil and snow water content and temperature. The observations used to infer state information range from direct field measurements of these quantities to more indirectly related measurements like radiances or backscatter values in remote sensing products. A few studies have also tried to assimilate state-dependent diagnostic fluxes, like discharge or remotely sensed heat fluxes. The success of assimilation of observations which are indirectly related to the state is largely dependent on a good characterization of the observation operator, $\mathcal{H}(.)$. This section refers to terminology discussed in the chapter *Mathematical Aspects of Data Assimilation* (Nichols).

## *7.1 Soil Moisture Observations*

### 7.1.1 Direct Insertion

At the point scale, Heathman et al. (2003) directly inserted daily gravimetric ground measurements of surface soil moisture (0–5 cm) as a surrogate for remote sensing data to estimate the profile water content (0–60 cm) at four locations in a large watershed. Four soil layers were modelled of 15 cm each and an additional fifth layer was a top 5 cm layer. The results were compared to time domain reflectometer (TDR) measurements. They found no significant improvement in soil water estimates below 30 cm depth. They also stated that daily observations were needed for good results.

Montaldo et al. (2001) presented an operational assimilation framework for crudely assimilating surface soil moisture measurements in a simple SVAT model: biases between observed and modelled time rates of change of surface soil moisture were used to quantify biases between modelled and actual root-zone-average soil moisture contents. They tested the framework for misspecification of a parameter, the saturated hydraulic conductivity, and for uncertain initial conditions, and found improvements through assimilation. The assimilation frequency was found to be of limited importance: infrequent corrections were reported to be sustained by the internal model dynamics. It should be noted that data assimilation intervals of 3–120 h only were considered. In a subsequent study, Montaldo and Albertson (2003) recognized that large errors in the saturated hydraulic conductivity resulted in persistent bias in the predictions and proposed a multi-scale (in time) assimilation system in which the root zone soil moisture was updated at the observation time scale and the parameter was adjusted at a coarser time-scale, since it would be questionable to adapt parameters as frequent as observations would be available.

At a coarser scale, Li and Islam (1999) assimilated gravimetric measurements of soil moisture as a surrogate for remote sensing data through daily hard-updating over a single unit region in a four-layer model. They used site-averaged data over an area of $15 \times 15$ km$^2$. They focused on the role of surface soil moisture assimilation in the partitioning of fluxes and found that assimilation of surface soil moisture had a positive impact, under the assumption of zero error in the observations and forcings. For deeper layers the improvement in profile predictions decreased. They speculated that in the presence of commonly encountered random measurement errors, daily assimilation of microwave measurements of soil moisture would not improve the profile estimate and the partitioning of the fluxes. They studied three different frequencies of assimilation: 12, 24 and 48 h and found a limited sensitivity to the data assimilation frequency, with slightly better results for more intensive data assimilation.

### 7.1.2 Statistical Correction, Nudging, Optimal Interpolation

Houser et al. (1998) compared different assimilation strategies with TOPLATS, using off-line inverted remotely sensed microwave observations in a distributed

model set-up. They found that Newtonian nudging assimilation was preferable to statistical corrections assimilation and optimal interpolation. Pauwels et al. (2001) assimilated real ERS images in TOPLATS to assess the impact on discharge predictions. Through comparison of results from a nudging and a statistical correction technique in both a lumped and a distributed model version, they found that assimilating the statistics (spatial mean and variance) of remote sensing data in lumped models sufficed to improve discharge predictions. Paniconi et al. (2003) used the Newtonian nudging technique in a pure synthetical study over an idealized artificial study area to assimilate surface soil moisture in a 3-D Richards equation-based distributed model. They stated that four-dimensional weighting functions used in the nudging approach provide a simple way to incorporate knowledge on characteristic length scales and spatio-temporal variability of the state variables. Hurkmans et al. (2006) tested the sensitivity of this dynamical relaxation technique for the different nudging parameters.

### 7.1.3 Kalman Filter

*Point profile estimation*: Entekhabi et al. (1994) were among the pioneers to estimate time-dependent 1 m soil moisture and temperature profiles under bare soil from synthetic measurements of microwave and infrared radiation, using an EKF. The direct use of emitted radiation by including a complex observation model, i.e., a radiative transfer model (RTM), in the Kalman filter procedure was a key significant feature of their study. They found that starting from an intentionally poor initial guess and with hourly updates, the estimates improved in time and, eventually, the dynamics of the true profile were captured down to depths far beyond the penetration depth of the observations. This work, with inclusion of an RTM in the estimation procedure, was extended by Galantowicz et al. (1999) using daily field data of L-band radio brightness over a period of 8 days at a Beltsville Agricultural Research Center bare soil test plot, and synthetic data over a 4 month period at an observation interval of 3 days. They studied a soil column of 1 m depth with 31 layers. They initialized the a priori state error covariance matrix $\mathbf{P}^b$ as a diagonal matrix with large diagonal values. Through time, deeper soil moisture could be retrieved as the interdepth covariances had been adapted through modelled moisture percolation and redistribution (i.e., the off-diagonal elements increased). Likewise, Crosson et al. (2002) estimated soil moisture distributions by assimilating brightness temperatures with a Kalman filter incorporating an RTM. Each time brightness temperature data were available, the modelled soil moisture profiles were used as input in a forward RTM and combined with the observations to update the soil profile.

Active microwave observations were assimilated by Hoeben and Troch (2000) in a synthetic study to estimate the soil profile with an EKF. They studied the effect of system and observational noise and found that in the presence of realistic system noise, the retrieval is feasible with an acceptable accuracy, but for observational noise which approaches the real world satellite errors, the accuracy of the profile retrievals drops to the level of the reference run without data assimilation. Based

on their investigation of the update interval (from hourly to every 2 days), they suggested that daily radar images would be necessary for accurate updates.

Wendroth et al. (1999) applied a Kalman filter to the surface layer of a 3-layer (10, 30 and 50 cm) soil profile and found that assimilation of pressure head observations improved the soil moisture estimates for deeper layers, even when the model showed clear shortcomings in the simulation of evapotranspiration.

Walker et al. (2001a) discussed a 1-D soil moisture profile retrieval by assimilation of synthetic near-surface soil moisture ground measurements, which greatly simplified the observation operator. The KF scheme was found to be superior to a direct insertion scheme. They found that the observation interval was not important for profile estimation with a Kalman filter, when the forcing data was accurate (to ensure correct predictions). The observation depth did not have a significant effect on the profile retrieval time with a Kalman filter. This synthetic study was extended by a study using real field data from the Nerrigundah catchment. Walker et al. (2001b) developed a simplified soil moisture model (ABDOMEN) and studied assimilation of near-surface soil moisture measurements for profile soil moisture retrieval. They found that the presence of bias hampered the success of the Kalman filter procedure and that less frequent updating improved the total soil moisture profile, while near-surface soil moisture was poorly predicted. Therefore, they stressed the need for an appropriate forecasting model and suggested that assimilation of near-surface soil moisture would be useful only to correct for errors in soil moisture forecasts as result of errors in initial conditions and/or atmospheric forcing data, and not as a result of errors in the physics of the soil moisture model.

*Lumped spatial field estimation*: Georgakakos and Baumer (1996) used the Kalman filter to assimilate basin-integrated ground surface soil moisture observations to estimate soil moisture in a deeper layer with a simple conceptual 2-layer model. Through a sensitivity study, they found that even when the upper soil water measurements contained substantial (added) noise, the estimates of lower soil water contents were improved.

In the framework of the European AIMWATER project on the Seine catchment, the assimilation of Synthetic Aperture Radar (SAR) observations was considered (Francois et al. 2003), mainly aiming at updating discharge flows. Oudin et al. (2003) concluded that the current SAR instruments have a repeat time that is too low to enhance their parameter updating procedure efficiency. In a very simple study on four subcatchments of the Seine catchment, they also found that the optimal soil moisture depth (of TDR measurements) for parameter updating was dependent on the subcatchment considered. Streamflow and soil moisture were estimated for the same study area by a sequential Kalman filter by Aubert et al. (2003). They simulated time repetitivity of remote sensing data by eliminating part of their TDR measurements and found that through assimilation the efficiency remained higher than without assimilation for a repeat time as high as 1 week.

Crow (2003) found successful results through daily assimilation of brightness temperature observations via an EnKF to correct for the impact of poorly sampled rainfall data on land surface predictions of root-zone soil moisture and surface energy fluxes. Plot-scale simulations were run with the TOPLATS SVAT on 2

sites, both for 2 approaches: using synthetic data in an identical twin data assimilation experiment (these experiments are discussed in chapter *Observing System Simulation Experiments*, Masutani et al.), and using real data obtained during the Southern Great Plains 1997 (SGP97) experiment. They indicated that an increased observation frequency (up to once every 5 days) reduced the sensitivity of the results to the frequency in rainfall observations. The filter performance was evaluated with regard to the assumptions that underlie the optimality of the KF update equations. Crow and Wood (2003) also applied an EnKF to assimilate remotely sensed soil brightness temperatures using point-scale TOPLATS results at 2 sites to compensate predictions in surface latent heat flux and root-zone water storage for errors due to use of climatological rainfall data. They discussed inadequacies in model physics, and the contrasts of spatial support between model predictions and sensor retrievals. They found little improvement when increasing the ensemble sizes and suggested that for larger ensemble sizes, alternative error sources and shortcomings in the reference results themselves are more important than the errors arising from finite ensemble sizes.

Wilker et al. (2006) conducted a single column (single site) SGP97 assimilation experiment with an EKF and the operational Numerical Weather Prediction (NWP) system of the European Centre for Medium-Range Weather Forecasts (ECMWF). They showed that, in the case of non-uniform soil moisture profiles, the typical top layer vertically integrated simulated soil moisture will introduce errors, because the top surface layer is not resolved properly to represent the soil moisture corresponding (through a forward operator) to the observed brightness temperature. Therefore, they advised to correct the observations for this representativeness error.

*Distributed spatial field estimation*: Distributed applications of the Kalman filter are often limited by computational constraints and hence reformulated as a collection of individual 1-D applications.

Reichle et al. (2002a, b) compared synthetic experiments results using an EnKF to the variational approach and the Extended Kalman filter, respectively. They gave insights into the theoretical and practical aspects of these techniques, and illustrate them for distributed case studies with different LSMs over different testbed regions in the northern USA.

Margulis et al. (2002) discussed the EnKF in a field test with assimilation of real brightness data into a 1-D model, applied over the study area of the SGP97 experiment. They aggregated the observational data to $4 \times 4$ km$^2$ pixels to reduce the computational load. Through assimilation, they found that surface soil moisture and latent heat flux estimates were nearly always closer to ground truth measurements and more consistently within the measurement error bars than the open loop simulations (i.e., without data assimilation).

While most studies on soil moisture assimilation focused on filtering techniques, Dunne and Entekhabi (2005) argued that the soil moisture estimation problem should be treated as a reanalysis-type problem, as observations beyond the estimation time still provide useful information, as long as subsequent precipitation events are avoided. Dunne and Entekhabi (2005) compared the performance of an Ensemble Kalman smoother to that of an EnKF in an Observing

System Simulation Experiment (OSSE; see chapter *Observing System Simulation Experiments*, Masutani et al.), using a 1-D model (uncorrelated grid cells) and an RTM to merge model results with synthetic brightness data. Because of the occurrence of precipitation, a hybrid smoother/filter approach was presented to break the study interval into a series of smoothing windows of single drydowns. They found that including future observations could improve the initial conditions at depth, resulting in improved latent heat flux estimates. Dunne and Entekhabi (2006) compared the EnKF and its smoother variant for estimation of soil moisture and surface energy fluxes by assimilation of real L-band brightness data over the SGP97 study area.

Walker and Houser (2001) avoided spin-up of a land surface model by initializing soil moisture through a 1-D EKF of synthetic near-surface soil moisture data over the North American continent. This study illustrated the essential goal of data assimilation for state estimation, i.e., find the best state (analysis, initial condition) to initiate future predictions.

Walker and Houser (2004) addressed requirements for soil moisture satellite accuracy, repeat time and spatial resolution through a twin experiment with a 1-D EKF in the Catchment model with 3 moisture prognostic variables. Each catchment (average area of 4,400 $km^2$) was taken as a calculation unit. The resolution and accuracy requirements were found to be much more important than repeat time. They found that the soil moisture observations should have accuracy better than 5 vol%; the resolution of the assimilated data should be less than the resolution of the land surface model; and repeat times should be from 1 to 5 days.

Crow and van Loon (2006) applied TOPLATS to a watershed, assuming that sub-basin scale variability in water table depth was solely driven by the local soil-topography index. In a synthetic filtering experiment (without any need for detailed spatial validation), they showed possible pitfalls with adaptive filtering, the consequences of an improper selection of model error and the benefit of combined soil moisture and runoff data for adaptive filtering.

Some studies have tried to approximate the full 3-D problem by sophisticated mathematical techniques. Reichle and Koster (2003) compared the performance of a 1-D and 3-D EnKF in a synthetic twin experiment with the Catchment model, considering only 3 state variables related to soil moisture per catchment. Since non-zero off-diagonal elements would necessitate a simultaneous update for all catchments or grid cells, this would require immense computational effort. However, for continental or global soil moisture fields, the scale for horizontal error correlations is much smaller than the domain size, and covariance localization can be used in combination with a parallel implementation. They found that information was spread from observed to unobserved catchments, when taking into account the horizontal error correlations.

Walker et al. (2002) applied a 3-D KF to assimilate near-surface measurements from the Nerrigundah catchment for soil moisture profile estimation in a full 3-D soil moisture model. Because the spatial coupling necessitated a computationally efficient methodology to propagate the state error covariances, a simplified system dynamics approach was used.

De Lannoy et al. (2009) implemented a full 3-D adaptive EnKF system, which was parallellized both in the forecast part and the update part, but not in the calculation of the Kalman gain. This allowed finding spatial error correlations between individually simulated soil profiles, which could then be used to propagate observational information in a single profile to all other profiles in a small-scale field.

### 7.1.4 3D/4D-Var

*Point scale estimation*: Calvet et al. (1998) assimilated field measurements of surface soil moisture through a (strong constraint) variational data assimilation approach, and found that 4 or 5 observations spread over an assimilation window of 15 days were enough to retrieve total soil water content by inverting the Interactions between Soil, Biosphere and Atmosphere (ISBA) scheme, given correct knowledge of the forcings. Use of the soil temperature was more of a problem, because it was not always found to be sensitive to the total water content. Their study was conducted at one site at the point scale and used averaged gravimetric soil moisture measurements (0–5 cm) during an intensive observation period (IOP) of 30 days. Wingeron et al. (1999) used the same ISBA model in combination (variational) with a surface soil moisture data set of 3 months during the vegetation cycle of soybean to study requirements for the use of remote sensing measurements of soil moisture to accurately estimate the root zone soil moisture.

*Lumped spatial field estimation*: Li and Islam (2002) applied a model inversion technique suggested by Calvet et al. (1998) for assimilation of surrogate remotely sensed data (averaged gravimetric samples) over a $15 \times 15$ km$^2$ area, and found that the estimation of surface soil moisture was very sensitive to the initialization of deeper layer soil moisture. They also found that initialization of the soil moisture profile in such a way that it optimizes the error in the surface soil moisture, may not lead to optimal estimation of surface fluxes and accurate retrieval of deeper layer soil moisture. This was attributed to a decoupling of the surface and deeper soil moisture.

Pathmathevan et al. (2003) estimated 1-D soil moisture profiles through daily variational assimilation of microwave radiometer measurements in the Land Surface Scheme (LSS) of the Simple Biosphere Model 2 (SiB2) with 3 layers of soil moisture. No adjoint model was developed, and a heuristic optimization technique, simulated annealing, was used instead to minimize a variational cost function.

*Distributed spatial field estimation*: Reichle and McLaughlin (2001) discussed a pioneering synthetic study on the feasibility of a 4-D variational assimilation algorithm, where they used a representer method to account for both model and observational error. They developed a relatively simple model to allow development of a numerically well-behaved adjoint model. A non-linear radiative transfer model (RTM) was included on-line in the assimilation procedure to allow direct assimilation of brightness temperatures. In a closely related 4-D weak constraint variational data assimilation study, Reichle et al. (2001) showed that synthetic radio

brightness measurements could be used to estimate soil moisture at scales finer than the resolution of the brightness images, provided that information on land surface characteristics and micro-meteorological inputs were available.

## *7.2 Soil Temperature Observations*

Soil temperature can be used to update the total land surface state, including soil moisture, and its dependent land surface fluxes, like evapotranspiration. Lakshmi (2000) used surface temperature to validate model surface temperatures and adjust model simulated soil moisture. A two-layer land surface model was applied to 66 $1^\circ \times 1^\circ$ pixels over the area of the Red-Arkansas study region. For a period of 1 year, radiance data, which were available as gridded fields and converted to temperature data, were equally weighted with model simulations twice a day to obtain an adjusted estimate of the soil temperature. They found that, through assimilation, on average (over the studied area) improved estimates for soil moisture were found and the effect of errors in the forcings was reduced. Spatially distributed comparisons of soil moisture fields showed a reduced difference between observed (satellite brightness data converted to soil moisture through an RTM) and simulated soil moisture, and also a reduced standard deviation in the difference.

Land surface temperature may be used to provide estimates of components of the surface energy balance and land surface control on evaporation. Kumar and Kaleita (2003) used an EKF to assimilate top layer temperature measurements for the estimation of a soil temperature profile at a 1-point site at the western edge of the Little Washita River Watershed. Data were assimilated at $^1/_2$ and 24-h increments in a soil column of 6 layers. When data were available every $^1/_2$ h, the lower layers responded much more rapidly to the inclusion of observed data. They also found that the correlation structure between the different layers was more complicated than could be described with a simple diagonal matrix.

Castelli et al. (1999) applied a variational methodology with an adjoint to assimilate area-averaged soil surface temperature for retrieving surface fluxes and a soil moisture index, which depended on soil wetness and aerodynamic conditions.

Caparrini et al. (2004) discussed the determination of turbulent heat fluxes by variational assimilation of remotely sensed land surface temperature into a surface energy balance model and showed an application to a large area within the US Great Plains. They showed how to assimilate measurements with varying scales and with overlapping coverages.

Boni et al. (2001) assimilated half-hourly in situ ground temperature observations to generate a reference (validation truth) to explore the value of satellite data assimilation by a variational technique with an adjoint. The performance was found to vary with the timing of the satellite overpass and the estimation was most improved when measurements were available close to the time of peak ground temperature. The study was conducted on 2 sites (not distributed in space) within the SGP97

experiment area. The satellite brightness data were area-averaged and converted to temperatures off-line, before assimilation.

## 7.3 Low-Level Atmospheric Observations

Research on the assimilation of screen-level measurements of relative humidity and temperature (Bouttier et al. 1993; Viterbo and Beljaars 1995) has mostly focused on variational studies, which tried to find an optimal initial state by searching for a best match between the resulting model simulations and the observations. These low-level atmospheric data have been used because they are widely available and very sensitive to soil moisture. Generally, soil moisture estimates have been integrated over the root zone. Mahfouf (1991) introduced the assimilation of low-level atmospheric variables such as relative humidity and air temperature to initialize soil moisture for improved short- and medium-range weather forecasts. He tested a strong constraint variational approach and a sequential nudging approach, which was a statistical algorithm based on linear regressions. Based on this work, several authors investigated these two approaches with atmospheric observations. Bouttier et al. (1993), Hu et al. (1999), Douville et al. (2000), Pleim and Xiu (2003), for example, explored and adapted the sequential nudging technique, while the studies of Callies et al. (1998), Bouyssel et al. (1999) and Hess (2001) followed the variational approach. Douville et al. (2000) found that the nudging technique was very sensitive to model bias. Rhodin et al. (1999) applied the technique of Callies et al. (1998) to a regional weather forecast model, while neglecting all horizontal correlations to facilitate the 3-D assimilation problem. It should be remarked that in all these studies, soil moisture has little physical meaning and it is rather used as a parameter in NWP models. Assimilation of atmospheric variables is interesting because these data are readily available in an operational system for NWP, but Bouyssel et al. (1999), for example, reported that operational implementation of surface analyses is difficult with these kind of data in general weather conditions including precipitation, cloudy conditions or large-scale advection. Seuffert et al. (2004) found that synergistic assimilation of screen-level parameters and microwave brightness temperatures yielded more consistent results than assimilation of either observation type separately.

## 7.4 Land Surface Flux Observations

To date, only a few studies have explored the assimilation of remotely sensed land surface flux observations. Schuurmans et al. (2003) converted $1 \times 1$ km$^2$ resolution remotely sensed data into latent heat flux estimates to assimilate them into a distributed hydrological model to improve the model water balance. They applied an optimal interpolation scheme (constant gain Kalman filter), to study the spatial

distribution of model latent heat flux estimates and found improvements in areas with higher elevations.

## *7.5 Vegetation-Based Observations*

Specification of seasonal variations of vegetation properties can significantly affect the simulation of several near-surface variables. Mahfouf and Viterbo (2001) indicated that the difficulty in capturing the variability of vegetation is the relation between satellite reflectances to input parameters, such as leaf area index (LAI) and albedo in land surface schemes. Pauwels et al. (2007) have shown that through assimilation of LAI values and soil moisture observations the results from coupled hydrological/crop growth models can be improved.

## *7.6 Discharge Observations*

The possibility to use discharge data to update the state variables of a hydrological model has been explored, either using only discharge data (Pauwels and De Lannoy 2006; Vrugt et al. 2006; Komma et al. 2008), or a combination of discharge and soil moisture data (Aubert et al. 2003).

A fundamental difference between the assimilation of runoff rates and the assimilation of other variables (for example soil moisture values) is the fact that observed catchment runoff rates are the integrated result of runoff generating processes occurring between the moment of the observation and a number of time steps before the observation. This implies that, if a discharge observation at a certain time is assimilated, the state variables of a number of time steps before the assimilation need to be updated as well, to assure fully optimized discharge forecasts.

## *7.7 Snow Water Equivalent/Snow Cover Observations*

Snow on land is an important variable affecting the global energy and water budgets, because of its high albedo, low thermal conductivity, considerable spatial and temporal variation and medium-term capacity for water storage. The amount of water in snow, i.e., the snow water equivalent (SWE), can be observed in situ or derived from brightness temperatures, e.g. obtained from the Advanced Microwave Scanning Radiometer – EOS (AMSR-E) or Special Sensor Microwave/Imager (SSM/I). Another commonly used observation is the snow cover area or fraction (SCA or SCF), which can be measured relatively accurately through remote sensing, e.g. with the Moderate Resolution Imaging Spectroradiometer (MODIS). Some challenges related to snow assimilation have been discussed by Walker and Houser (2005).

Cosgrove and Houser (2002) showed that large water balance errors could occur when directly inserting SWE observations into imperfectly modelled snow melting processes. Assimilation of remotely sensed SWE was studied in a synthetic study with a 1-D EKF by Sun et al. (2004). Slater and Clark (2006) assimilated real SWE data in a conceptual model with an Ensemble Square Root Kalman filter and showed that merging of information was better than either the model results or interpolated observations. Durand and Margulis (2006) conducted a feasibility study using a variety of point-scale synthetically generated observations in combination with a LSM and a RTM to assess the contribution of each channel (brightness temperature $T_b$ of SSM/I, $T_b$ of AMSR-E, and broadband albedo from MODIS satellite products) to recovering the true SWE. Other interesting follow-up studies on SWE assimilation were reported by Durand and Margulis (2007, 2008). Dong et al. (2007) used a 1-D EKF to assimilate SWE data, which were obtained after conversion of Scanning Multichannel Microwave Radiometer (SMMR) brightness temperature to SWE. They excluded data with potential high errors. Andreadis and Lettenmaier (2006) found that assimilation of AMSR-E SWE data into the VIC model was not very successful, due to errors in the AMSR-E product. De Lannoy et al. (2010) successfully downscaled AMSR-E-scale synthetic SWE to retrieve fine-scale variability in several 1-D and 3-D EnKF setups.

Modelling results by Déry et al. (2005) showed improved runoff timing and runoff amounts when MODIS fractional snow cover data were incorporated. Rodell and Houser (2004) used satellite-derived MODIS SCA to update the SWE in a land surface model by a rule-based assimilation scheme, but they found that SCA contained very little information about SWE. Andreadis and Lettenmaier (2006) applied an EnKF to update SWE by assimilation of SCA from MODIS data and used a simple snow depletion curve as the observation operator to relate SWE to SCA.

## *7.8 Ground Water Storage Observations*

Subsurface observations have only seen limited use to estimate the land surface state, most likely because using these observations only yields a limited observability of the land surface system (the information content in the observations does not allow to fully reconstruct the land surface state). Zaitchik et al. (2008) assimilated Gravity Recovery and Climate Experiment (GRACE)-derived monthly terrestrial water storage (TWS) anomalies for each of the four major subbasins of the Mississippi into the Catchment Land Surface Model (CLSM) using an Ensemble Kalman smoother, and obtained improved water storage and fluxes dynamics.

Several applications of the Kalman filter in the field of groundwater modelling have been studied by Eigbe et al. (1998) and Porter et al. (2000), but most studies do not consider the land surface state, but rather focus on the inverse problem of determining the hydraulic properties, assuming a perfect groundwater model.

## 8 Case Studies

Significant advances in hydrological data assimilation have been made over the past decade from which we have selected a few case studies to demonstrate the utility of hydrological data assimilation.

### *8.1 Case Study 1: Soil Moisture Assimilation*

The estimation of soil moisture profiles has received considerable attention, because a correct assessment of the soil moisture state is crucial to estimate the partitioning of surface fluxes, for weather predictions and climate analyses. For this case study, the EnKF was used with the Community Land Model (CLM2.0) to assimilate ground measurements for soil moisture profile estimation (De Lannoy et al. 2007a). The focus was on the determination of the best observation conditions for assimilation and on the optimization of the method with real data.

An Ensemble Kalman filter for state estimation and a dynamic bias estimation algorithm was applied to estimate individual soil moisture profiles in a small corn field with the CLM2.0 model through the assimilation of measurements from capacitance probes. Both without and with inclusion of forecast bias correction, the effect of the assimilation frequency, the assimilation depths, and the number of observations assimilated per profile, were studied. Assimilation of complete profiles had the highest impact on deeper soil layers, and the optimal assimilation frequency was about 1–2 weeks, if bias correction was applied (Fig. 4). Without bias correction, a higher assimilation frequency always further improved the results (Fig. 4). Bias correction on top of state estimation extracts more information from the observations and thus a limited assimilation frequency is sufficient for optimal results. The optimal assimilation depth depended on the calibration results. Assimilation in

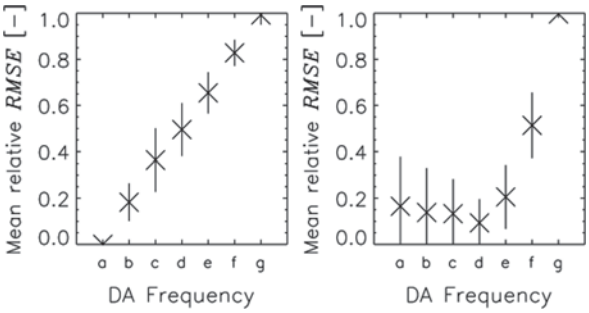



**Fig. 4** Mean (and spatial standard deviation over 36 field profiles) relative profile-integrated RMSE (root mean square error) for profile assimilation with (*left*) EnKF and (*right*) EnBKF (with inclusion of dynamic forecast bias estimation) as a function of varying numbers of observations in time. Assimilation intervals a, b, c, d, e, f and g are 1, 2, 4 days, 1, 2, 4 and 8 weeks, respectively

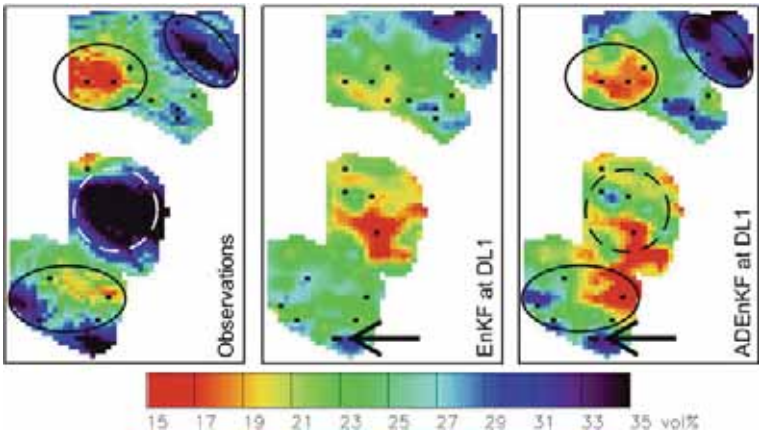



**Fig. 5** Spatially interpolated fields of (*left*) observations, (*middle*) ensemble mean forecasts and single-profile EnKF analysis at the indicated point location only and (*right*) full 3-D adaptive EnKF analyses after assimilation at the indicated point location only. The *black dots* indicate observed locations. The *full ellipses* show areas with an improved impact through adaptive 3-D EnKF filtering. The simulated (*middle panel*) moisture is underestimated in the *dashed circular area* and the adaptive 3-D filter (without bias estimation) cannot overcome the large bias

the surface layer had typically less impact than assimilation in other layers. In general, the correct propagation of the innovations for the bias-blind state as well as for the bias filtering from any layer to other layers was insufficient. The approximate estimation of the a priori (bias) error covariance and the choice of a zero-initialized persistent bias model made it impossible to accurately estimate the bias in layers for which no observations were available.

In a subsequent study by De Lannoy et al. (2009), horizontal propagation of assimilated profiles information in space was achieved after optimizing (training) spatial forecast error covariances in an adaptive three-dimensional (3-D) EnKF. In Fig. 5 an interpolated field of point-scale measurements is shown, together with a one-profile EnKF analysis and a 3-D EnKF analysis after spatial error covariance training. The one-profile EnKF updates all observed and unobserved profile layers only at the assimilation location. The full 3D EnKF spreads this information to all unobserved locations in space.

These studies show the importance of both a good first and second order error characterization for Kalman filtering with real data, i.e. soil moisture forecast bias estimation and spatial forecast error covariance estimation.

## *8.2 Case Study 2: Streamflow Assimilation*

Rüdiger et al. (2005) have shown the potential for assimilating streamflow measurements to retrieve soil moisture. A synthetic study was undertaken on three nested catchments (sequentially draining into each other) within the Goulburn River

experimental catchment in south-eastern Australia (Rüdiger et al., 2007). Three scenarios are presented: (i) only streamflow observations are available for the outlet of the lowest catchment; (ii) streamflow observations are not available and surface soil moisture observations are only available for one of the catchments under the assumption that the other two catchments are too densely vegetated to allow a reliable retrieval of soil moisture; and (iii) streamflow observations are available for the lower catchment and surface soil moisture observations for the middle catchment. This synthetic study identifies the potential of using different observations, where and when available, for the retrieval of soil moisture initial states. The assimilation type here is performed as an initial state optimization through minimizing an objective function penalizing the deviation from the observed soil moisture and/or streamflow over some assimilation window.

The assessment is based on a comparison between assimilated, truth and non-assimilated (control) simulations in Fig. 6. In the control simulation, the root zone soil moisture content was subjected to a wet bias. It was found that the assimilation of streamflow has a significant improvement in the retrieval of profile and root zone soil moisture in all three catchments, but displays limitations in retrieving the surface soil moisture state. In contrast, the assimilation of surface soil moisture in the lower catchment alone does not have any effect on the upstream catchments, as there

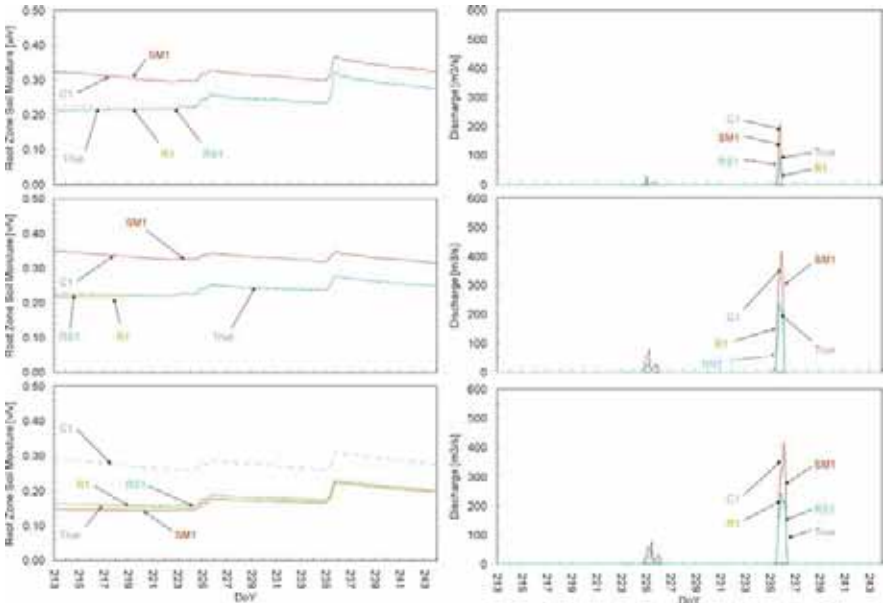


**Fig. 6** Assimilation results for root zone soil moisture and runoff for (R1) discharge (observed at the *lower catchment*) assimilation only, (SM1) soil moisture (observed at the *lower catchment*) assimilation only, and (RS1) simultaneous assimilation of soil moisture (observed in the *middle catchment*) and discharge (*lower catchment*). Individual catchments are shown in rows (*upper*, *middle*, and *lower catchments*). C1 represents the control run without assimilation

is no feedback between the downstream and upstream soil moisture and respective runoff. Finally, the joint assimilation of both streamflow and surface soil moisture observations leads to a further improvement from the streamflow assimilation alone.

The comparison between the results from the degraded and the assimilation runs show a good improvement of the overestimated soil moisture and runoff values through streamflow assimilation. The best performance is observed for the lower catchment, with slight inaccuracies for the two upstream catchments. The main difference between the "truth" observations and the assimilation run is the retrieval of the surface soil moisture content, which is underestimated. This is due to the initial surface soil moisture content not having a significant impact on the runoff and, hence, the objective function, when the profile moisture is well retrieved. While the infiltration capacity excess mechanism is still the main process contributing to runoff (runoff is only produced when saturation of the surface soil moisture is achieved), there is no runoff occurring in the first 10 days of the assimilation window, so that changes to the initial soil moisture states cannot generate runoff events. The precipitation events causing runoff occur over a short period, but during these events sufficient water is introduced into the catchment to wet up the surface layer to the point of saturation and allow runoff to be produced. Because the root zone soil moisture is accurately retrieved, all subsequent soil moisture values are close to the true observations, and, therefore, the initial value of the surface soil moisture before its saturation during the first precipitation event is irrelevant.

The study of Rüdiger et al. (2005) was undertaken as a proof-of-concept twin-study for streamflow assimilation, in which only the initial states were perturbed. In Rüdiger et al. (2007) wet and dry biases and white noise were added to the forcing data to simulate uncertainties in the observational data base, while assuming that there is no knowledge about observational or background errors.

## *8.3 Case Study 3: Snow Assimilation*

Accurate prediction of snowpack status is important for a range of environmental applications, yet model estimates are typically poor and in situ measurement coverage is inadequate. Moreover, remote sensing estimates are spatially and temporally limited due to complicating effects, including distance to open water, presence of wet snow, and presence of thick snow (Dong et al. 2005). However, through assimilation of remote sensing estimates into a land surface model, it is possible to capitalize on the strengths of both approaches (Dong et al. 2007). In order to achieve this, reliable estimates of the uncertainty in both remotely sensed and model simulated snow water equivalent (SWE) estimates are critical. For practical application, the remotely sensed SWE retrieval error is prescribed with a spatially constant but monthly varying value, with data omitted for: (1) locations closer than 200 km to significant open water; (2) times and locations with model-predicted presence of liquid water in the snowpack; and (3) model SWE estimates greater than 100 mm.

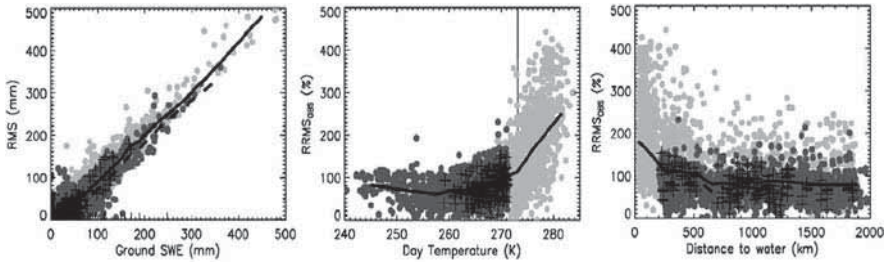



**Fig. 7** SMMR passive microwave SWE retrieval root mean square (RMS) error of the in situ SWE estimates (*left panel*), average monthly daytime temperature (*middle panel*), and "distance" to water (*right panel*). The *light grey dots* show all the data and *dark gray dots* show the data remaining after omitting pixels closer than 200 km to water and with an average monthly daytime temperature above 2°C; the *lines* show the mean values respectively. The *pluses* represent data for pixels including 5 or more ground stations

The model error is estimated using standard error propagation with a calibrated spatially and temporally constant model error contribution. The SWE estimates from assimilation were found to be superior to both the model simulation and remotely sensed estimates alone, except when model SWE estimates rapidly and erroneously crossed the 100 mm SWE cut-off early in the snow season.

Based on an extensive evaluation of SMMR SWE estimates Dong et al. (2005) suggest that SMMR SWE retrievals should not be used for: (1) regions within 200 km of significant open water bodies due to mixed pixel contamination; (2) times when monthly mean air temperature is above 2°C due to potential meltwater contamination; and (3) times and locations where in situ SWE values are above 100 mm due to microwave signal saturation (Fig. 7). Restricting the use of remotely sensed SWE on this basis was found to result in a nearly unbiased SWE estimate with a seasonal maximum RMS (root-mean-square) median error of 20 mm.

A set of numerical experiments were undertaken by Dong et al. (2007) to evaluate the impact of assimilating quality controlled SMMR SWE retrievals on snowpack state variables. The first simulation is a straight model simulation run (i.e., without data assimilation, and referred to as the open-loop run) to show how the model performed in the absence of assimilation. The second and third simulations are two Extended Kalman filter (EKF) assimilation experiments (referred to as assimilation run-I and run-II), started with the same initial conditions as the open-loop run, but assimilating the remotely sensed SMMR SWE estimates when available. The difference between these two runs is that run-I assimilates all available SMMR SWE data while run-II only assimilates quality-controlled data. The median predicted and observed SWE estimates for pixels with five or more in situ stations are shown in Fig. 8. For the simulations starting in the middle of winter, it was found that assimilation run-II outperformed both of the other snowpack simulations, with the results from assimilation run-I approaching the unmasked SMMR SWE values. This was expected, as erroneous SWE observations when not eliminated (as in run-I), or adequately characterized by their error covariances, act to degrade the snowpack

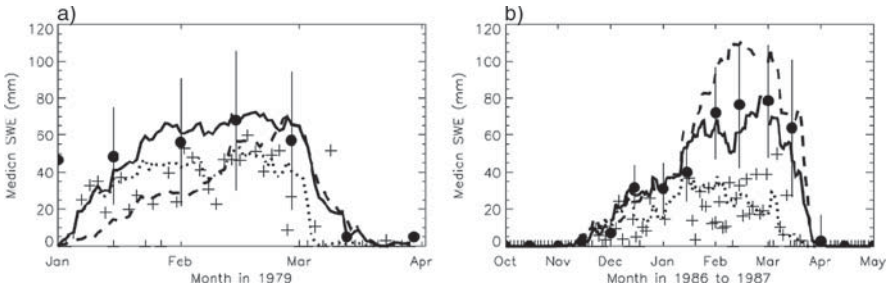



**Fig. 8** Comparison of the median SWE for pixels including five or more stations; ground observations (*black dots*), SMMR observations (*plus symbols*), model forecast (*dash lines*), model forecast with assimilation run-I (*dotted lines*) and run-II (*solid lines*) from: (**a**) January to March in 1979 (*left panel*), and (**b**) July 1986–June 1987 (zoomed to the winter months from October 1986 to April 1987 – *right panel*). The *vertical lines* show the plus one and minus one standard deviation from the median of the ground observations

simulation through their assimilation. The open-loop simulation significantly underestimates the snowpack SWE throughout the entire simulation due to the zero snow initialization. The resulting median estimates from assimilation run-II are in close agreement with the ground observations. Statistical analysis shows that bias error has been largely reduced, and RMS error has been slightly reduced.

## *8.4 Case Study 4: Skin Temperature Assimilation*

The land surface skin temperature state is a principal control on land-atmosphere fluxes of water and energy. It is closely related to soil water states, and is easily observable from space and aircraft infrared sensors in cloud-free conditions. The usefulness of skin temperature in land data assimilation studies is limited by its very short memory (on the order of minutes) due to the very small heat storage it represents. Radakovich et al. (2001) have demonstrated skin temperature data assimilation in a land surface model using three-hourly observations from the International Satellite Cloud Climatology Project (ISCCP) – see Fig. 9. Incremental and semi-diurnal bias correction techniques based on Dee and da Silva (1998) were developed to account for biased skin temperature forecasts. The assimilation of ISCCP-derived surface skin temperature significantly reduced the bias and standard deviation between model predictions and the National Centers for Environmental Prediction (NCEP) reanalysis (Kalnay et al. 1996). However, the assimilation of ISCCP-derived surface skin temperature has a substantial impact on the sensible heat flux, due to an enhanced gradient between the surface and 2 m air temperatures. If the near-surface air temperature were interactive, as in a coupled land-atmosphere model, then it would respond to this enhanced flux rather than maintaining the artificial temperature gradient.

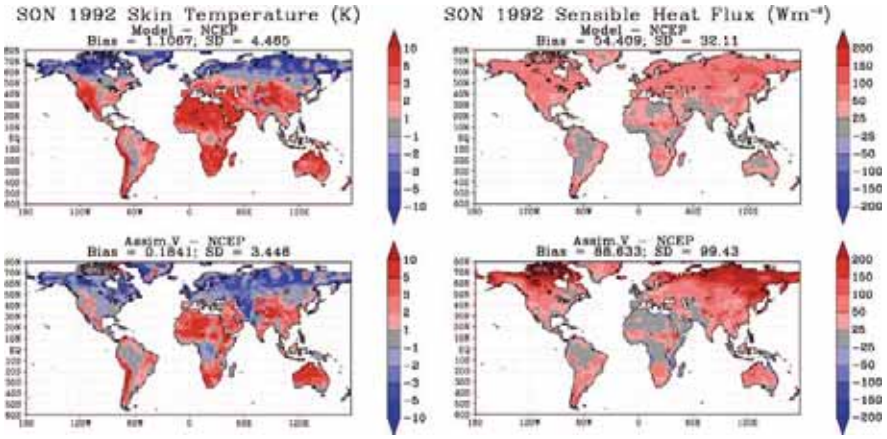



**Fig. 9** Differences between simulated and reanalysis (*top left*), assimilated and reanalysis (*bottom left*) mean skin temperatures (K), and the resulting differences between simulated and reanalysis (*top right*), and assimilated and reanalysis (*bottom right*) mean sensible heat fluxes ($Wm^{-2}$) for September–November 1992. Global terrestrial mean bias and standard deviation (SD) for September–November are also noted (Radakovich et al. 2001)

This study was extended by Bosilovich et al. (2007), where remotely sensed surface temperature was assimilated into a coupled atmosphere/land global data assimilation system, with explicit accounting for biases in the model state. In this scheme, an incremental bias correction term is introduced in the model's surface energy budget. The method was validated against the assimilated observations, as well as independent near surface air temperature observations. In many regions, not accounting for the diurnal cycle of bias caused degradation of the diurnal amplitude of background model air temperature. Energy fluxes collected through the Coordinated Enhanced Observing Period (CEOP) were used to more closely inspect the surface energy budget. In general, sensible heat flux is improved with the surface temperature assimilation, and two stations show a reduction of bias by as much as 30 $Wm^{-2}$.

## 9 Summary

Hydrological data assimilation is an objective method to estimate the hydrological system states from irregularly distributed observations. These methods integrate observations into numerical prediction models to develop physically consistent estimates that better describe the hydrological system state than the raw observations alone. This process is extremely valuable for providing initial conditions for hydrological system prediction and/or correcting hydrological system prediction, and for increasing our understanding and improving parametrization of hydrological system behaviour through various diagnostic research studies.

Hydrological data assimilation has still many open areas of research. Development of hydrological data assimilation theory and methods is needed to: (i) better quantify and use model and observational errors; (ii) create model-independent data assimilation algorithms that can account for the typical non-linear nature of hydrological models; (iii) optimize data assimilation computational efficiency for use in large operational hydrological applications; (iv) use forward models to enable the assimilation of remote sensing radiances directly; (v) link model calibration and data assimilation to optimally use available observational information; (vi) create multivariate hydrological assimilation methods to use multiple observations with complementary information; (vii) quantify the potential of data assimilation downscaling; and (viii) create methods to extract the primary information content from observations with redundant or overlaying information. Further, the regular provision of snow, soil moisture, and surface temperature observations with improved knowledge of observational errors in time and space are essential to advance hydrological data assimilation. Hydrological models must also be improved to: (i) provide more "observable" land model states, parameters, and fluxes; (ii) include advanced processes such as river runoff and routing, vegetation and carbon dynamics, and groundwater interaction to enable the assimilation of emerging remote sensing products; (iii) have valid and easily updated adjoints; and (iv) have knowledge of their prediction errors in time and space. The assimilation of additional types of hydrological observations, such as streamflow, vegetation dynamics, evapotranspiration, and groundwater or total water storage must be developed.

As with most current data assimilation efforts, we describe data assimilation procedures that are implemented in uncoupled models. However, it is well known that the high-resolution time and space complexity of hydrological phenomena have significant interaction with atmospheric, biogeochemical, and oceanic processes. Scale truncation errors, unrealistic physics formulations, and inadequate coupling between hydrology and the overlying atmosphere can produce feedbacks that can cause serious systematic hydrological errors. Hydrological balances cannot be adequately described by current uncoupled hydrological data systems, because large analysis increments that compensate for errors in coupling processes (e.g. precipitation) result in important non-physical contributions to the energy and water budgets. Improved coupled process models with improved feedback processes, better observations, and comprehensive methods for coupled assimilation are needed to achieve the goal of fully coupled data assimilation systems that should produce the best and most physically consistent estimates of the Earth system.

## References

Andreadis, K.M. and D.P. Lettenmaier, 2006. Assimilating remotely sensed snow observation into a macroscale hydrology model. *Adv. Water Resour.*, **29**, 872–886.

Arya, L.M., J.C. Richter and J.F. Paris, 1983. Estimating profile water storage from surface zone soil moisture measurements under bare field conditions. *Water Resour. Res.*, **19**, 403–412.

Aubert, D., C. Loumagne and L. Oudin, 2003. Sequential assimilation of soil moisture and streamflow data into a conceptual rainfall-runoff model. *J. Hydrol.*, **280**, 145–161.

Bennett, A.F., 1992. *Inverse Methods in Physical Oceanography,* Cambridge University Press, Cambridge, 346 pp.

Bergthorsson, P. and B. Döös, 1955. Numerical weather map analysis. *Tellus*, **7**, 329–340.

Bernard, R., M. Vauclin and D. Vidal-Madjar, 1981. Possible use of active microwave remote sensing data for prediction of regional evaporation by numerical simulation of soil water movement in the unsaturated zone. *Water Resour. Res.*, **17**, 1603–1610.

Beven, K., 1989. Changing ideas in hydrology: The case of physically-based models. *J. Hydrol.*, **105**, 157–172.

Boni, G., D. Entekhabi and F. Castelli, 2001. Land data assimilation with satellite measurements for the estimation of surface energy balance components and surface control on evaporation. *Water Resour. Res.*, **37**, 1713–1722.

Bosilovich, M.G., J.D. Radakovich, A.D. Silva, R. Todling and F. Verter, 2007. Skin temperature analysis and bias correction in a coupled land-atmosphere data assimilation system. *J. Meteorol. Soc. Jpn.*, **85A**, 205–228.

Bouttier, F. and P. Courtier, 1999. Data assimilation concepts and methods. *ECMWF training course notes*.

Bouttier, F., J.-F. Mahfouf and J. and Noilhan, 1993. Sequential assimilation of soil moisture from atmospheric low-level parameters. Part I: Sensitivity and calibration studies. *J. Appl. Meteorol.*, **32**, 1335–1351.

Bouyssel, F., V. Cassé and J. Pailleux, 1999. Variational surface analysis from screen level atmospheric parameters. *Tellus*, **51A**, 453–468.

Bras, R. and I. Rodriguez-Iturbe,1985. *Random Functions and Hydrology,* Addison Wesley, Reading, MA, 590 pp.

Bratseth, A.M., 1986. Statistical interpolation by means of successive corrections. *Tellus*, **38A**, 439–447.

Bruckler, L. and H. Witono, 1989. Use of remotely sensed soil moisture content as boundary conditions in soil-atmosphere water transport modeling: 2. Estimating soil water balance. *Water Resour. Res.*, **25**, 2437–2447.

Callies, U., A. Rhodin and D. Eppel, 1998. A case study on variational soil moisture analysis from atmospheric observations. *J. Hydrol.*, **212–213**, 95–108.

Calvet, J.-C., J. Noilhan and P. Bessemoulin, 1998. Retrieving the root-zone soil moisture from surface soil moisture or temperature estimates: A feasibility study based on field measurements. *J. Appl. Meteorol.*, **37**, 371–386.

Caparrini, F., F. Castelli and D. Entekhabi, 2004. Variational estimation of soil and vegetation turbulent transfer and heat flux parameters from sequences of multisensor imagery. *Water Resour. Res.*, **40**, W12515.1–W12515.15.

Castelli, F., D. Entekhabi and E. Caporali, 1999. Estimation of surface heat flux and an index of soil moisture using adjoint-state surface energy balance. *Water Resour. Res.*, **35**, 3115–3125.

Charney, J.G., M. Halem and R. Jastrow, 1969. Use of incomplete historical data to infer the present state of the atmosphere. *J. Atmos. Sci.*, **26**, 1160–1163.

Chen, Y. and D. Zhang, 2006. Data assimilation for transient flow in geologic formations via ensemble Kalman filter. *Adv. Water Resour.*, **29**, 1107–1122.

Cosgrove, B.A. and P.R. Houser, 2002. The effect of errors in snow assimilation on land surface modeling. Preprints, *16th Conference on Hydrology*, Orlando, FL, American Meteor Society, J136–J137.

Cressman, G.P., 1959. An operational objective analysis system. *Mon. Weather Rev.*, **87**, 367–374.

Crosson, W.L., C.A. Laymon, R. Inguva and M.P. Schamschula, 2002. Assimilating remote sensing data in a surface flux-soil moisture model. *Hydro. Processes*, **16**, 1645–1662.

Crow, W., 2003. Correcting land surface model predictions for the impact of temporally sparse rainfall rate measurements using an ensemble Kalman filter and surface brightness temperature observations. *J. Hydrometeorol.*, **4**, 960–973.

Crow, W.T. and E. van Loon, 2006. Impact of incorrect model error assessment on the sequential assimilation of remotely sensed surface soil moisture. *J. Hydrometeorol.*, **7**, 421–432.

Crow, W.T. and E.F. Wood, 2003. The assimilation of remotely sensed soil brightness temperature imagery into a land surface model using ensemble Kalman filtering: A case study based on ESTAR measurements during SGP97. *Adv. Water Resour.*, **26**, 137–149.

Daley, R., 1991. *Atmospheric Data Analysis,* Cambridge University Press, Cambridge, 460 pp.

Dee, D.P. and A. da Silva, 1998. Data assimilation in the presence of forecast bias. *Q. J. R. Meteorol. Soc.*, **124**, 269–295.

Dee, D.P. and R. Todling, 2000. Data assimilation in the presence of forecast bias: The GEOS moisture analysis. *Mon. Weather Rev.*, **128**, 3268–3282.

De Lannoy, G.J.M., P.R. Houser, V.R.N. Pauwels and N.E.C. Verhoest, 2006. Assessment of model uncertainty for soil moisture through ensemble verification. *J. Geophys. Res.*, **111**, D10101.1–D10101.18.

De Lannoy, G.J.M., P.R. Houser, V.R.N. Pauwels and N.E.C. Verhoest, 2007a. State and bias estimation for soil moisture profiles by an ensemble Kalman filter: Effect of assimilation depth and frequency. *Water Resour. Res.*, **43**, W06401, doi:10.1029/2006WR005100.

De Lannoy, G.J.M., P.R. Houser, N.E.C. Verhoest and V.R.N. Pauwels, 2009. Adaptive soil moisture profile filtering for horizontal information propagation in the independent column-based CLM2.0. *J. Hydrometeorol.*, **10**, 766–779.

De Lannoy, G.J.M., R.H. Reichle, P.R. Houser, K.R. Arsenault, V.R.N. Pauwels and N.E.C. Verhoest, 2010. Satellite-scale snow water equivalent assimilation into a high-resolution land surface model. *J. Hydrometeorol.*, **11**, 352–369, doi:10.1175/2009JHM1194.1.

De Lannoy, G.J.M., R.H. Reichle, P.R. Houser, V.R.N. Pauwels and N.E.C. Verhoest, 2007b. Correcting for forecast bias in soil moisture assimilation with the ensemble Kalman filter. *Water Resour. Res.*, **43**, W09410, doi:10.1029/2006WR00544.

Déry, S.J., V.V. Salomonson, M. Stieglitz, D.K. Hall and I. Appel, 2005. An approach to using snow areal depletion curves inferred from MODIS and its application to land surface modelling in Alaska. *Hydrol. Processes*, **19**, 2755–2774.

Dirmeyer, P., 2000. Using a global soil wetness dataset to improve seasonal climate simulation. *J. Climate*, **13**, 2900–2921.

Dong, J., J.P. Walker and P.R. Houser, 2005. Factors affecting remotely sensed snow water equivalent uncertainty. *Remote Sens. Environ.*, **97**, 68–82, doi:10.1016/j.rse.2005.04.010.

Dong, J., J.P. Walker, P.R. Houser and C. Sun, 2007. Scanning multichannel microwave radiometer snow water equivalent assimilation. *J. Geophys. Res.*, **112**, D07108, doi:10.1029/2006JD007209.

Douville, H., P. Viterbo, J.-F. Mahfouf and A.C.M. Beljaars, 2000. Evaluation of the optimum interpolation and nudging techniques for soil moisture analysis using fife data. *Mon. Weather Rev.*, **128**, 1733–1756.

Duan, Q., S. Sorooshian and V.K. Gupta, 1992. Effective and efficient global optimization for conceptual rainfall-runoff models. *Water Resour. Res.*, **28**, 1015–1031.

Dunne, S. and D. Entekhabi, 2005. An ensemble-based reanalysis approach to land data assimilation. *Water Resour. Res.*, **41**, W02013.1–W02013.18.

Dunne, S. and D. Entekhabi, 2006. Land surface state and flux estimation using the ensemble Kalman smoother during the Southern Great Plains 1997 field experiment. *Water Resour. Res.*, **42**, W01407.1–W01407.15.

Durand, M. and S.A. Margulis, 2006. Feasibility test of multifrequency radiometric data assimilation to estimate snow water equivalent. *J. Hydrometeorol.*, **7**, 443–457.

Durand, M. and S.A. Margulis, 2007. Correcting first-order errors in snow water equivalent estimates using a multifrequency, multiscale radiometric data assimilation scheme. *J. Geophys. Res.*, **112**, D13121.1–D13121.15.

Durand, M. and S.A. Margulis, 2008. Effects of uncertainty magnitude and accuracy on assimilation of multiscale measurements for snowpack characterization. *J. Geophys. Res.*, **113**, D02105.1–D02105.17.

Eigbe, U., M. Beck, H. Weather and F. Hirano, 1998. Kalman filtering in groundwater flow modelling: Problems and prospects. *Stochast. Hydrol. Hydraul.*, **12**, 15–32.

Entekhabi, D., H. Nakamura and E.G. Njoku, 1994. Solving the inverse problem for soil moisture and temperature profiles by sequential assimilation of multifrequency remotely sensed observations. *IEEE Trans. Geosci. Rem. Sens.*, **32**, 438–448.

Evensen, G., 1994. Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *J. Geophys. Res.*, **99**, 10143–10162.

Evensen, G., 2003. The ensemble Kalman filter: Theoretical formulation and practical implementation. *Ocean Dyn.*, **53**, 343–367.

Francois, C., A. Quesney and C. Ottlé, 2003. Sequential assimilation of ERS-1 SAR data into a coupled land surface-hydrological model using an extended Kalman filter. *J. Hydrometeorol.*, **4**, 473–487.

Friedland, B., 1969. Treatment of bias in recursive filtering. *IEEE Trans. Autom. Control*, **AC-14**, 359–367.

Galantowicz, J.F., D. Entekhabi and E.G. Njoku, 1999. Tests of sequential data assimilation for retrieving profile soil moisture and temperature from observed L-band radiobrightness. *IEEE Trans. Geosci. Rem. Sens.*, **37**, 1860–1870.

Georgakakos, K.P. and O.W. Baumer, 1996. Measurement and utilization of on-site soil moisture data. *J. Hydrol.*, **184**, 131–152.

Gove, J.H. and D.Y. Hollinger, 2006. Application of a dual unscented Kalman filter for simultaneous state and parameter estimation in problems of surface atmosphere exchange. *J. Geophys. Res.*, **111**, D08S07.1–D08S07.21.

Heathman, G., P. Starks, L. Ahuj and T. Jackson, 2003. Assimilation of soil moisture to estimate profile soil water content. *J. Hydrol.*, **279**, 1–17.

Hebson, C. and E. Wood, 1985. Partitioned state and parameter estimation for real-time flood forecasting. *Appl. Math. Comput.*, **17**, 357–374.

Hess, R., 2001. Assimilation of screen-level observations by variational soil moisture analysis. *Meteorol. Atmos. Phys.*, **77**, 145–154.

Hoeben, R. and P.A. Troch, 2000. Assimilation of active microwave observation data for soil moisture profile estimation. *Water Resour. Res.,* **36**, 2805–2819.

Hollingsworth, A. and P. Lönnberg, 1989. The verification of objective analyses: Diagnostics of analysis system performance. *Meteorol. Atmos. Phys.*, **40**, 3–27.

Houser, P., M.F. Hutchinson, P. Viterbo, J. Hervé Douville and S.W. Running, 2004. Terrestrial data assimilation, Chapter C.4. In *Vegetation, Water, Humans and the Climate,* Global Change – The IGB Series, Kabat, P. et al. (eds.), Springer, Berlin, pp 273–287.

Houser, P.R., W.J. Shuttleworth, J.S. Famiglietti, H.V. Gupta, K.H. Syed and D.C. Goodrich, 1998. Integration of soil moisture remote sensing and hydrologic modeling using data assimilation. *Water Resour. Res.*, **34**, 3405–3420.

Houtekamer, P.L. and H.L. Mitchell, 1998. Data assimilation using a Ensemble Kalman filter techniques. *Mon. Weather Rev.*, **126**, 796–811.

Hu, Y., X. Gao, W. Shuttleworth, H. Gupta and P. Viterbo, 1999. Soil moisture nudging experiments with a single column version of the ECMWF model. *Q. J. R. Meteorol. Soc.*, **125**, 1879–1902.

Hurkmans, R., C. Paniconi and P.A. Troch, 2006. Numerical assessment of a dynamical relaxation data assimilation scheme for a catchment hydrological model. *Hydrol. Processes*, **20**, 549–563.

Jackson, T.J., T.J. Schmugge, A.D. Nicks, G.A. Coleman and E.T. Engman, 1981. Soil moisture updating and microwave remote sensing for hydrological simulation. *Hydrol. Sci. Bull.*, **26**, 305–319.

Jazwinski, A.H., 1970. *Stochastic Processes and Filtering Theory,* Vol. 64. Academic Press, New York, 376 pp.

Kalman, R.E., 1960. A new approach to linear filtering and prediction problems. *Trans. ASME, Ser. D, J. Basic Eng.*, **82**, 35–45.

Kalnay, E., M. Kanamitsu, R. Kistler, et al., 1996. The NCEP/NCAR 40-year reanalysis project. *Bull. Amer. Meteorol. Soc.*, **77**, 437–471.

Katul, G.G., O. Wendroth, M.B. Parlange, C.E. Puente, M.V. Folegatti and D.R. Nielsen, 1993. Estimation of in situ hydraulic conductivity function from nonlinear filtering theory. *Water Resour. Res.*, **29**, 1063–1070.

Komma, J., G. Blöschl and C. Reszler, 2008. Soil moisture updating by ensemble Kalman filtering in real-time flood forecasting. *J. Hydrol.*, **357**, 228–242.

Koster, R.D., M. Suarez, P. Liu, U. Jambor, A. Berg, M. Kistler, R. Reichle, M. Rodell and J. Famiglietti, 2004. Realistic initialization of land surface states: Impacts on subseasonal forecast skill. *J. Hydrometeorol.*, **5**, 1049–1063.

Kostov, K.G. and T.J. Jackson, 1993. Estimating profile soil moisture from surface layer measurements – A review. In: *Proceedings of the International Society for Optical Engineering,* Vol. 1941. Orlando, FL, pp 125–136.

Kumar, P. and A.L. Kaleita, 2003. Assimilation of near-surface temperature using extended Kalman filter. *Adv. Water Resour.*, **26**, 79–93.

Lakshmi, V., 2000. A simple surface temperature assimilation scheme for use in land surface models. *Water Resour. Res.*, **36**, 3687–3700.

Li, J. and S. Islam, 1999. On the estimation of soil moisture profile and surface fluxes partitioning from sequential assimilation of surface layer soil moisture. *J. Hydrol.*, **220**, 86–103.

Li, J. and S. Islam, 2002. Estimation of root zone soil moisture and surface fluxes partitioning using near surface soil moisture measurements. *J. Hydrol.*, **259**, 1–14.

Lorenc, A., 1981. A global three-dimensional multivariate statistical interpolation scheme. *Mon. Weather Rev.*, **109**, 701–721.

Lorenc, A.C., R.S. Bell and B. Macpherson, 1991. The meteorological office analysis correction data assimilation scheme. *Q. J. R. Meteorol. Soc.*, **117**, 59–89.

Mahfouf, J.-F., 1991. Analysis of soil moisture from near-surface parameters: A feasibility study. *J. Appl. Meteorol.*, **30**, 1534–1547.

Mahfouf, J. and P. Viterbo, 2001. Land surface assimilation. *Meteorological Training Course Lecture Series ECMWF*.

Margulis, S.A., D. McLaughlin, D. Entekhabi and S. Dunne, 2002. Land data assimilation of soil moisture using measurements from the Southern Great Plains 1997 field experiment. *Water Resour. Res.*, **38**, 35.1–35.18.

Margulis, S.A., E.F. Wood and P.A. Troch, 2006. A terrestrial water cycle: Modeling and data assimilation across catchment scales. *J. Hydrometeorol.*, **7**, 309–311.

Maybeck, P.S., 1979. *Stochastic Models, Estimation, and Control,* Vol. 1 (Vol. 141). Academic Press, Toronto, 423 pp.

McLaughlin, D., 1995. Recent developments in hydrologic data assimilation. In U.S. National Report to the IUGG (1991–1994). *Rev. Geophys.*, **33**(supplement), 977–984.

McLaughlin, D., 2002. An integrated approach to hydrologic data assimilation: Interpolation, smoothing, and filtering. *Adv. Water Resour.*, **25**, 1275–1286.

Milly, P.C.D., 1986. Integrated remote sensing modelling of soil moisture: Sampling frequency, response time, and accuracy of estimates. *Integrated Design of Hydrological Networks – Proceedings of the Budapest Symposium*, IAHS Publication No. 158, 201–211.

Milly, P. and Z. Kabala, 1986. Integrated modelling and remote sensing of soil moisture. In *Hydrologic applications of space technology – Proceedings of the Cocoa Beach Workshop,* Vol. 158. Florida, pp 201–211.

Montaldo, N. and J.D. Albertson, 2003. Multi-scale assimilation of surface soil moisture for robust root zone moisture predictions. *Adv. Water Resour.*, **26**, 33–44.

Montaldo, N., J.D. Albertson, M. Mancini and G. Kiely, 2001. Robust simulation of root zone soil moisture with assimilation of suface soil moisture data. *Water Resour. Res.*, **37**, 2889–2900.

Moradkhani, H., S. Sorooshian, H.V. Gupta and P.R. Houser, 2005. Dual state-parameter estimation of hydrological models using ensemble Kalman filter. *Adv. Water Resour.*, **28**, 135–147.

Nichols, N.K., 2001. State estimation using measured data in dynamic system models, *Lecture notes for the Oxford/RAL Spring School in Quantitative Earth Observation*.

Ottlé, C. and D. Vidal-Madjar, 1994. Assimilation of soil moisture inferred from infrared remote sensing in a hydrological model over the HAPEX-MOBILHY Region. *J. Hydrol.*, **158**, 241–264.

Oudin, L., A. Weisse, C. Loumage and S. Le Hégarat-Mascle, 2003. Assimilation of soil moisture into hydrological models for flood forecasting: A variational approach. *Can. J. Rem. Sens.*, **29**, 679–686.

Pan, M. and E.F. Wood, 2006. Data assimilation for estimating the terrestrial water budget using a constrained ensemble Kalman filter. *J. Hydrometeorol.*, **7**, 534–547.

Paniconi, C., M. Marrocu, M. Putti and M. Verbunt, 2003. Newtonian nudging for a Richards equation-based distributed hydrological model. *Adv. Water Resour.*, **26**, 161–178.

Parrish, D. and J. Derber, 1992. The national meteorological center's spectral statistical interpolation analysis system. *Mon. Weather Rev.*, **120**, 1747–1763.

Pathmathevan, M., T. Koike, X. Lin and H. Fujii, 2003. A simplified land data assimilation scheme and its application to soil moisture experiments in 2002 (SMEX02). *Water Resour. Res.*, **39**, SWC6.1–SWC6.20.

Pauwels, V.R.N. and G.J.M. De Lannoy, 2006. Improvement of modeled soil wetness conditions and turbulent fluxes through the assimilation of observed discharge. *J. Hydrometeorol.*, **7**, 458–477.

Pauwels, V.R.N., R. Hoeben, N.E.C. Verhoest and F.P. De Troch, 2001. The importance of the spatial patterns of remotely sensed soil moisture in the improvement of discharge predictions for small-scale basins through data assimilation. *J. Hydrol.*, **251**, 88–102.

Pauwels, V.R.N., N.E.C. Verhoest, G.J.M. De Lannoy, V. Guissard, C. Lucau and P. Defourny, 2007. Optimization of a coupled hydrology/crop growth model through the assimilation of observed soil moisture and LAI values using an Ensemble Kalman Filter. *Water Resour. Res.*, **43**, W04421, doi:10.1029/2006WR004942.

Pleim, J.E. and A. Xiu, 2003. Development of a land surface model. Part II: Data assimilation. *J. Appl. Meteorol.*, **42**, 1811–1822.

Porter, D., B. Gibbs, W. Jones, P. Huyakorn, L. Hamm and G. Flach, 2000. Data fusion modeling for groundwater systems. *J. Contam. Hydrol.*, **42**, 303–335.

Prevot, L., R. Bernard, O. Taconet, et al., 1984. Evaporation from a bare soil evaluated using a soil water transfer model and remotely sensed surface soil moisture data. *Water Resour. Res.*, **20**, 311–316.

Radakovich, J.D., P.R. Houser, A. da Silva and M.G. Bosilovich, 2001. Results from global land-surface data assimilation methods. *Proceedings AMS 5th Symposium on Integrated Observing Systems*, Albuquerque, NM, 14–19 January, pp 132–134.

Reichle, R.H., W.T. Crow and C.L. Keppenne, 2008. An adaptive ensemble Kalman filter for soil moisture data assimilation, *Water Resour. Res.*, **44**, W03423, doi:10.1029/2007WR006357.

Reichle, R.H., D. Entekhabi and D.B. McLaughlin, 2001. Downscaling of radiobrightness measurements for soil moisture estimation: A four-dimensional variational data assimilation approach. *Water Resour. Res.*, **37**, 2353–2364.

Reichle, R.H. and R. Koster, 2003. Assessing the impact of horizontal error correlations in background fields on soil moisture estimation. *J. Hydrometeorol.*, **4**, 1229–1242.

Reichle, R.H. and R. Koster, 2004. Bias reduction in short records of satellite soil moisture. *Geophys. Res. Lett.*, **31**, L19501.1–L19501.4.

Reichle, R.H. and D.B. McLaughlin, 2001. Variational data assimilation of microwave radiobrightness observations for land surface hydrologic applications. *IEEE Trans. Geosci. Rem. Sens.*, **39**, 1708–1718.

Reichle, R.H., D.B. McLaughlin and D. Entekhabi, 2002a. Hydrologic data assimilation with the ensemble Kalman filter. *Mon. Weather Rev.*, **120**, 103–114.

Reichle, R.H., J.P. Walker, P.R. Houser and R.D. Koster, 2002b. Extended versus ensemble Kalman filtering for land data assimilation. *J. Hydrometeorol.*, **3**, 728–740.

Rhodin, A., F. Kucharski, U. Callies, D. Eppel and W. Wergen, 1999. Variational analysis of effective soil moisture from screen-level atmospheric parameters: Application to a short-range weather forecast model. *Q. J. R. Meteorol. Soc.*, **125**, 2427–2448.

Rodell, M. and P.R. Houser, 2004. Updating a land surface model with MODIS-derived snow cover. *J. Hydrometeorol.*, **5**, 1064–1075.

Rood, R.B., S.E. Cohn and L. Coy, 1994. Data assimilation for EOS: The value of assimilated data, Part 1. *Earth Observer*, **6**, 23–25.

Rüdiger, C., G. Hancock, H.M. Hemakumara, B. Jacobs, J.D. Kalma, C. Martinez, M. Thyer, J.P. Walker, T. Wells and G.R. Willgoose, 2007. The Goulburn River experimental catchment data set. *Water Resour. Res.*, **43**, W10403, doi:10.1029/2006WR005837.

Rüdiger, C., J.P. Walker, J.D. Kalma, G.R. Willgoose and P.R. Houser, 2005. Root zone soil moisture retrieval using streamflow and surface soil moisture data assimilation. In *MODSIM 2005 International Congress on Modelling and Simulation*, Zerger, A. and Argent, R.M. (eds.), Modelling and Simulation Society of Australia and New Zealand, Inc., Melbourne, Australia, 12–15 December, 2005, pp 1458–1464.

Schuurmans, J., P. Troch, A. Veldhuizen, W. Bastiaansen and M. Bierkens, 2003. Assimilation of remotely sensed latent heat flux in a distributed hydrological model. *Adv. Water Resour.*, **26**, 151–159.

Seuffert, G., H. Wilker, P. Viterbo, M. Drusch and J.-F. Mahfouf, 2004. The usage of screen-level parameters and microwave brightness temperature for soil moisture analysis. *J. Hydrometeorol.*, **5**, 516–531.

Slater, A.G. and M. Clark, 2006. Snow data assimilation via an ensemble Kalman filter. *J. Hydrometeorol.*, **7**, 478–493.

Stauffer, D.R. and N.L. Seaman, 1990. Use of four-dimensional data assimilation in a limited-area mesoscale model. Part I: Experiments with synoptic-scale data. *Mon. Weather Rev.*, **118**, 1250–1277.

Stieglitz, M., D. Rind, J. Famiglietti and C. Rosenzweig, 1997. An efficient approach to modeling the topographic control of surface hydrology for regional and global climate modeling. *J. Climate*, **10**, 118–137.

Sun, C., J.P. Walker and P.R. Houser, 2004. A methodology for snow data assimilation in a land surface model. *J. Geophys. Res.*, **109**, D08108.1–D08108.12.

Thiemann, M., M. Trosset, H. Gupta and S. Sorooshian, 2001. Bayesian recursive parameter estimation for hydrological models. *Water Resour. Res.*, **37**, 2521–2535.

Turner, M.R.J., J.P. Walker and P.R. Oke, 2007. ensemble member generation for sequential data assimilation. *Remote Sens. Environ.*, **112**, doi:10.1016/j.rse.2007.02.042.

van Loon, E.E. and P.A. Troch, 2001. Directives for 4-D soil moisture data assimilation in hydrological modelling. *IAHS*, **270**, 257–267.

Viterbo, P. and A. Beljaars, 1995. An improved land surface parameterization scheme in the ECMWF model and its validation. *J. Climate*, **8**, 2716–2748.

Vrugt, J.A., H.V. Gupta, B. O'Nualláin and W. Bouten, 2006. Real-time data assimilation for operational ensemble streamflow forecasting. *J. Hydrometeorol.*, **7**, 548–565.

Walker J.P. and P.R. Houser, 2001. A methodology for initialising soil moisture in a global climate model: Assimilation of near-surface soil moisture observations. *J. Geophys. Res.*, **106**, 11761–11774.

Walker, J.P. and P.R. Houser, 2004. Requirements of a global near-surface soil moisture satellite mission: Accuracy, repeat time, and spatial resolution. *Adv. Water Resour.*, **27**, 785–801.

Walker, J.P. and P.R. Houser, 2005. Hydrologic data assimilation. In *Advances in Water Science Methodologies*, Aswathanarayana, A. (ed.), A.A. Balkema, The Netherlands, 230 pp.

Walker, J.P., P.R. Houser and R. Reichle, 2003. New technologies require advances in hydrologic data assimilation. *EOS*, **84**, 545–551.

Walker, J.P., G.R. Willgoose and J.D. Kalma, 2001a. One-dimensional soil moisture profile retrieval by assimilation of near-surface observations: A comparison of retrieval algorithms. *Adv. Water Resour.*, **24**, 631–650.

Walker, J.P., G.R. Willgoose and J.D. Kalma, 2001b. One-dimensional soil moisture profile retrieval by assimilation of near-surface measurements: A simplified soil moisture model and field application. *J. Hydrometeorol.*, **2**, 356–373.

Walker, J.P., G.R. Willgoose and J.D. Kalma, 2002. Three-dimensional soil moisture profile retrieval by assimilation of near-surface measurements: Simplified Kalman filter covariance forecasting and field application. *Water Resour. Res.*, **38**, 1301, doi:10.1029/2002WR001545.

Wendroth, O., H. Rogasik, S. Koszinski, C.J. Ritsema, L.W. Dekker and D.R. Nielsen, 1999. State-space prediction of field-scale soil water content time series in a sandy loam. *Soil & Till. Res.,* **50**, 85–93.

Wilker, H., M. Drusch, G. Seuffert and C. Simmer, 2006. Effects of the near-surface soil moisture profile on the assimilation of L-band microwave brightness temperature. *J. Hydrometeorol.*, **7**, 433–442.

Wingeron, J.-P., A. Olioso, J.-C. Calvet and P. Bertuzzi, 1999. Estimating root zone soil moisture from surface soil moisture data and soil-vegetation-atmosphere transfer modeling. *Water Resour. Res.*, **35**, 3735–3745.

WMO, 1992. Simulated real-time intercomparison of hydrological models (*Tech. Rep. No. 38*). Geneva.

Zaitchik, B.F., M. Rodell and R. Reichle, 2008. Assimilation of GRACE terrestrial water storage data into a land surface model: Results for the Mississippi river basin. *J. Hydrometeorol.*, **9**, 535–548.

Zhang, H. and C.S. Frederiksen, 2003. Local and nonlocal impacts of soil moisture initialization on AGCM seasonal forecasts: A model sensitivity study. *J. Climate*, **16**, 2117–2137.

# Assimilation of GPS Soundings in Ionospheric Models

**Boris Khattatov**

## 1 Introduction

The ionosphere, usually defined to extend upward from about 80 km, is the region of the Earth's atmosphere where concentration of ionized particles becomes sufficiently high to become easily observable. Below 80 km absorption of solar radiation by the atmosphere above decreases the probability of a neutral atmospheric molecule being ionized and results in negligible concentration of ionized particles. At altitudes higher than about 400 km, the density of neutral particles that are subject to ionization decreases substantially and absolute concentration of charged particles gets smaller with altitude.

Unlike the low and middle atmosphere, forecasting ionospheric conditions must involve forecasting the external drivers, primarily solar activity in the form of solar flux and coronal mass ejections (CMEs). Accurate deterministic forecasting of solar behaviour is an extremely difficult task due to both complexity of the physical processes involved and lack of observations. Therefore, practical ionospheric data assimilation systems mainly focus on nowcasting rather than forecasting. Efforts are underway to apply methods of data assimilation to solar physics and the magnetosphere; however such efforts are not the subject of this chapter. Even the relatively simpler objective of nowcasting the ionospheric conditions is made difficult by the short characteristic time-scales (often minutes and sometimes seconds) and scarcity of observations.

Accurate knowledge of the ionospheric conditions was historically a low priority. This situation changed when our increasing reliance on ground-space communications became important for both military and civilian infrastructure. Adverse ionospheric effects on certain types of communications led to an increased need for knowing the current ionospheric state, primarily electron densities. This, in turn, resulted in an increased investment in observing and accurate modelling of the

B. Khattatov (✉)
Fusion Numerics Inc, Boulder, CO, USA
e-mail: boris@fusionnumerics.com

W. Lahoz et al. (eds.), *Data Assimilation*, DOI 10.1007/978-3-540-74703-1_22,

ionosphere, naturally spawning efforts to involve data assimilation methodologies; for details of these methodologies see chapters in Part I, *Theory*.

In this chapter we discuss basic processes controlling ionospheric composition and dynamics at low and mid latitudes; give an overview of the numerical modelling approaches; describe Global Positioning System (GPS) observations available for assimilation into models; and discuss some practical uses of reliable ionospheric specification.

## 2 Background

For the purpose of this chapter we define the ionosphere as the region of the Earth's atmosphere where charged particles, ions and electrons, become sufficiently abundant to noticeably influence dynamics and chemistry of the atmosphere. The primary source of ions (and electrons) is ionization of neutral particles by solar short wave ultraviolet (UV) radiation. At altitudes lower than about 80 km, the intensity of the UV radiation is small due to absorption by the dense neutral atmosphere and thus concentration of charged particles is low. At high altitudes the concentration of neutral particles that can be ionized decreases exponentially with altitude leading to a decrease in the concentration of charged particles. These conditions result in a pronounced peak of ion density at 200–400 km known as the *F-region* of the ionosphere.

Assimilation of observations in ionospheric physics-based models is a relatively new area of research. First practical attempts to apply data assimilation methods common in numerical weather prediction (see chapter *Numerical Weather Prediction*, Swinbank) to the ionosphere took place in early 1990s. There are a number of reasons for this late adoption of the data assimilation methods.

The scarcity of observations available for assimilation into ionospheric models naturally affected adversely the introduction of data assimilation methods to ionospheric modelling. Making observations of the ionospheric state is a difficult task due to the high altitude of the ionosphere and relatively low concentration of ions compared to neutral particles.

Development of practical, affordable, and reliable means of observing the ionosphere was a low priority until our communication and navigation infrastructure became vulnerable to adverse ionospheric conditions. Indeed, arguably the largest volume of data related to ionospheric conditions is currently derived from measuring ionosphere-induced dispersion that adversely affects the accuracy of GPS operations.

Yet another reason is that unlike meteorological conditions in the troposphere or the stratosphere, where initial conditions largely determine the future system state, ionospheric conditions are strongly influenced by external forcing. This is similar to forecasting chemical weather in the troposphere, where one has to take account of sources and sinks (see chapter *Inverse Modelling and Combined State-Source Estimation for Chemical Weather*, Elbern et al.).

Generally, any meaningful medium- or long-term ionospheric forecast must be capable of forecasting this forcing. The external forcing comes primarily in two

forms – increased intensity of UV and X-ray radiation from the Sun due to solar flares, and impact of clouds of plasma emitted by the Sun as a result of coronal mass ejections. The first process results in enhanced density of ionized particles in the ionosphere, the second one can potentially alter the ionospheric magnetic field in a drastic manner and lead to highly anomalous ionospheric transport, which in turn affects the composition and temperature of ionospheric charged particles. Since both processes are generally poorly understood, and since near real time observations of the solar activity suitable for forecasting are even less abundant than observations of the ionosphere, reliable deterministic forecasting of such events is currently impossible.

While efforts are under way to include methods of data assimilation for nowcasting and forecasting solar influences and magnetospheric conditions, these efforts are outside the scope of this chapter. Here we will focus on describing challenges involved in nowcasting ionospheric conditions, primarily estimating three-dimensional (3-D) electron densities, using a numerical model of ionospheric composition and GPS observations.

As of February 2009 the author was aware of only four physics-based assimilation systems designed for nowcasting and eventually forecasting ionospheric conditions. Development of all four has been funded by either or both the US Navy and the US Air Force Research Laboratory. One such system that uses a sequential assimilation approach has been developed at Utah State University (Schunk et al. 2004; Scherliess et al. 2004). The Jet Propulsion Laboratory, in collaboration with the University of Southern California, developed both sequential and variational assimilation systems (Pi et al. 2003; Hajj et al. 2004). Fusion Numerics Inc has built an operational system (Khattatov et al. 2005) consisting of a numerical time-dependent model and a sequential data assimilation scheme based on the approach described in Ménard et al. (2000) and Khattatov et al. (2000).

In this chapter we briefly describe elementary ionospheric processes controlling ionospheric composition and dynamics at low and middle latitudes (Sect. 3), numerical modelling procedures (Sect. 4), GPS data (Sect. 5), challenges with data assimilation for the ionosphere (Sect. 6), the impact of the ionosphere on telecommunications (Sect. 7), and practical utilization of ionospheric specifications obtained in the process of data assimilation (Sect. 8). Section 9 discusses future directions.

## 3 Overview of Ionospheric Processes

### *3.1 Elementary Processes*

Here we present a short overview of elementary ionospheric processes. For a more comprehensive description the reader is referred to the excellent descriptions in Banks and Kockarts (1973) or Schunk and Nagy (2000).

*Solar radiation*, in particular short-wave ultra violet radiation, is the primary source of ionization in the atmosphere. Absorption of a photon by a neutral particle,

either atom or a molecule, results in the appearance of an ion and one or more electrons, e.g. (where O is an oxygen atom and $O^+$ an oxygen ion):

$$O + h\nu \rightarrow O^+ + e$$

Only a few neutral particles have ionization absorption spectra that result in noticeable ionization yields; these particles are N (nitrogen atom), O (oxygen atom), He (Helium atom), $N_2$ (nitrogen molecule), and $O_2$ (oxygen molecule). However, once an ion particle has been produced, it can transfer its charge to another neutral particle when it collides with a neutral particle. Typically, the following charge transfer reactions are taken into account in modern ionospheric models (e.g. Huba et al. 2000):

$$H^+ + O \rightarrow O^+ + H$$

$$He^+ + N_2 \rightarrow N_2^+ + He$$

$$He^+ + N_2 \rightarrow N^+ + N + He$$

$$He^+ + O_2 \rightarrow O^+ + O + He$$

$$He^+ + O_2 \rightarrow O_2^+ + He$$

$$N^+ + O_2 \rightarrow NO^+ + O$$

$$N^+ + O_2 \rightarrow {O_2}^+ + N(^2D)$$

$$N^+ + O \rightarrow O^+ + N$$

$$N^+ + NO \rightarrow NO^+ + O$$

$$O^+ + H \rightarrow H^+ + O$$

$$O^+ + N_2 \rightarrow NO^+ + N$$

$$O^+ + O_2 \rightarrow {O_2}^+ + O$$

$$O^+ + NO \rightarrow NO^+ + O$$

$${N_2}^+ + O \rightarrow NO^+ + N(^2D)$$

$${N_2}^+ + O_2 \rightarrow {O_2}^+ + N_2$$

$${N_2}^+ + O_2 \rightarrow NO^+ + NO$$

$${N_2}^+ + NO \rightarrow NO^+ + N_2$$

$${O_2}^+ + N \rightarrow NO^+ + O$$

$${O_2}^+ + N(^2D) \rightarrow N^+ + O_2$$

$$O_2{}^+ + NO \rightarrow NO^+ + O_2$$

$$O_2{}^+ + N^2 \rightarrow NO^+ + NO.$$

When an ionized particle collides with a free electron, it can become neutral again via the charge recombination reaction (e.g. Huba et al. 2000):

$$H^+ + e = H$$

$$He^+ + e = He$$

$$N^+ + e = N$$

$$O^+ + e = O$$

$$N_2{}^+ + e = N_2$$

$$NO^+ + e = NO$$

$$O_2{}^+ + e = O_2.$$

While other ionized particles are present in the Earth's atmosphere, the ionosphere primarily consists of seven ions that constitute the majority of charged particles in this region ($H^+$, $He^+$, $N^+$, $O^+$, $N_2{}^+$, $NO^+$, and $O_2{}^+$), with $O^+$ being the most abundant overall.

Once charged particles originate in the atmosphere due to the ionization, they begin to interact with other charged and neutral particles. Collisions and other types of interactions impose a drag on the ions and electrons. The magnitude of the drag is proportional to the difference between the velocity of the ion and the velocity of the interacting particle:

$$\frac{\partial v_j}{\partial t} = \nu_{ij}.(v_i - v_j)$$

In the above equation, $v_i$ and $v_j$ are the velocities of the particle imposing the drag, which can be another charged or neutral particle, and of the ionized particle itself, respectively; $\nu_{ij}$ is the *collision drag coefficient*.

Collisions between ionized and neutral particles also result in a variety of heating processes which include ion-neutral frictional heating; ion-ion and ion-electron collisional heating; elastic and non-elastic heating; and rotational, vibrational, fine structure and photoelectron heating/cooling for electrons. Parametrizations approximating the effect of these quantum mechanical processes have been developed (e.g. Banks and Kockarts 1973) and are widely used in numerical models of the ionosphere. Such parametrizations aim at estimating the time derivative of the temperature of a particular charged particle $j$ as a function of its temperature ($T_j$) and the temperatures of other charged particles, $T_i$:

$$\frac{\partial T_j}{\partial t} = \sum_i Q_{ij} \cdot f(T_i, T_j)$$

In addition to these processes, heat transfer must also be modelled using the standard heat transfer equation with particle-specific thermal diffusivity coefficients.

## *3.2 Transport and Solar Effects*

Transport of charged particles is primarily driven by their interaction with the Earth's magnetic field and the superimposed external magnetic fields originating from space. These effects are dramatically different in the neighbourhood of magnetic poles, where extraterrestrial charged particles can enter the atmosphere, and at low and middle latitudes, where the Earth's magnetic field shields the atmosphere from space plasma.

Primarily, charged particles move along the magnetic field lines. Neutral particles, whose dynamics is determined mainly by solar heating, rotation of the Earth, and wave-driven momentum transfer between the troposphere and upper atmosphere, affect the ionosphere through neutral-ion drag. When neutral particles drag charged particles across the magnetic field lines, either originating from the Earth's own magnetic field or superimposed by space plasma, under the influence of the *Lorentz force* the charged particles move in a direction perpendicular to both the magnetic field lines and the drag direction.

The horizontal component of the neutral winds is significantly larger than the vertical one, at least at ionospheric altitudes. Since the Earth magnetic field lines are oriented mainly vertically near the poles, the Lorentz force makes charged particles move in a horizontal direction, forming vortex-like cells. External magnetic fields due to solar wind and coronal mass ejections significantly influence these circulation patterns in a complex fashion. Data on the spatial and temporal behaviour of the solar plasma is scarce and efforts have been undertaken to develop parametrizations aiming at describing the superimposed magnetic field as a function of easily observable solar activity, such as the *solar flux* at 10.7 cm and other parameters (e.g. Weimer 2001). An example of horizontal trajectories of electrons in the ionosphere over the North Pole resulting from such processes, and computed using the Weimer parametrization is shown in Fig. 1.

The complexity of ionospheric processes at high altitudes results in a somewhat artificial, but often practically necessary, separation of numerical modelling efforts into high latitude, and mid and low latitude parts. In this chapter we focus on the low and mid latitude modelling. For a description of high latitude processes, the reader is referred, for example, to Schunk (1988) and Schunk and Nagy (2000).

At low latitudes and the magnetic Equator, the Earth magnetic field lines are aligned nearly horizontally. The neutral winds alternate between easterlies and westerlies at day and night. This results in plasma being driven upward by the Lorentz force during the day and downward at night (the so-called "$\mathbf{E} \times \mathbf{B}$ drift").

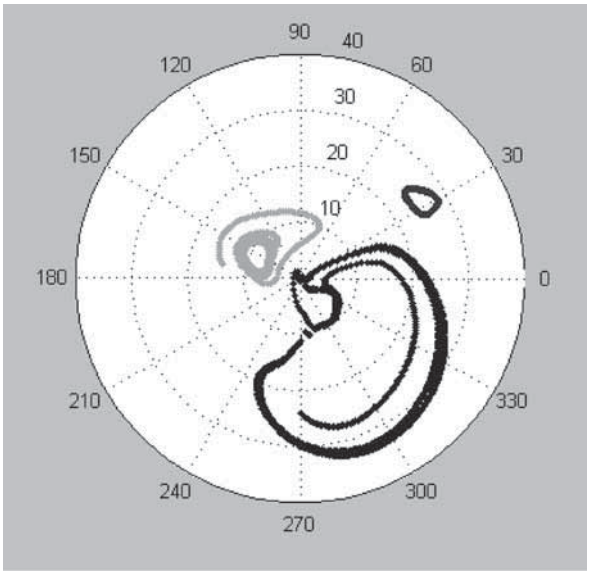



**Fig. 1** Examples of trajectories of electrons moving in the ionosphere over the North Pole computed using the Weimer (2001) parametrization

The magnitude of this effect decreases with altitude as the density of the neutral atmosphere decreases. Once the plasma is transported upwards by this so-called equatorial fountain, it descends northward and southward along the Earth's magnetic field lines forming two pronounced maxima, located at about $\pm 15°$ off the magnetic Equator. An example of the spatial distribution of the plasma densities obtained in the described numerical model is shown in Fig. 2.

In order to avoid complexities arising from modelling these electromagnetic interactions from first principles, parametrizations have been developed allowing one to readily approximate the vertical velocity of the equatorial plasma as a function of neutral wind velocity, time of day, and solar flux (e.g. Fejer and Scherliess 1995).

**Fig. 2** A 3-D iso-surface of electron densities in the ionosphere illustrating the effect of the equatorial fountain. Note the two pronounced maxima located north and south of the magnetic Equator. The calculation uses the Fejer and Scherliess (1995) empirical model

Ionospheric numerical models must take into account both the elementary processes and transport processes described above in order to approximate the spatial distribution and temporal behaviour of the ionospheric plasma. In the following sections we describe a practical implementation of a numerical system taking these processes into account and a related data assimilation methodology.

## 4 Modelling the Ionosphere

The ionospheric model described here is a numerical global model of the ionosphere loosely based on the description given in Bailey and Balan (1996), Fuller-Rowell et al. (1996), Millward et al. (1996), and Huba et al. (2000). More information about the model and its validation can be found in Khattatov et al. (2005).

The model computes the spatial distribution and temporal evolution of seven major ions ($H^+$, $O^+$, $He^+$, $O_2{}^+$, $NO^+$, $N_2{}^+$ and $N^+$) and electrons. Other prognostic variables include ion and electron temperatures and velocities.

The model solves the plasma dynamics equations for the seven ion species and electrons, and the energy conservation equation for the three major ions and electrons. It includes chemical interactions with neutral particles and ion-ion and ion-neutral collision rates, photoionization, and several different types of heating. The model variables are described in Table 1 below.

The model domain covers all latitudes and longitudes; however, polar transport and source terms are currently turned off. As is customary in ionospheric applications, the dynamic equations are solved in *magnetic coordinates* since in the

**Table 1** Model variables and constants

| Symbol | Quantity |
|---|---|
| $N_i$ | Density of ion $i$ |
| $V_i$ | Field-aligned velocity of ion $i$ |
| $T_i$ | Temperature of ion $i$ |
| $N_e$ | Electron density |
| $V_e$ | Electron field aligned velocity |
| $T_e$ | Electron temperature |
| $P_i$ | Chemical production for ion $i$ |
| $L_i$ | Chemical loss for ion $i$ |
| $I$ | Magnetic inclination angle |
| $D$ | Magnetic declination angle |
| $U_n$ | Zonal neutral velocity |
| $V_n$ | Meridional neutral velocity |
| $\nu_{ij}$ | Ion-ion collision frequency |
| $\nu_{in}$ | Ion-neutral collision frequency |
| $Q,F$ | Ion heating rates |
| $g$ | Acceleration of gravity |
| $m_i$ | Mass of ion $i$ |
| $\kappa$ | Thermal conductivity |
| $k$ | Boltzmann constant |

absence of electric fields plasma predominantly moves parallel to the direction of the magnetic field. A detailed discussion of the coordinate transformation and related equations can be found, for instance, in Millward et al. (1996). Here we give only a brief overview for the benefit of readers not familiar with this subject.

The Earth's magnetic field is approximated as that of a tilted eccentric dipole. The first magnetic coordinate is magnetic longitude. For each magnetic longitude we consider a "stack" of magnetic field lines characterized by the distance of the apex of each line from the Earth's centre at the magnetic Equator. This distance, normalized by the Earth's radius, is the second magnetic coordinate, $p$. Finally, for each field line the distance from the apex to a particular point along the line gives the third coordinate, $s$.

A portion of the model grid for low latitudes only is shown in Fig. 3 below. For clarity, only select gridpoints are shown; the regular model configuration is $100 \times 100 \times 100$.

Once the field-aligned transport is solved for, the model computes the plasma evolution due to cross-field transport. As discussed above, cross-field transport (the so-called "$\mathbf{E} \times \mathbf{B}$ transport") is forced by electric fields either imposed externally from the magnetosphere or generated internally from the action of the neutral wind. In the lower thermosphere the mobility of the ions is inhibited by collisions with the neutral atmosphere. The dynamo action of the neutral winds drives currents that through continuity create polarized electric fields. The ions respond to these electric fields by drifting perpendicularly to both the electric and magnetic fields. In non-fully coupled models, like the one described here, the $\mathbf{E} \times \mathbf{B}$ plasma velocity is specified from external empirical models (e.g. Fejer and Scherliess 1995). Once this velocity is known, solving the plasma advection equation is relatively straightforward.

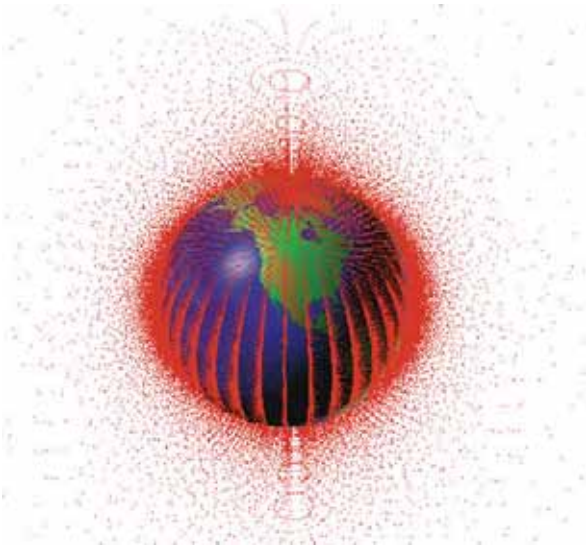


**Fig. 3** A portion of the 3-D model grid

The densities and velocities of neutral particles in the model domain are obtained from the well-known empirical models of Hedin (1991) and Hedin et al. (1996).

We now look at some specific features of the model.

*Continuity equation for each ion*: The numerical solution of the continuity equation should generate an ion density $N_i$ given all related variables at time $t$. This equation can be written in the magnetic coordinates as follows

$$\frac{\partial N_i}{\partial t} - b_s^2 \frac{\partial \left(\frac{N_i V_i}{b_s}\right)}{\partial s} + N_i \cdot \nabla V_\perp + \nabla N_i \cdot V_\perp = P_i - L_i \cdot N_i$$

where

$$b_s = \sqrt{1 + 3\cos^2(\text{eccLat})} \cdot \left(\frac{\text{R}_\text{e}}{\text{eccRadius}}\right)^3$$

$$\nabla V_\perp = \frac{6 \cdot V_\perp^{\text{eq}} \sin^2(\text{eccLat}) \cdot (1 + \cos^2(\text{eccLat}))}{p \cdot R_e \cdot (1 + 3 \cdot \cos^2(\text{eccLat}))^2}$$

eccRadius and eccLat are the radius and latitude of a particular point on the field line in eccentric coordinates (for definition of eccentric coordinates see, for example, Bailey and Balan 1996) and $V_\perp^{\text{eq}}$ is the value of $\mathbf{E} \times \mathbf{B}$ drift at the magnetic Equator.

*Momentum conservation equation*: The numerical solution of the momentum conservation equation generates the ion velocity $V_i$ given all related variables at time $t$. In general,

$$V_i = \frac{1}{\sum_{n=1}^{N\_\text{Neutrals}} \nu_{in} + \sum_{j=1}^{N\_\text{Ions}} \nu_{ij}} \cdot \left[ \begin{array}{l} -g \sin I + \frac{b_s k_i}{m_i} \left( \frac{T_i}{N_i} \frac{\partial N_i}{\partial s} + \frac{T_e}{N_e} \frac{\partial N_e}{\partial s} + \frac{\partial (T_i + T_e)}{\partial s} \right) \\ + \sum_{n=1}^{N\_\text{Neutrals}} \nu_{in} (V_n \cos D - U_n \sin D) \cos I + \sum_{j=1}^{N\_\text{Ions}} \nu_{ij} V_j \end{array} \right]$$

*Energy conservation equation*: The solution of this equation generates an ion temperature $T_i(t+\Delta t)$ given all related variables at time $t$.

$$\frac{3}{2} k N_i \left( \frac{\partial T_i}{\partial t} + V_\perp \nabla T_i \right) = k N_i T_i b_s^2 \frac{\partial}{\partial s} \left( \frac{V_i}{b_s} \right) - k N_i T_i \cdot \nabla V_\perp + b_s^2 \frac{\partial}{\partial s} \left( \kappa \frac{\partial T_i}{\partial s} \right) + \frac{3}{2} k N_i V_i b_s \frac{\partial T_i}{\partial s} + Q + F$$

Since the plasma heating rates $Q$ and $F$ are, generally, non-linear, this equation has to be solved iteratively.

*Electron temperature equation*: The electron temperature equation is similar to the ion temperature equation (see immediately above), except that the conductivities ($\kappa$) and heating rates ($Q$, $F$) are computed for electrons.

*Electron density equation*: The electron density is computed using the assumption that overall the plasma is neutral, that is, electrons mainly originate when a neutral particle is ionized. Thus electron densities are simply a sum of the densities of all ion particles.

$$N_e = \sum_{i=1}^{\text{Number of Ions}} N_i$$

*Electron velocity equation*: The electron velocity equation assumes that there are no field-aligned currents:

$$V_e = \frac{\sum_{i=1}^{\text{Number of Ions}} V_i N_i}{N_e}$$

## 5 GPS Data

Continuous soundings of the ionosphere by the Global Positioning System (GPS) yield a convenient source of data for assimilation into ionospheric models. The GPS (Parkinson and Spilker 1996) consists of several segments. As of mid 2008, the space part of the GPS is a constellation of 30 satellites located in orbits about 20,000 km above the Earth's surface. At any given time at least 4 satellites are visible from most locations on Earth. Put simply, each GPS satellite emits a signal that contains a unique identification code (*pseudo random number*, PRN), accurate time from an onboard atomic clock, satellite position, clock corrections, and auxiliary information. A GPS receiver that has 4 or more GPS satellites in view can, by "triangulation" in space and time, determine both precise time and position.

The ground-based portion of the GPS consists of a network of several hundred stations hosting highly accurate receivers whose locations are known to within a millimetre. The stations continuously receive GPS satellite signals and deliver the received data to several GPS computing centres. These centres utilize precise station coordinates and the received GPS data in order to estimate GPS satellite ephemeris, satellite clock corrections, satellite and station differential code biases (described below), and other parameters. At regular intervals this information is uploaded to each GPS satellite for broadcasting to end users.

GPS L-band frequency signals are delayed by the ionosphere by a time approximately proportional to the total integrated electron content along the line of sight between a receiver and a GPS satellite. These delays can result in positional errors

of tens and occasionally hundreds of metres. To mitigate this effect GPS satellites and high-end GPS receivers use signals at 2 different frequencies referred to as L1 and L2. Since the ionosphere is a dispersive media at these frequencies, it is possible to estimate and remove the ionospheric delays and therefore estimate the slant *total electron content* (TEC) for each dual-frequency receiver-satellite pair.

Estimation of the slant TEC using dual-frequency GPS receivers is made difficult by the presence of so-called *differential code biases* (DCBs) in both satellite transmitters and the receivers. Slight differences in the hardware channels of the antennas and the receiver or transmitter introduce additional propagation time differences to signals at the L1 and L2 frequencies. Satellite DCBs are generally small, equivalent to several centimetres. GPS receiver DCBs can be quite large, often reaching metres or tens of metres. Traditionally, these biases are estimated in the off-line mode after accumulating a time series of measurements from a number of reference stations (e.g. Mannucci et al. 1998). A combination of a 3-D numerical model with a specialized data assimilation scheme allows one to estimate these biases on-line, for example in a scheme where the unknown state (electron densities at each model grid point) is augmented with the DCB values for each utilized receiver.

## 6 Ionospheric Data Assimilation

The data assimilation approach adopted here resembles that of Khattatov et al. (2000). Let us assume that model estimates of electron densities at all grid points at time $t$ are arranged in a vector $\mathbf{x}$ with dimension $N_x$. Formally, integration of the non-linear model $\mathcal{M}$ can be written as

$$\mathbf{x}_{t+\Delta t} = \mathcal{M}(t, \mathbf{x}_t) \tag{1}$$

Let vector $\mathbf{y}$ contain observations of a quantity linearly related to electron densities at the same time. In the case of GPS reference station data such quantities are slant TEC from each station to all satellites in view.

The connection between $\mathbf{x}$ and $\mathbf{y}$ is established via linear interpolation and integration of the non-linear observation operator $\mathcal{H}$ as follows:

$$\mathbf{y} = \mathcal{H}(\mathbf{x}). \tag{2}$$

Under assumptions of linearity (e.g. $\mathcal{H}$ is replaced by the linear $\mathbf{H}$, $\mathcal{M}$ is replaced by the linear $\mathbf{M}$) and Gaussian statistics, the optimal value of $\mathbf{x}$ that inverts Eq. (2) given a set of observations $\mathbf{y}$ and model estimates of $\mathbf{x}$ is given by the Kalman filter (for example, see chapter *Mathematical Concepts of Data Assimilation*, Nichols):

$$\mathbf{x}_t^a = \mathbf{x}_t + \mathbf{K}(\mathbf{y} - \mathbf{H}\mathbf{x}_t) \tag{3}$$

$$\mathbf{K} = \mathbf{B}_t\mathbf{H}^{\mathrm{T}}(\mathbf{H}\mathbf{B}_t\mathbf{H}^{\mathrm{T}} + \mathbf{O} + \mathbf{R})^{-1} \tag{4}$$

Here $\mathbf{B}_t$ is the forecast error covariance at time $t$. $\mathbf{O}$ is the error covariance matrix of the observations and $\mathbf{R}$ is the representativeness error covariance associated with errors of interpolation and discretization. Matrix $\mathbf{K}$ is called the Kalman gain matrix.

The analysis error covariance is expressed as:

$$\mathbf{B}_t^a = \mathbf{B}_t - \mathbf{B}_t\mathbf{H}^{\mathrm{T}}(\mathbf{H}\mathbf{B}_t\mathbf{H}^{\mathrm{T}} + \mathbf{O} + \mathbf{R})^{-1}\mathbf{H}\mathbf{B}_t. \tag{5}$$

Once inversion of Eq. (2) is performed, the derived electron densities, $\mathbf{x}_t{}^a$, can be used as the initial condition for the model $\mathbf{M}$ to predict electron densities at a later time (at the beginning of the next assimilation window) according to Eq. (1).

Since the model domain contains about $10^6$ points, direct matrix manipulations described by Eqs. (3), (4), and (5) are generally impossible to implement explicitly in practice. As customary in large-scale Kalman filter implementations in NWP and other areas of atmospheric sciences (see, e.g., chapter *Constituent Assimilation*, Lahoz and Errera), we track only the evolution of the diagonal of the background error covariance matrix and parametrize the off-diagonal elements. If we also assume that correlations between variations of electron densities at two different points become negligible at certain distances (i.e., we assume compactly supported error covariance models), the matrices become sparse and the above-mentioned calculations become possible. Further details can be found in Khattatov et al. (2000).

We argue that since plasma equations are solved in magnetic coordinates, in order for the background error covariance to be separable, the correlation lengths and distances between points need to be specified in magnetic rather than geographic coordinates. This is the approach adopted here.

Let us now recall that measurements of slant TECs by GPS reference stations often contain large unknown biases. Formally, slant TEC measured by a particular station is a sum of "true" and "unknown" slant TEC and receiver and satellite biases (denoted by $b_r$ and $b_s$, respectively):

$$y_{\text{observed}} = y_{\text{true}} + b_r + b_s$$

Depending on the local time, the magnitude of these biases is often significantly larger than the actual slant TEC. Clearly, in order for the data to be useful, these biases need to be determined and accounted for (see chapter *Bias Estimation*, Ménard). This can be accomplished by augmenting the state vector $\mathbf{x}$ with satellite and receiver biases. Since the size of $\mathbf{x}$ is $\sim 10^6$, the number of satellite biases is ~30 (corresponding to the number of operational GPS satellites), and the number of GPS stations we currently use is ~100, this does not lead to any significant increase in the number of unknowns.

In principle, both satellite and station biases can be continuously computed using this method. At this stage in the development process we choose to use fixed broadcast satellite biases and only determine receiver DCBs. This can be justified by

noting that satellite biases are perceived to be generally better known than the receiver biases and are slower varying.

The assimilation system consisting of the numerical model and the GPS data assimilation scheme described above has been extensively validated using various methodologies. Further discussion, including validation results can be found in Khattatov et al. (2005).

## 7 Impact of Ionosphere on Telecommunications, Scintillations

High frequency (metre length) electromagnetic waves can bounce off the ionosphere to the surface and back several times, thus potentially travelling thousands of kilometres between the transmitter and a receiver and enabling very-long range communications. On the other hand, when ionospheric conditions are disturbed and the radio signal is absorbed or scattered instead of being reflected, such a propagation mode may not be possible.

The need for such long-range communications decreased with the advent of ground-to-satellite and satellite-to-ground communications. Yet the ground-space communications also are affected by the ionosphere, at least at some frequencies. In particular, our increasing reliance on the GPS makes adverse ionospheric conditions interfering with such communications particularly perceptible. This has driven the need for accurate nowcasting and forecasting of ionospheric conditions.

One of the most disruptive phenomena affecting transmission of electromagnetic signals in the ionosphere are scintillations, or plasma bubbles. These are small scale perturbations of the ambient electron densities that, under certain conditions (*Rayleigh-Taylor instability*), can rapidly grow and result in partial or complete fading of the GPS signal. This effect is akin to not being able to see through a body of water when waves are present at the surface.

An example of such interruption is shown in Fig. 4. The figure depicts slant electron content between a stationary GPS receiver and several GPS satellites over 24 h. Different lines correspond to different GPS satellites. As the satellites appear above the horizon or move to lower elevation angles the TEC values are large. At local night time (hours 15–24 in Fig. 4) the TEC values are quite low. Near sunset, at around 1200 UTC, the GPS signals are completely interrupted. This is followed by period of abrupt variability in the GPS signal, indicative of ionospheric disturbances.

Thus, the capability to forecast scintillations and to warn the end user about possible upcoming service interruptions can be rather valuable.

One can separate approaches to modelling and forecasting scintillations into two major classes. The first approach aims at explicitly resolving growth and evolution of these instabilities with very high-resolution fluid dynamics code, and diagnosing favourable conditions for appearance of scintillations. This approach is likely not to be practically feasible on a global scale for a number of years due to the necessity to resolve very small spatial scales.

The “probabilistic” second approach (e.g. Sultan 1996; Secan et al. 1995) uses ambient properties of the ionospheric plasma such as ion and electron densities,

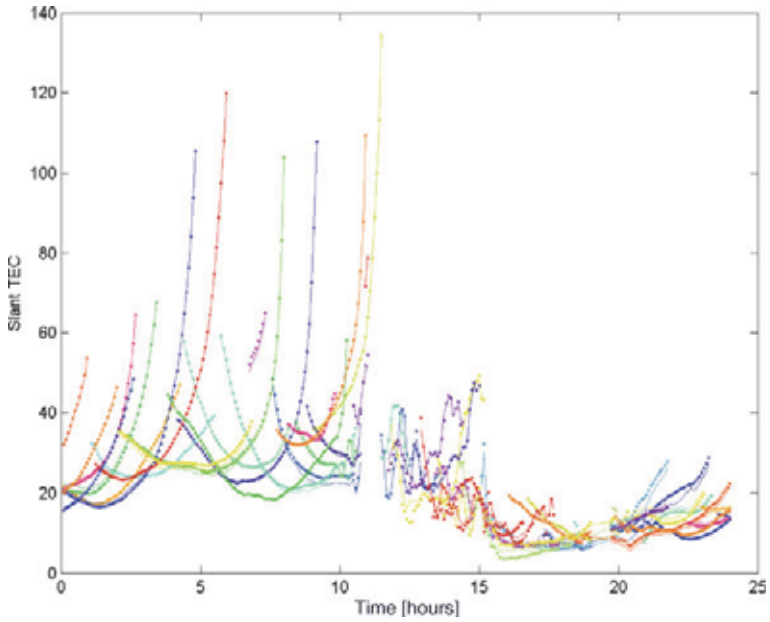



**Fig. 4** Slant TEC estimates from the model (*solid lines*) and GPS receiver (*dots*) as a function of GPS time for August 24, 2005. Note the sharp onset of scintillations at GPS time (UTC time) 1100 just after the local sunset

and density gradients; ion and density collision frequencies; recombination rates; inclination and declination of the magnetic field lines; and ion velocities and temperatures, to compute the linear growth rate of plasma instabilities as a function of geographic location and time. The greater the growth rate, the higher the likelihood of severe scintillations.

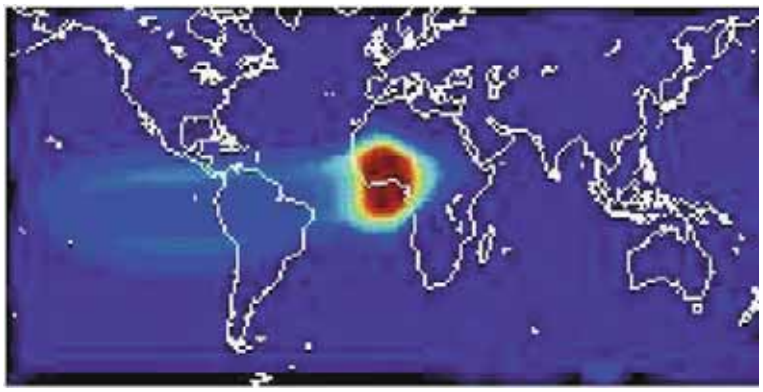


**Fig. 5** A map showing diagnosed scintillation growth rates. Relatively higher rates are marked in *red*

Since all the necessary parameters are routinely computed in the assimilative numerical model described here, it is relatively straightforward to diagnose the scintillations' growth rates. An example of the calculated average growth rates is shown in Fig. 5. Regions evidenced clearly distinguishable in the picture have favourable conditions for occurrence of scintillations.

## 8 Application to Single-Frequency GPS Positioning

One of the main practical applications of the developed ionospheric specification system is the augmentation of positioning capabilities of single-frequency GPS receivers. Free electrons in the ionosphere delay the group velocity and advance the phase velocity of the GPS signals emitted by GPS satellites. This and other sources of errors (e.g. offsets in satellite clocks, inaccurate ephemeris) impact the accuracy of the derived distance between the receiver and a satellite (the "pseudorange"). Since the ionosphere is normally non-homogeneous, ionospheric delays are different for different visible GPS satellites. If these delays are not properly accounted for, they will interfere with the GPS receiver's ability to accurately compute its position.

Slant electron content of 1 TEC unit ($10^{16}$ electrons $m^{-2}$) between a receiver and a GPS satellite results in approximately a 16 cm positioning error on the L1 frequency along the line of sight between the receiver and the satellite. Thus, a relatively common slant ionospheric electron content of 50 TEC units results in approximately an 8 m positioning error, which is significant for many applications. Ionosphere-induced signal propagation delays are the major source of errors for single-frequency GPS receivers.

Dual-frequency receivers can estimate and remove ionospheric delays from their measurements; however, this process requires additional time. This is related to the fact that the TEC estimated from pseudorange measurements is very noisy. In order to obtain usable TEC values, pseudorange-derived TEC needs to be smoothed using phase measurements in a process called "phase levelling" (see Blewitt 1990 for a discussion). This process can require between 10 and 60 min of data collection during which the receiver will not be able to compute its position accurately.

Therefore, externally-supplied accurate estimates of slant delays between a GPS receiver and visible GPS satellites can be very useful to users of both single-frequency and dual-frequency receivers.

The ionospheric assimilation system described in Sect. 6 has been used in a GPS augmentation system consisting of several integrated components:

- A software package that communicates with a user's GPS receiver attached to a laptop or a PDA (Personal Digital Assistant). The software collects raw pseudorange and phase data from the GPS receiver;
- A software package that communicates this collected data to a backend computing server via the Internet and receives processed results from the backend server;
- An ionospheric modelling and assimilation system that generates ionospheric delays for the approximate receiver location to visible GPS satellites;

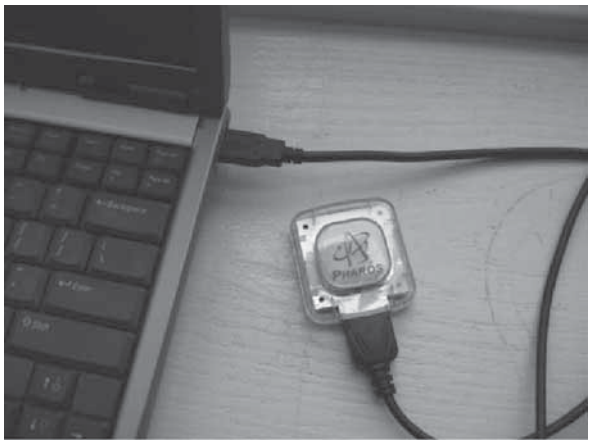



**Fig. 6** A photograph of a Pharos single frequency GPS receiver attached to a notebook computer

- A positioning engine (courtesy of GPS Solutions Inc) that computes accurate receiver coordinates, accounting for ionospheric delays and precise GPS satellite orbits and clocks.

The augmentation system's primary application is enhancing accuracy and robustness of positioning with inexpensive single frequency receivers such as those produced by Pharos or Garmin manufacturers. Preliminary results have demonstrated meaningful practical improvements in positioning quality. The rest of this section outlines the system design and presents sample positioning results.

At the hardware level, the end user has a laptop computer or a PDA device with a single frequency GPS receiver attached to it via a USB cable or Bluetooth. Results shown here demonstrate positioning functionality with an off-the-shelf $200 Pharos GPS receiver shown in Fig. 6. The setup used to collect data in the field is illustrated in Fig. 7.

**Fig. 7** Collecting data in the field with a Pharos receiver attached to a laptop computer

The end-user software allows the user to start and stop the collection of raw data from the GPS receiver and shows visible GPS satellites. After completing data collection the user can upload collected data files to the server via the Internet. This can be done either in the field, if the user has wireless Internet connection, or later in a post-processing mode. The software running on the backend server authenticates the user, and archives the collected data files and related metadata. It then invokes a program that computes slant total electron content between the approximate user receiver position and all visible GPS satellites at that time, from archived 3-D distributions of the ionospheric electron content.

Once slant TEC values are computed, these data together with the receiver raw data and metadata files are passed on to the positioning engine. The positioning engine uses predicted precise GPS satellite orbits and clock corrections to compute satellite positions and then computes an accurate user position after adjusting raw receiver data for ionospheric delays. The computed position together with graphical auxiliary data is transmitted back to the user.

Figure 8 shows typical errors in the positioning solutions computed by the receiver itself (blue dots) and by the described augmentation service (red dots). Each dot corresponds to a single receiver measurement, normally reported every second.

Note that the receiver solution shown here was not aided by the US Federal Aviation Administration's Wide Area Augmentation Service (WAAS). Clearly, the service is capable of significantly improving positioning accuracy. Figure 9 shows similar data, but in the case when the receiver was receiving the WAAS signal.

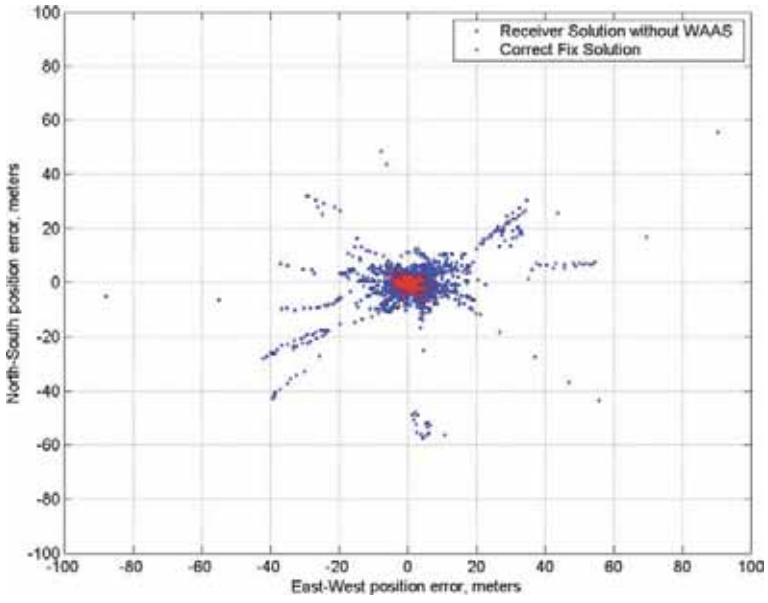



**Fig. 8** Positioning errors (metres) of standalone receiver (*blue*) the augmentation service (*red*). No WAAS augmentation is employed by the receiver. *x*-axis: East-West position error; *y*-axis: North-South position error

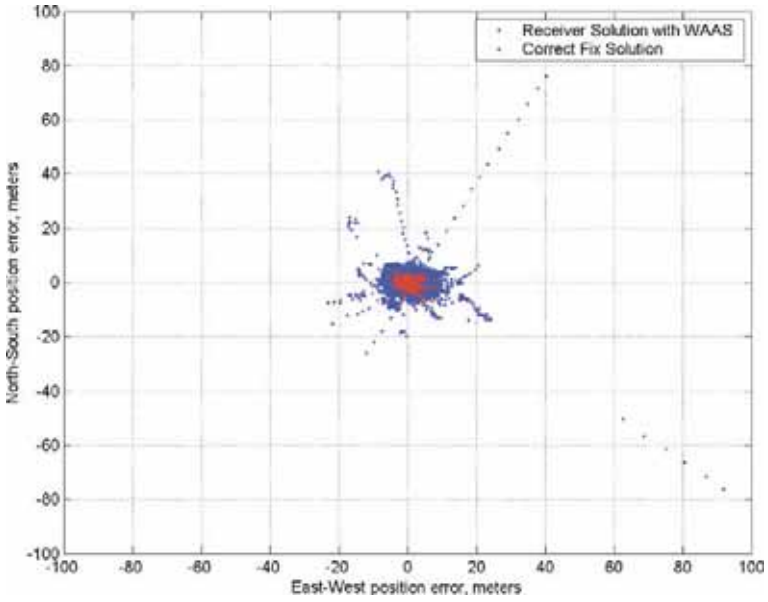



**Fig. 9** Similar to Fig. 8 but with WAAS augmentation

As these results show, inexpensive single frequency GPS receivers can be used to achieve positioning accuracy of 1–2 m with appropriate augmentation relying on data from ionospheric assimilation models and using precise GPS orbits and clock corrections.

## 9 Future Directions

A lot of research, resources and time are required for proper forecasting of the solar drivers that affect ionospheric conditions. Thus, mature medium- to long-term forecasting of the ionosphere, which is a very worthy scientific and engineering goal, is only likely to be realized far into the future. Yet, as we show here, nowcasting and accurate post-processing are achievable at the present time and have practical uses. In this author's opinion near future advancement of ionospheric assimilation will likely proceed along the following directions:

- *Increasing amounts of assimilated GPS data*: This is dependent on installing new ground GPS stations and making their data available to users. Since significant financial expenditure is needed for installing a station, this process is likely to be gradual;
- *Assimilation of data from new sources*, primarily GPS occultation measurements such as those obtained from the recently launched COSMIC (Constellation Observing System for Meteorology Ionosphere and Climate) satellite (see, e.g.,

Rocken et al. 2000). Addition of occultation and in situ measurements will provide a wealth of information on the vertical distribution of electron densities which is not readily obtained from ground-based GPS data;

- *Making ionospheric assimilation models operational and their results readily available to interested end users*: As is well known from numerical weather forecasting, which has evolved from a research setting to a robust and reliable operational implementation, this evolution is a non-trivial engineering task that requires significant human and technical resources. Yet such evolution is necessary in order to make ionospheric nowcasting usable in practice;
- *Decreasing data latencies*: A large class of GPS users are interested in having ionospheric information in real time for purposes such as surveying, high-precision agriculture, construction and others. Yet a majority of freely available, public GPS observations that can be assimilated into models are available with delays ranging from minutes, to hours to days. Many brands of modern geodesic quality GPS receivers can simply be "plugged in" to the Internet and stream their data in real time. However, in practice, even if the owner of such a receiver decides to make the data publicly available, authentication and data access to a network of such receivers present a hurdle. The recently developed NTRIP protocol (http://igs.bkg.bund.de/ntrip/NTRIP.htm) and related software provide a way for moving toward a solution to this problem. In addition, obtaining such data, assimilating it, and delivering to users the resulting product with very short latencies (~1–10 s) is a complex engineering task.

**Acknowledgments** The author is very thankful to Dr. T. Fuller-Rowell and Dr. O. de la Beaujardiere for their interest in the work described here, encouragement, and guidance. A number of skilled computer scientists and software engineers helped to develop the numerical model and data assimilation scheme described here in a short time. GPS Solutions Inc was an integral part of developing the positioning augmentation service and provided the positioning engine. This work has primarily been funded by the US Air Force Research Laboratory in Hanscom, MA.

## References

Bailey, G.J. and N. Balan, 1996. A low-latitude ionosphere-plasmosphere model. In *STEP: Handbook of Ionospheric Models*, Schunk, R.W. (ed.), STEP Report.

Banks, P.M. and G. Kockarts, 1973. *Aeronomy, Parts A and B*, Academic Press, Inc. New York, 430 pp (Part A), 355 pp (Part B).

Blewitt, G., 1990. An Automatic editing algorithm for GPS data. *Geophys. Res. Lett.*, **17**, 199–205.

Fejer, B.G. and L. Scherliess, 1995. Time dependent response of equatorial ionospheric electric fields to magnetospheric disturbances. *Geophys. Res. Lett.*, **22**, 851–854.

Fuller-Rowell T.J., D. Rees, S. Quegan, R.J. Moffett, M.V. Codrescu and G.H. Millward, 1996. A coupled thermosphere ionosphere model (CTIM). In *Handbook of Ionospheric Models*, Schunk, R.W. (ed.), STEP Report.

Hajj, G.A., B.D. Wilson, C. Wang, X. Pi and G. Rosen, 2004. Data assimilation of ground GPS total electron content into a physics-based ionospheric model by use of the Kalman filter. *Radio Sci.*, **39**, doi:10.1029/2002RS002859.

Hedin, A.E., 1991. Extension of the MSIS thermosphere model into the middle and lower atmosphere. *J. Geophys. Res.*, **96**, 1159–1172.

Hedin, A.E., E.L. Fleming, A.H. Manson, F.J. Scmidlin, S.K. Avery, R.R. Clark, S.J. Franke, G.J. Fraser, T. Tsunda, F. Vial and R.A. Vincent, 1996. Empirical wind model for the upper, middle, and lower atmosphere. *J. Atmos. Terr. Phys.*, **58**, 1421–1447.

Huba, J., G. Joyce and J. Fedder, 2000. SAMI2 is another model of the ionosphere (SAMI2): A new low-latitude ionosphere model, *J. Geophys. Res*., **105**, 23035–23053.

Khattatov, B.V., J.-F. Lamarque, L.V. Lyjak, R. Ménard, P. Levelt, X.X. Tie, G.P. Brasseur, G.P. and J.C. Gille, 2000. Assimilation of satellite observations of long-lived chemical species in global chemistry transport models. *J. Geophys. Res.*, **105**, 29135–29144.

Khattatov, B.V., M. Murphy, M. Gnedin, J. Sheffel, J. Adams, B. Cruickshank, V. Yudin, T. Fuller-Rowell and J. Retterer, 2005. Ionospheric nowcasting via assimilation of GPS measurements of ionospheric electron content in a global physics-based time-dependent model. *Q. J. R. Meteorol. Soc.*, **131**, 3543–3559.

Mannucci, A.J., B.D. Wilson, D.N. Yuan, C.H. Ho, U.J. Lindqwister and T.F. Runge, 1998. A global mapping technique for GPS-derived ionospheric total electron content measurements. *Radio Sci.*, **33**, 565–582.

Ménard, R., S.E. Cohn, L.P. Chang and P.M. Lyster, 2000. Stratospheric assimilation of chemical tracer observations using a Kalman filter, part I: Formulation. *Mon. Weather Rev.*, **128**, 2654–2671.

Millward, G.H., R.J. Moffett, S. Quegan and T.J. Fuller-Powell, 1996. A coupled thermosphere-ionosphere-plasmosphere model (CTIP). In *STEP: Handbook of Ionospheric Models*, Schunk, R.W. (ed.), STEP Report.

Parkinson, B.W. and J.J. Spilker Jr., 1996. *Global Positioning System: Theory and Applications* (Vols. 1 and 2). American Institute of Aeronautics, 370 L'Enfant Promenade, SW, Washington, DC.

Pi, X., C. Wang, G.A. Hajj, I.G. Rosen, B.D.Wilson and G. Bailey, 2003. Estimation of E_B drift using a global assimilative ionospheric model: An observation system simulation experiment. *J. Geophys. Res.*, **108**, 1075–1087.

Rocken, C., Y.-H. Kuo, W. Schreiner, D. Hunt, S. Sokolovskiy and C. McCormick, 2000. COSMIC System Description. *Terr. Atmos. Ocean. Sci.*, **11**, 21–52.

Scherliess, L., R.W. Schunk, J.J. Sojka and D.C. Thompson, 2004. Development of a physics-based reduced state Kalman filter for the ionosphere. *Radio Sci.*, **39**, RS1S04, doi:10.1029/2002RS002797.

Schunk, R.W., 1988. A mathematical model of the middle and high latitude ionosphere. *Pure Appl. Geophys*., **127**, 255–303.

Schunk, R. and A. Nagy, 2000. *Ionospheres,* Cambridge University Press, Cambridge, 570 pp.

Schunk, R.W., L. Scherliess, J.J. Sojka, D.C. Thompson, D.N. Anderson, M. Codrescu, C. Minter, T.J. Fuller-Rowell, R.A. Heelis, M. Hairston and B.M. Howe, 2004. Global Assimilation of Ionospheric Measurements (GAIM). *Radio Sci.*, **39**, doi:10.1029/2002RS002794.

Secan, J.R., R.M. Bussey, E.J. Fremouw and S. Basu, 1995. An improved model of equatorial scintillation. *Radio Sci*., **30**, 607–617.

Sultan, P.J., 1996. Linear theory and modeling of the Rayleigh-Taylor instability leading to the occurrence of equatorial spread. *J. Geophys. Res.*, **101**, 26875–26891.

Weimer, D.R., 2001. An improved model of ionospheric electric potentials including substorm perturbations and application to the GEM November 24, 1996 event. *J. Geophys. Res*., **106**, 407–416.

# Part VI
# The Longer View

# Reanalysis: Data Assimilation for Scientific Investigation of Climate

**Richard B. Rood and Michael G. Bosilovich**

## 1 Introduction

Reanalysis is the assimilation of long time series of observations with an unvarying assimilation system to produce datasets for a variety of applications; for example, climate variability, chemistry-transport, and process studies. Reanalyses were originally proposed for atmospheric observations as a method to generate "climate" datasets from "weather" observations. As the satellite records of chemical, land and oceanic parameters build with time, "reanalyses" are being developed for other types of observations. Coupled reanalyses, for example atmospheric-ocean reanalyses, are possible. In addition, very long reanalyses that use no satellite observations are being planned (e.g. Compo et al. 2006). Reanalysis datasets have become one of the most important datasets for scientific and application communities. As of July 2009, the Kalnay et al. (1996) paper, which describes one of the first reanalysis datasets, has more than 6,600 recorded citations. In this chapter discussion will be drawn from the experience of atmospheric reanalysis, and the issues raised are relevant to all types of reanalysis.

The provision of reanalyses was advocated by Bengtsson and Shukla (1988) and Trenberth and Olson (1988) in order to provide homogeneous datasets for climate applications and to encourage research in the use of satellite observations without the operational constraints of Numerical Weather Prediction. Trenberth and Olson (1988) calculated derived products, such as the Hadley circulation, from assimilation analyses used in operational weather forecasting. They found large discontinuities in time series of these derived quantities. The discontinuities were clearly linked to changes in the assimilation system, such as changes in the forecast model. Given the four-dimensional (time and space) nature of assimilated datasets and the success of assimilation in providing initial conditions for weather forecasting, it was logical to propose using a single, non-varying assimilation system to generate a long time series for the purpose of investigating the Earth's climate.

R.B. Rood (✉)
University of Michigan, Ann Arbor, MI, USA
e-mail: rbrood@umich.edu

W. Lahoz et al. (eds.), *Data Assimilation*, DOI 10.1007/978-3-540-74703-1_23,

Kalnay and Jenne (1991) proposed that a reanalysis be performed as a partnership between the National Meteorological Center (NMC, now part of the National Centers of Environmental Prediction, NCEP) and the National Center for Atmosphere Research (NCAR). This project required the preparation of the input datasets, the definition of the analysis system, and a data distribution plan. The analysis system was a version of the operational system used for weather prediction, but at lower resolution.

Three organizations performed a first generation of reanalyses in the spirit of Bengtsson and Shukla (1988) and Kalnay and Jenne (1991). Aside from the NCEP/NCAR reanalysis (Kalnay et al. 1996), the European Centre for Medium-Range Weather Forecasts (ECMWF) executed the ERA-15 project (Gibson et al. 1997) and the Data Assimilation Office (DAO, now the Global Modeling and Assimilation Office, GMAO) at NASA's Goddard Space Flight Center provided the 17-year Goddard Earth Observing System, Version 1 (GEOS-1) reanalysis (Schubert et al. 1993). These three reanalyses have been cited in many studies, which document successes as well as identifying a series of shortcomings that stand at the core of future research. New reanalyses have come from these and additional organizations.

The quality of the first-generation reanalyses is documented in the proceedings from two workshops (WCRP 1998, 2000; see also, Newson 1998). Kistler et al. (2001) gives an excellent overview of the NCEP/NCAR reanalysis project, and the discussions in that paper are relevant to all of the projects. Quantities that are directly constrained by the observations, i.e., temperature, geopotential, and the rotational component of the wind, are consistent across the three reanalyses. At the other extreme, quantities that are only weakly constrained by the observations or are dependent upon the physical parametrizations of the assimilating models differ greatly. Further, these derived quantities, which include the divergent component of the wind, precipitation, evaporation, clouds, fresh-water runoff, and surface fluxes, have significant uncertainties, as revealed either by independent validation or through applications in scientific studies.

Following this first set of reanalyses there is a second generation that strives to address some of the deficiencies of the first generation of reanalyses as well as to extend the reanalyses to earlier times. Kanamitsu et al. (2002) describe the incremental evolution of the original NCEP/NCAR reanalysis, which was performed in partnership with the United States Department of Energy (hence, NCEP/DOE reanalysis). ECMWF produced a 40 year reanalysis (Uppala et al. 2005) with significant incorporations of lessons learned. Both of these reanalyses did benefit from improvements to the assimilation system and from better treatment of the observations. However, there remain in these datasets some deficiencies that are, perhaps, intrinsic to reanalysis datasets. These deficiencies are related to the variability of the observational data stream and to the representation of the hydrometeorological and energy cycles. These subjects will be explored in more detail in this chapter. Reanalysis datasets, especially those generated at NASA's GMAO, have been used extensively in constituent transport applications (e.g. Bey et al. 2001; Douglass et al. 2003). Like studies involving the hydrometeorological cycle, constituent transport

studies require closed, physically consistent budgets. That is, the reanalysis products need to satisfy fundamental conservation equations. The results from these studies highlight that assimilated datasets do not satisfy conservation principles and, hence, are not physically consistent. The development of physically consistent assimilated datasets remains a research challenge (see chapter *The Role of the Model in the Data Assimilation System*, Rood).

In addition to the extensions of the original reanalysis efforts, there have been new reanalysis efforts. The JRA-25 was generated by the Japan Meteorological Agency and described by Onogi et al. (2007). This reanalysis has paid specific attention to improvement of precipitation, and the representation of global precipitation is improved relative to the ERA-40 and NCEP/DOE reanalyses. Mesinger et al. (2006) document NCEP's North American Regional Reanalysis (NARR), which is a high resolution regional reanalysis that uses the global reanalysis as the boundary conditions for a regional model-assimilation system.

Recently, first results from two new products were released. These are NASA's Modern Era Retrospective-analysis for Research and Applications (MERRA) and ECMWF's ERA-Interim (Links are provided at the end of the chapter.). These reanalyses have had significant attention paid to the input data stream, data quality control, bias correction, and the interface between the model and the analysis system. They are designed to address many of the problems discussed below. The first results suggest significant progress has been made.

Trenberth et al. (2008b) and Bengtsson et al. (2007) summarize the state of the art at the time of this writing and argue for continuing research to improve reanalysis. Based on the successes of reanalysis in climate science, there is broad agreement that the improvement, the extension, and the production of reanalyses are an essential element of the business of climate research. This is a rapidly changing field, with much of the current information found online from institutional and project websites. A snapshot of these activities is given at the end of this chapter to provide an introduction into current information. The chapter will next highlight some of the special aspects of the problem of data assimilation intended to be used in the scientific investigation of climate and constituent transport. This will be followed by sections on hydrometeorological applications of reanalysis and constituent transport applications. Finally, a discussion of the challenges for future reanalysis projects is presented. The references and examples here are expository and by no means comprehensive.

## 2 Special Aspects of the Reanalysis Problem

This section presents, first, lessons learned from reanalysis activities. Then two related aspects that provide difficult challenges to reanalysis, heterogeneity of the input data stream and bias, are discussed. The impact of data heterogeneity and bias on trends derived from assimilated datasets is then highlighted.

*Lessons learned.* The lessons learned from the first-generation reanalyses provided the foundation for a second generation of reanalyses. These lessons can

be summarized as general success in defining the major modes of variability on synoptic and planetary scales, as well as credible representation of the variability associated with longer-term, large-scale phenomena: e.g. monsoons, El Niño – La Niña, and the Madden-Julian oscillation. The deficiencies include fundamental problems in the hydrological cycle and the general circulation as well as artifacts in the reanalysis datasets that are directly related to changes in the observing network. The representation of tropical meteorological features is not as robust as the representation of the middle latitudes. The quality in the Arctic and Antarctic is highly variable (Bromwich et al. 2007).

Most of the primary references that describe reanalysis datasets and workshop reports have stated that reanalyses are not appropriate for trend studies (WCRP 1998, 2000; Newson 1998; Kistler et al. 2001). This is attributed, first, to the sensitivity of the assimilated dataset to changes in the observing system. Variables that are prescribed by the physical parametrizations are more sensitive to variability in the observing system than those variables that are directly specified. Furthermore, as was revealed by the deliberations of the Ozone Trends Panel (1988), the best trend determination is often determined by explicitly computing the behaviour of separate observational streams. Bengtsson et al. (2004) and Santer et al. (2004) perform trend analyses with reanalysis datasets; their work will be discussed more thoroughly below.

An important product from the first and second generation reanalyses is the quality-controlled input data record (Onogi 2000; Haimberger 2006). This examination of the input data record comes from comparing the input data stream with model estimates of expected values as well as with neighbouring observations. Information is provided on both global and local observing systems. For instance, it is possible to establish jumps in mean quantities as satellite instrumentation changes as well as to quantify changes in instrument performance. For the radiosonde network measurement differences between the instruments used by different countries and provided by different vendors are quantified. For other types of observations, for example shipboard observations, it is possible to identify systematic errors that establish that the observing sensor is not at the reported altitude above the sea's surface. The quality control information obtained from reanalysis projects is a potentially rich research product that is underutilized. As institutions push forward with new reanalyses, they are committed to sharing these quality-controlled input datasets. This will improve the robustness of future conclusions drawn from reanalysis datasets as one source of non-geophysical variability will be reduced.

There are other unique lessons learned from the reanalysis activities. One is that modern assimilation systems applied to the historical observations improve forecasts. A number of notable forecast failures in the pre-satellite era have been studied and forecast quality is greatly improved (see Kistler et al. 2001). This validates that research investments in model development and the evolution of assimilation methodology have beneficial impact. Another result of note is that methods of data treatment that have been applied in weather prediction might have to be reconsidered in climate applications. For instance, direct consideration of aerosol radiative effects on infrared observations might be important during periods of volcanic activity to assure the accurate use of radiances. Finally, the reanalyses help

to focus attention of those observations needed to address the key uncertainties in energy, moisture and constituent budgets, providing guidance for future observing systems.

*Heterogeneity of the input data stream.* The input observations used in reanalysis come from many sources. Historically, the bulk of the measurements to be assimilated are extracted from those collected, operationally, for weather forecasting. These measurements include observations of the surface conditions over land and ocean, observations from weather balloons and airplanes, and remotely sensed observations from satellites. The instruments used to make these observations were not designed with calibration standards to establish long-term, climate-quality datasets. Further, observing systems deployed by different countries and different agencies within countries were not (and are not) procured and deployed in a way to assure consistent accuracy.

Added to the observations collected for operations are those observations collected for research. A present and growing practice is to use research observations in operational applications (see chapter *Research Satellites*, Lahoz). Reanalysis projects are ideally suited to include research-data streams that were not appropriate for real-time applications when they were originally collected. Some of these research observations were collected in campaigns of limited temporal span and spatial extent. Others have been collected during multiyear satellite missions. Data archeology, pioneered by Roy Jenne at the National Center for Atmospheric Research in the United States, recovers some of these research data so they can be brought to bear on reanalysis problems. These recovered datasets are especially important for the quality of the reanalyses prior to 1960.

One of the most obvious discontinuities in the observing system is the beginning of the record in 1979 of operational, polar-orbiting satellites. Prior to this time, the upper air observing system was dominated by order $10^5$ radiosonde observations per day. The radiosonde observations were (and are) concentrated in the Northern Hemisphere. Besides differences in spatial and temporal coverage, the jump in 1979 is related to specific characteristics of the profile-by-profile observations. For example, the vertical resolution of the radiosondes is much higher than that of the satellite observations. One result of this is that near the tropical tropopause the poorer resolution of the satellite observations manifests itself as a positive temperature bias. There are numerous sources of bias between radiosonde and satellite temperature observations, and these vary with space and time.

A specific, subtle example of the impact of input data heterogeneity is from the radiosonde network itself (Lait 2002; Redder et al. 2004; Haimberger 2006). Radiosondes provide what some consider to be the single most important class of observations of the upper air. This might be arguable in the current era of high quality satellite observations and as satellite assimilation techniques improve, but there is no argument that the radiosonde network is of paramount importance prior to the satellite era (see chapter *The Global Observing System*, Thépaut and Andersson). The radiosonde measurements have benefited from much scrutiny, and strategies to develop climate quality datasets have been exercised. Different countries use different types of radiosondes, and within a country, several manufacturers of radiosondes are used. There is no consistent calibration of radiosondes.

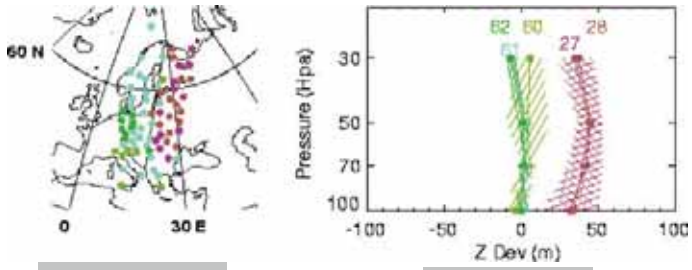



**Fig. 1** From Lait (2002). The *left panel* shows the distribution of radiosondes observations over eastern Europe, colour-coded by manufacturer. The *right panel* shows the difference of the radiosonde heights from the zonal mean analysis. The different types of radiosondes group together, and a spurious circulation separates the different types of radiosondes. See also Rood (2003)

Lait (2002) examines the impact of the heterogeneity of the radiosonde network on the quality of the assimilation analysis. Lait subtracts the zonal mean geopotential height from that of the radiosonde observation. This reveals persistent anomalies clustered by radiosonde type. A regional aspect of this impact is shown in Fig. 1. The left panel shows the radiosondes over eastern Europe, colour coded by manufacturer. The right panel shows the difference of the geopotential height from the zonal mean, still, colour coded by manufacturer. The eastward lying observations are between 30 and 40 geopotential metres higher than the westward lying observations. This height gradient is persistent with altitude. A wind error of order 5 $ms^{-1}$ is consistent with this height gradient in a part of the atmosphere where the expected wind speed is order 10 $ms^{-1}$. Lait (2002) identifies persistent wind patterns, seemingly spurious rivers of air, surrounding regions of differing radiosonde instrumentation. Again, this is directly related to biases in the observations of fundamental geophysical parameters (see chapter *Bias Estimation*, Ménard).

The discussion above brackets the extremes of the issues associated with data heterogeneity. At one extreme, when a new global observation type is added to the observing system, large changes in the assimilated data product are realized. In the case of the radiosondes, subtle biases between different types of radiosondes were shown to have large enough impact on the analysis of wind to impact the quantification of atmospheric transport. Between these two extremes are a whole set of impacts that might be expected when new data types are introduced. For example, the introduction of scatterometry data to define the ocean surface winds or precipitation observations to define the hydrological cycle will, no doubt, improve the quality of the assimilated data product. However, these improvements will be accompanied by changes in mean quantities such as surface pressure, precipitation, and outgoing longwave radiation; hence, leaving a signal in the reanalysis time series that is not of geophysical origin.

Alternatively, the exquisite sensitivity of the reanalysis to the input data stream suggests that the assimilation process is an outstanding monitor of the quality of the observing system. Štajner et al. (2004) provide one example of using assimilation to monitor the observing system by detecting variability as a function of satellite scan angle, changes in retrieval techniques, and orbital degradation.

*Impact of bias.* Data assimilation theory has been implemented, primarily, under the assumption that the information from the observations is unbiased relative to the information from the model (see chapters in Part I, *Theory*). That is, given a parameter such as temperature, the time mean of the observations subtracted from the mean of the model prediction is zero. However, as the previous discussion on heterogeneity in the observing system shows, the observations themselves are biased relative to each other even within the same nominal instrument type, e.g. the radiosondes and the succession of operational satellites. Different observing systems measuring the same geophysical parameter are expected to have bias between each other. There are systematic errors in the models. These systematic errors have regional and temporal dependencies. The assimilation quality is impacted by the bias between model prediction and observations as well as the bias between different pieces of the observing system.

One of the classic bias problems of data assimilation is known as the "spin-up" problem. Precipitation is determined to first order by the estimation of temperature and humidity and the use of these estimates by the physical parametrizations of the model. In the absence of assimilation the model determines precipitation based upon the model's temperature and humidity. Often when the observation-corrected temperature and humidity that comes from the assimilation are used, precipitation far in excess of that which is observed is estimated. This biased estimate of precipitation suggests that fundamental processes in the model are not well represented on the scale of the observations; i.e., there is substantial model error. In this case, since temperature is relatively smooth and estimated well by the model, the errors can be linked to the moisture field. It is often the case that the vertical structure of the moisture field is in error. Specifically, there is a discrepancy between amount of moisture modelled and observed in the planetary boundary layer, as contrasted with the upper troposphere. Over the course of the forecast, excess moisture rains out and the model "spins up" to a balance.

From the point of view of short-term prediction, directly assimilating information that corrects the physical parametrizations can have a large positive impact. Hou et al. (2001, 2002) have shown that assimilating satellite precipitation observations improves both forecast skill and the estimate of important metrics of the climate system, for example, outgoing longwave radiation. Still, however, the physical processes in the model are always tending towards their biased state, and the correction by the insertion of observations is not without consequence. The general circulation, the time averaged, spatially averaged dynamics of the atmosphere, is where the consequence is usually realized (see, for example, Chen et al. 2008a). This will be discussed more thoroughly in the section on constituent transport modelling.

Ultimately, the quality of assimilation analyses will be dependent on eliminating the bias between the model and the observations. Assuming that the observations

can be corrected in some way to eliminate the bias between different instrument types, the elimination of bias between the model and observations relies on improved model quality. Much of this improvement will come from better physical parametrizations and will require reformulation of physical parametrizations. Such development will be based on improved, more complete observations and modelling algorithms that can utilize the observed information. In the meantime, however, there is potential benefit derived from bias correction.

Figure 2 demonstrates a prescribed, idealized system and an estimate of that system by model-data assimilation. The smooth line shows the known mean state, i.e., climate. The segmented line shows a series of model forecasts corrected intermittently by a set of observations that, over time, are randomly distributed around the known mean state. In the top plot the model forecast is unbiased; in the bottom plot the forecast is biased. In both plots the observations are unbiased. At a given time, 1979, the observing system is changed so that more observations are taken. This is symbolic of the increase in temporal and spatial resolution that occurred when satellite observations became operational. In the top plot when the model predictions are unbiased, the mean error in the analysis remains essentially the same before and after the change in the observing system. In the bottom plot, where the model prediction is biased, the increase in density of the observations reduced the mean error in the analysis by half, leaving a jump in the estimate of the mean state. Therefore, even if the mean state of the observations is homogenized prior to assimilation through some calibration procedure, as long as there is model error, reanalyses will be subject to errors based simply on improved data coverage.

Dee (2005) investigates the role of bias in assimilation. Dee posits that all components in the assimilation system are a potential source of bias and can propagate and

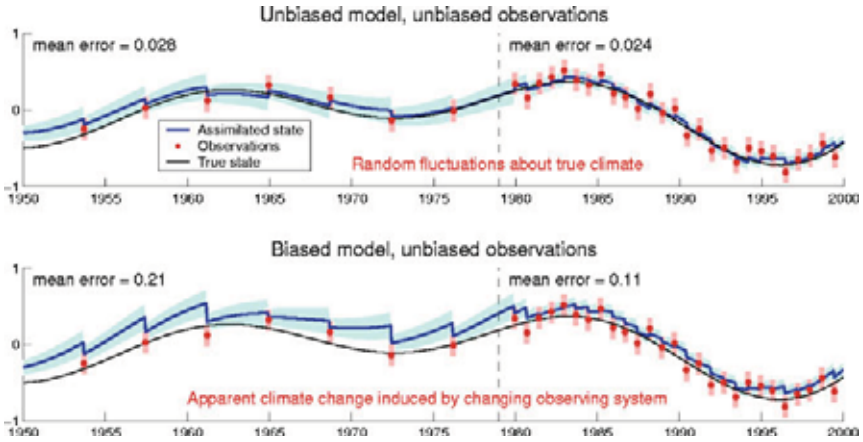



**Fig. 2** This figure is adapted from Dee (2005). The *solid line* represents a known true state of an idealized climate system. The *red dots* are observations of the system. The *blue lines* are model forecasts of the mean state following assimilation of the observations into the model. In the *top frame* the model is not biased. In the *bottom frame* the model is biased. Figure courtesy of D.P. Dee; see also Rood (2003)

enhance bias. Techniques to account for the bias require the use of ancillary information that may come from independent observations of known quality or theoretical evaluation of the source of the bias. In some cases it is not difficult or expensive to estimate bias and apply a correction algorithm (Dee and da Silva 1998). This can improve the quantitative integrity of the assimilated dataset and have positive impact, especially on prediction of parameters that are being assimilated. However, the bias correction is ultimately compensating for shortcomings in the system. This often implies that the model physics (or chemistry) are not correct, and this will ultimately manifest itself somewhere in the assimilated dataset.

Dee (2005) investigates the development of bias-aware assimilation techniques. With consideration of the possible sources of bias, it is possible to develop adaptive techniques to compensate for the bias. This is a formidable and, sometimes, imprecise task as bias is known to have multiple sources with spatial and temporal variability. As pointed out by Dee such techniques will "by construction, reduce the mean analysis increments, but not necessarily for the right reasons." The role of bias in data assimilation remains a fundamental problem (see chapter *Bias Estimation*, Ménard), and it is of particular importance to the development of reanalyses for climate and constituent transport applications.

The temporal averaging or smoothing that is intrinsic in the 4D-Var (four-dimensional variational) assimilation technique (see chapter *Variational Assimilation*, Talagrand) can reduce the effects of certain types of bias; however, there is nothing intrinsic in 4D-Var that eliminates the effect of bias through first principles. The type of bias that is most impacted is that where the model forecast is accurate and the statistics of the observations are such that a temporal average over the time interval of the forecast-assimilation cycle are unbiased relative to the model. Persistent biases that are related to the inadequacies of model representation of variables or instrumental characteristics will continue to impact negatively the assimilated data product in 4D-Var systems.

*Impact of data heterogeneity and bias on trend determination.* Simmons et al. (2004) investigate surface temperature trends in reanalyses and surface station observations and find complex relationships between the observation system and spatial and temporal scales. Bengtsson et al. (2004) assess the ability to determine trends with the ERA-40 reanalysis for several geophysical parameters. They investigate directly constrained variables (temperature), weakly constrained variables (integrated water vapour), and derived parameters (kinetic energy). If trends are calculated without regard to the observing system, then large spurious trends are found in all of the parameters. If the datasets are split into segments where the observing system is quasi-homogeneous then more convincing trends are determined. With special scrutiny, it is possible to provide corrections that improve the trend determination. Still, this study concludes "that there is significant uncertainty in the calculations of trends from present reanalyses data." (Bengtsson et al. 2004).

Bengtsson et al. (2004) studied, primarily, global trends. The global average has the potential for errors to compensate in the averaging process. Bromwich et al. (2007) compare ERA-40, NCEP/NCAR, and JRA-25, with a focus on representation of high latitudes; they also provide a good introductory summary of the

attributes of the different products. They note that the reanalyses are more accurate in the Arctic than the Antarctic, introducing the idea that there is regional heterogeneity in the quality. Further, they show that the summertime is more accurate than the wintertime, especially before the availability of satellite data. Hence, there is temporal heterogeneity in reanalysis products. There are significant differences between the reanalysis products. In the case of the Polar Regions, there are significant differences in atmospheric circulation and the propagation of weather-scale waves. Bromwich et al. (2007), also, point out significant sensitivity to the details of the satellite observing system revealed in the preparation for NASA's Modern Era Retrospective-Analysis for Research and Applications (MERRA); it is not simply a matter of satellite/no satellite.

In the case of the MERRA reanalysis, the Special Sensor Microwave/Imager (SSM/I) is a significant change in the satellite data observing system, being a new instrument yielding profiles of moisture and temperature. Onogi et al. (2007) show a change in precipitation (an improvement) with the availability of SSM/I. Bosilovich et al. (2008) tested the impact of SSM/I in July/August 1987, when it is initially available. Figure 3 shows that there is a change in the character of precipitation in the MERRA system. This leads to a 10% increase in tropical (15°S–15°N) precipitation when SSM/I radiances are assimilated.

Santer et al. (2004) use both first and second generation reanalysis products to investigate possible trends in tropopause height and the attribution of that trend to greenhouse gas global warming. This study provides a summary of the strengths and weaknesses of reanalysis products, and emphasizes the importance of the coherent dynamical structure provided by the reanalyses in determining trends. This coherency helps to define correlative behaviour between geophysical parameters and contributes to the definition of "fingerprints", which can be used to distinguish

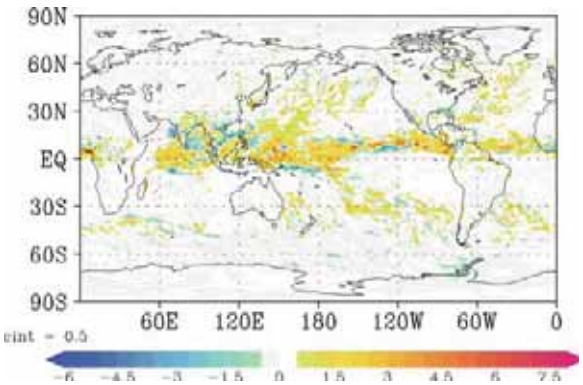



**Fig. 3** Data impact test of the inclusion of Special Sensor Microwave/Imager (SSM/I) in the GEOS-5 data assimilation system to be used for MERRA. August 1987 monthly mean precipitation difference between two experiments, with and without SSM/I radiance assimilation is shown. Units are mm day$^{-1}$. *Red* indicates positive differences (experiment with SSM/I radiance assimilation has higher values); *blue* indicates negative differences (experiment with SSM/I radiance assimilation has lower values)

cause and effect mechanisms for observed trends in warming. Santer et al. (2004) compare the temperature provided by the reanalyses with standard observational datasets that are used in trend detection. Using this comparison to verify the performance of the reanalysis, they derive the behaviour of the tropopause height. The data assimilation provides the estimate of the tropopause height, which is correlated with the temperature observations used in the verification process. This use of external observations and careful examination of correlated physics serves as an example of a strategy for applying reanalysis datasets to trend studies.

Chen et al. (2008a) demonstrate the complexities of using reanalysis products in the determination and attributions of trends. They explicitly discuss the impact of data discontinuities on the quality of the reanalysis. Using the idea that the reanalyses provide a dynamically coherent estimate of spatial and temporal variability, Chen et al. develop a technique to remove El Niño – La Niña variability from the longer-term time series. With this method they estimate the part of the temperature change due to global warming, including regional estimates. A fascinating result from the Chen et al. (2008a) study is that the NCEP/NCAR reanalysis shows an atmospheric response in the Walker Circulation, and the ECMWF ERA-40 shows the atmospheric response in the Hadley Circulation. These features of the general circulation, which are related to the divergence of the wind and the dissipation of waves, are the most difficult for assimilated datasets to represent.

## 3 Lessons from Applications

The following two applications, hydrometeorology and constituent transport modelling, will be used to demonstrate the scientific challenges that remain for reanalysis. Both of these problems are characterized by the fact that effective quantitative analysis requires the conservation of key physical variables: mass, momentum, and energy. The challenges that are faced and the deficiencies that are revealed demonstrate that in reanalysis datasets the insertion of the observations is a significant source or sink term in the conservation equation. In both applications, the conservation budgets with a non-assimilating model are more consistent, physically, than in the case of assimilated data. This fact points directly at the role of bias. To be clear, assimilated datasets are *not* consistent from a physical point of view as long as biases are being corrected by the insertion of observed information. The correction of bias through assimilation propagates and enhances biases throughout the system (see Dee 2005). The geophysical quantities from an assimilated dataset are constrained or informed by observations, perhaps they are a better match to observations than the unconstrained quantities, but the fabric that connects the variables, the correlated physics, is not the same as in the atmosphere. How well or how poorly correlated behaviour is represented is a function of both spatial and temporal scales. In particular, slow processes in the atmosphere – those features that are associated with residual circulations like the Hadley cell, the Brewer-Dobson circulation, and the Walker circulation (see chapter *General Concepts in Meteorology and Dynamics*, Charlton-Perez et al.), are not likely to be well represented.

*Hydrometeorology.* One of the key utilities in a reanalysis is that the output generated from the model physics provides information about variables that are not easily observed, but are informed by the analysed observed information. Uncertainties are a complex mix associated with observations, models, and implementation of analysis techniques. Betts et al. (2006) and Bengtsson et al. (2007) summarize strengths, weaknesses and the utility of reanalyses, especially regarding hydroclimate studies. Trenberth and Smith (2008, 2009) and Trenberth et al. (2008a) investigate, thoroughly, the energy budget in reanalyses. Betts (2004) provides a framework for using the correlated physics of hydrometeorological observations to analyse the underlying quality of global modelling and assimilation systems. This framework connects surface processes, radiative transfer, clouds, water, precipitation, and evaporation. While the method shows promise both in understanding the model and assimilation systems as well as the Earth's processes, challenges remain in verifying the connective processes. Betts and Bosilovich (2008) investigated the hydrometeorological connections in preliminary MERRA data compared to ERA-40. Figure 4 shows that coupling in the Amazon is quite different between the two systems. MERRA exhibits a wide dynamic range of evaporative fraction with little sensitivity to cloud fraction, while ERA-40 evaporative fraction increases steeply with cloud fraction. They caution that the coupling is also regionally dependent, and the differences between the systems indicate that users should take time to evaluate the processes in their region of interest.

Precipitation is an important validation metric for the climatology of reanalyses, being coupled into the energy and water cycles, as well as the dynamic circulation. Kalnay et al. (1996) classified precipitation as subject to large uncertainty. From a hydrometeorological perspective, observations are assimilated into reanalysis systems and the model parametrizations each affect the resulting estimate or forecast of precipitation. Newman et al. (2000) showed that there is internal consistency of precipitation, outgoing longwave radiation and upper level divergence within three different reanalyses, but the consistency between the reanalyses was very low.

Chen et al. (2008a, b) isolated the long-term trends in the NCEP/NCAR and ERA-40 reanalyses, evaluating the changes in both dynamics and thermodynamics. Figure 5 shows the long-term trend of the Hadley (top) and Walker (bottom) circulations. The Hadley circulation in ERA-40 has changed significantly in time, and this may be related to a spurious trend in latent heating by precipitation and the variations of the observing system. On the other hand, the NCEP/NCAR reanalysis shows change in the Walker circulation, correlated to changes of sea surface temperatures, which are prescribed by observations. The representation of these tropical circulations requires accurate representation of both dynamics and heating. The relationship between vertical motion and latent heating directly connects the divergence of the horizontal wind and the physical parametrizations. Both of these quantities are difficult to calculate. Precipitation is an integrated measure of this balance.

Bosilovich et al. (2009) used eight operational assimilation systems to investigate the uncertainties of the precipitation and outgoing longwave radiation. An ensemble average and variance were produced. Figure 6 shows the comparison of each of the analyses and ensemble average precipitation with precipitation

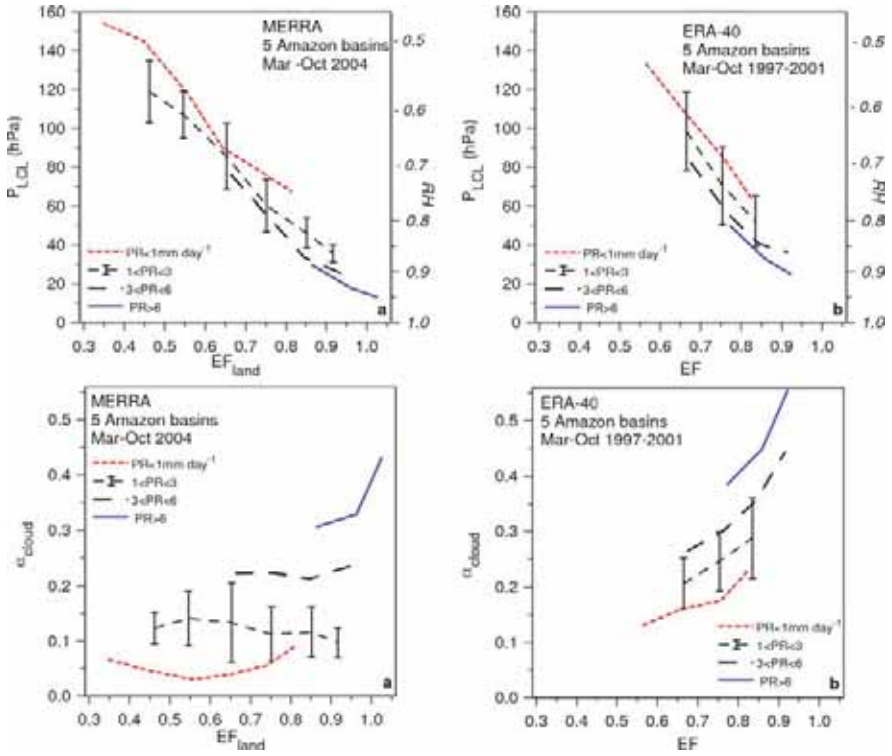



**Fig. 4** Functional relationships between lifted condensation level (LCL; *top plots*) and cloud albedo (closely related to *top* of atmosphere, TOA, albedo; *bottom plots*) with evaporative fraction (EF) and precipitation (PR) for MERRA (**a**; *left-hand plots*) and ERA-40 (**b**; *right-hand plots*). Note that the MERRA data is a short preliminary experiment, compared to a longer time series for ERA-40

from the Global Precipitation Climatology Project (GPCP, Adler et al. 2003, http://precip.gsfc.nasa.gov/). The global ensemble of analyses has lower error than any of the contributing members. Since, essentially the same observations are used in all of the analyses, the correlated features related to the observations should remain (both positive and negative). Uncorrelated model errors in the analyses can be minimized through the ensemble average. This suggests that an ensemble of reanalyses may provide some benefit. However, since this is a statistical formulation, it remains to be seen the degree to which such an ensemble may adhere to the physical principles that govern the Earth's processes.

Terrestrial drainage is a primary source of fresh water for the Arctic Sea, and so is an important component of the climate system (see chapter *Land Surface Data Assimilation*, Houser et al.). Several studies have applied reanalysis precipitation as forcing for river discharge models. Serreze and Hurst (2000) found reasonable spatial patterns at large scales and high northern latitudes in reanalyses. There were some notable seasonal biases (better in winter, worse in summer). Precipitation

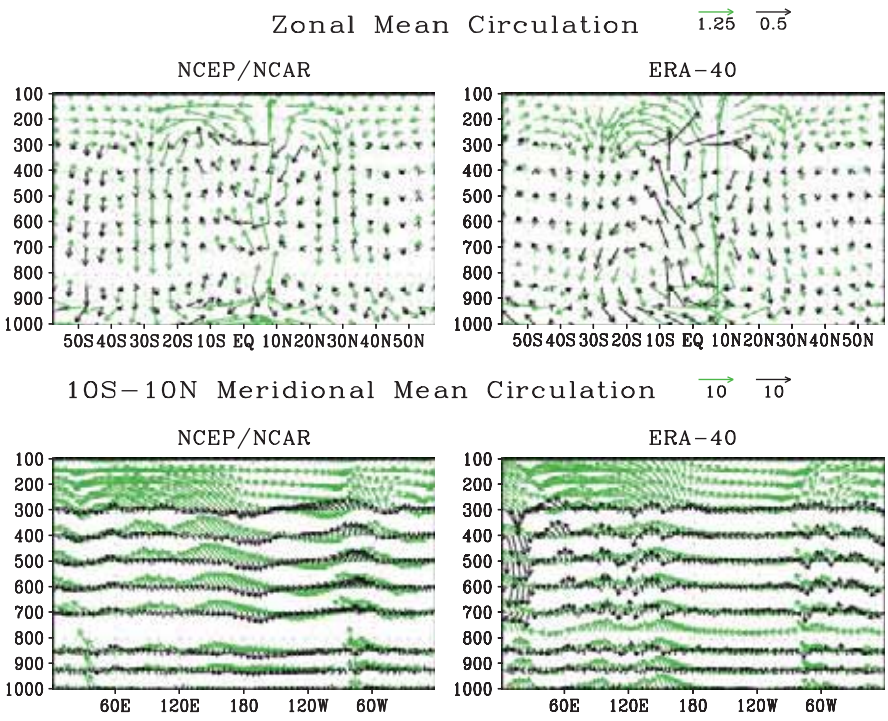



**Fig. 5** The circulation changes (*black vectors*) associated with the global warming trend mode in the zonal mean meridional-vertical cross section (*upper row*) and the 10°S–10°N meridional mean zonal-vertical cross section (*lower row*). *Left column*: NCEP/NCAR reanalysis data. *Right column*: ERA-40 reanalysis data. The climatology is drawn in *green vectors*. In the *upper plots*, showing the Hadley circulation, the *horizontal* component of the vectors is meridional wind with unit 1 $ms^{-1}$, and the *vertical* component of the vectors is negative ω with unit –1/60 hPa $s^{-1}$. In the *lower plots*, showing the Walker circulation, the *horizontal* component of the vectors is zonal wind with unit 1 $ms^{-1}$, and the *vertical* component of the vectors is negative ω with unit –1/120 hPa $s^{-1}$. The *arrow* lengths of the vectors are scaled as shown on the *top* of each row. From Chen et al. (2008a)

bias was related to high incoming shortwave radiation, which provided energy for evaporation and then precipitation. Pavelsky and Smith (2006) used two reanalyses and two observed precipitation data products, showing that a few positive aspects in the reanalyses were offset by substantial errors in variability and trends of the data. At high latitudes the quality and completeness of the direct observations have significant problems; for example, blowing snow leads to an underestimate of precipitation. Serreze et al. (2003) conclude that, while needing improvements, reanalyses are useful to study the high latitude water cycle. Cullather et al. (1998) find that reanalyses generally agree on the main features of Antarctic precipitation, but focusing on any region may lead to discrepancies. Teleconnections between ENSO (El Niño-Southern Oscillation) and Antarctic precipitation are influenced by how effectively observations input to the reanalysis are used (Bromwich et al. 2000).

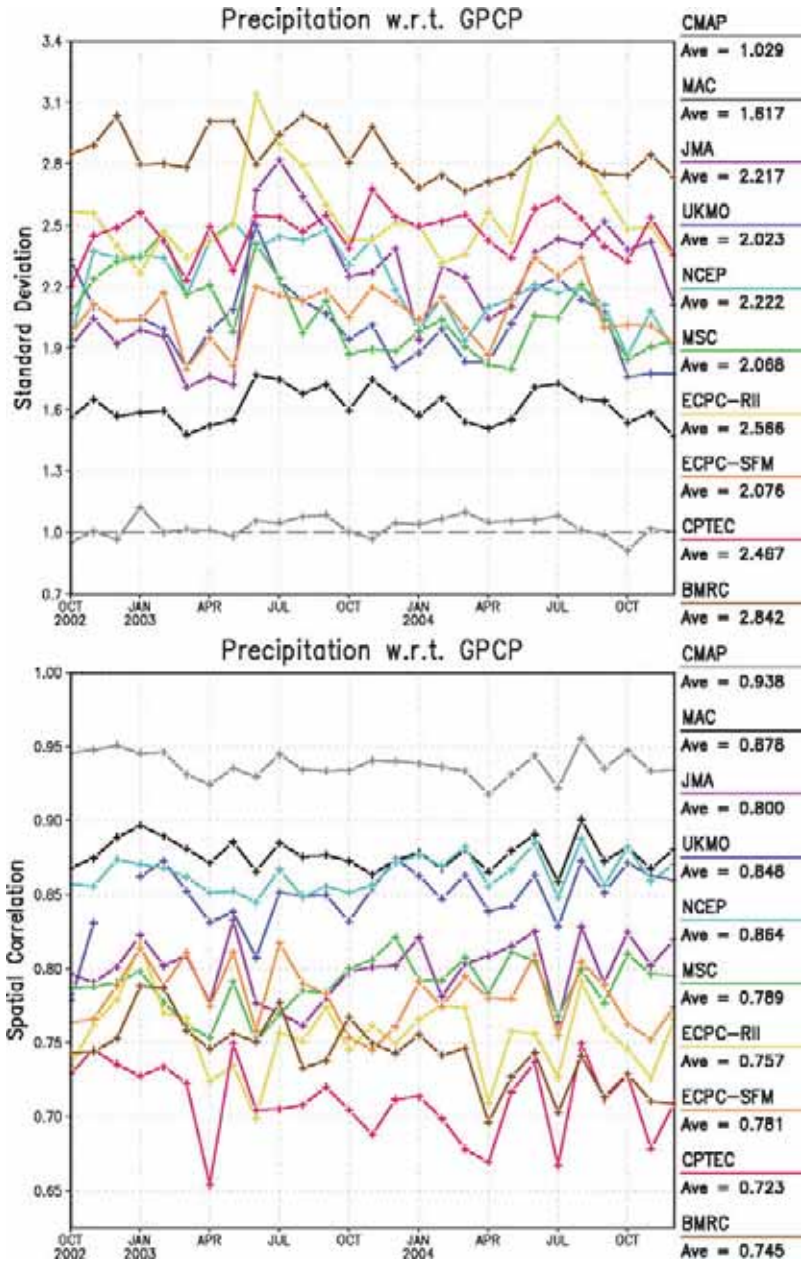



**Fig. 6** Standard deviation of the monthly global differences of eight operational analyses (identified by colour – see *right hand* of plots) and their ensemble average (labeled MAC) from the Global Precipitation Climatology Project (GPCP, Adler et al. 2003 http://precip.gsfc.nasa.gov/) (*top plot*) and spatial correlation to GPCP (*bottom plot*). The Climate Prediction Center Merged Analysis of Precipitation (CMAP, http://www.cpc.noaa.gov/products/global_precip/html/wpage.cmap.html) global precipitation observations are provided as a measure of observational uncertainty

Basin scale studies in well-instrumented regions allow comprehensive budget studies and the potential for independent observations to validate reanalysis systems. Hagemann and Gates (2001) used large basin scale discharge to compare reanalyses and identify weaknesses in the physics parametrizations. Fekete et al. (2004) also computed runoff from observed and reanalysis precipitation, and found the largest errors and sensitivity in arid and semi-arid regions. Basin scale studies allow for the evaluation of the coupling of the water and energy cycles in reanalyses (Roads and Betts 2000), but also the assessment of the impact of observations through the data assimilation and the spin-up in the subsequent forecast (Viterbo and Betts 1999). Schubert and Chang (1996) used multiple linear regression and the time series of analysis increments of atmospheric water and the atmospheric water budget to attribute the analysis increment contributions back to corrections of precipitation and evaporation. This method was later applied to monthly mean reanalysis water budgets with favourable comparisons to observations (Bosilovich and Schubert 2001).

Major issues remain in trying to improve the hydroclimate of reanalyses; these require continued research if they are to be addressed. One strategy takes advantage of the physical consistency realized in the stand-alone climate models. Then limited observing systems are used to constrain a particular attribute of the model. The model then evolves with this limited constraint. An example of this strategy is to use an ensemble of reanalyses using only surface pressure to provide 100 years of reanalyses data and include uncertainty estimates (Compo et al. 2006).

*Constituent transport modelling.* Rood et al. (1989) first used winds and temperatures from a meteorological assimilation to study stratospheric transport. Since that time there have been productive studies of both tropospheric and stratospheric transport. However, a number of barriers have been met in recent years, and the question arises – has a wall been reached where foundational elements of data assimilation are limiting the ability to do quantitative transport applications? Stohl et al. (2004; and the references therein) provide an overview of some of the limits that need to be considered in transport applications. Chapters *Constituent Assimilation* (Lahoz and Errera) and *Inverse Modelling and Combined State-source Estimation for Chemical Weather* (Elbern et al.) discuss the assimilation of constituents.

In transport applications, winds and temperatures are taken from a meteorological assimilation and used as input to a chemistry-transport model. The resultant distributions of trace constituents are then compared with observations. The constituent observations are telling indicators of atmospheric motions on all time-scales. Ultimately, how constituents are distributed in the atmosphere is related to the general circulation of the atmosphere. This is linked to the divergent component of the wind and/or vertical motion. The general circulation is determined by the dissipation of dynamical features. Data assimilation for weather prediction focuses on the propagation of dynamical features, and the dissipation of these features occurs on time-scales that are long compared with the forecast time-scale. In fact, dissipation often occurs outside of the model domain (e.g. the stratosphere), and dissipation is highly sensitive to the insertion of observations. There are fundamental, conflicting requirements of data assimilation for weather prediction and for climate diagnostics.

Constituent observations are often of very high quality and come from many observational platforms. They are markers of motion. As a community, rigorous quantitative Earth science has been significantly advanced by comparison of constituent observations and model estimates. Overall, it is found that the meteorological analyses do a very good job of representing variability associated with synoptic and planetary waves. This has been invaluable for accounting for dynamical variability, and allowing the evaluation of constituents from multiple observational platforms. On the other hand, those geophysical parameters that rely on the representation of the general circulation, for instance the lifetimes of long-lived constituents are poorly represented.

Douglass et al. (2003) and Schoeberl et al. (2003) each provide detailed studies that expose some of the foundational shortcomings of the physical consistency of data assimilation. In their studies they investigate the transport and mixing of atmospheric constituents in the upper troposphere and the lower stratosphere. Figure 7 from Douglass et al. (2003) shows ozone probability distribution functions in two latitude bands from four experiments using a constituent transport model. In three of these experiments, Panels B, D, and E, winds and temperatures are taken from an assimilation system. In Panel C are results from an experiment using winds from a general circulation model (GCM) simulation; that is, a free-running model without assimilation. Panel A shows ozonesonde observations; the sondes reflect similar distributions in the two latitude bands. In all of the numerical experiments, the means in the two latitude bands are displaced from each other, unlike the observations. In the three experiments using winds from different data assimilation systems (DAS), the half-width of the distributions is much too wide.

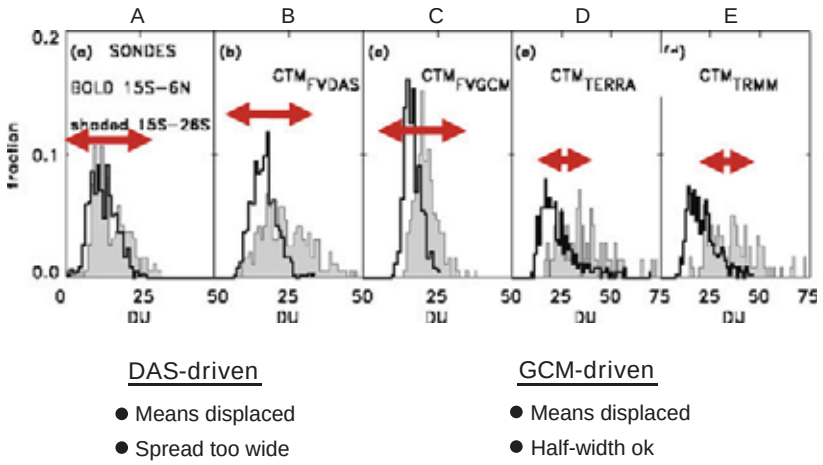



**Fig. 7** From Douglass et al. (2003). Probability distribution functions of ozone from: *Panel* (**A**) ozonesondes; and *Panels* (**B**–**E**) constituent-transport models, CTM, experiments. DAS-driven is from experiments using winds from a data assimilation system. GCM-driven is from experiments using winds from a general circulation model without data assimilation

There are a number of points to be made in this figure. First, the winds from the assimilation system in Panel B and the model in Panel C both use the finite-volume dynamics of Lin (2004). Therefore, these experiments are side-by-side comparisons that show the impact of inserting data into the model. Aside from developing a bias, the assimilation system shows much more mixing. As Douglass et al. show, the instantaneous representation of the wind is better in the assimilation, but the transport is worse. This is attributed to the fact that there are consistent biases in the model prediction of the tropical winds and the correction added by the data insertion causes spurious mixing. Tan et al. (2004) also investigate the dynamical mechanisms of the mixing in the tropics and the subtropics and find systematic errors consistent with these results. Second, the assimilation systems used for Panels D and E, have a different assimilation model, and their representation of transport is worse than that from the finite-volume model. This improvement is attributed to the fact that the finite-volume model represents the physics of the atmosphere better, in particular, the representation of the vertical velocity. Third, the results in Panel B show significant improvement compared to the older assimilation systems used in Panels D and E. Older assimilation systems have had enough deficiencies that scientists have shied away from doing tropical transport studies. Thus, this example demonstrates both the improvements that have been gained in recent years and indicate that the use of winds from assimilation in transport studies might have fundamental limitations.

Figure 8 is from Schoeberl et al. (2003). The Schoeberl et al. study is similar in spirit to the Douglass et al. study, but uses Lagrangian trajectories instead

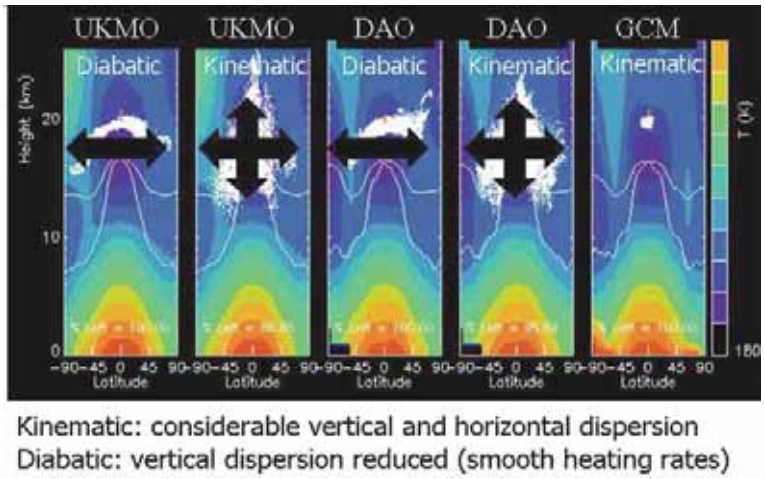



**Fig. 8** From Schoeberl et al. (2003). The dispersion of a tracer released at the tropopause from five numerical experiments. UKMO is from the United Kingdom Meteorological Office (now, The Met Office) assimilated data. DAO is from Data Assimilation Office (now, Global Modeling and Assimilation Office) assimilated data. GCM is from the general circulation model used in the DAO assimilation. Kinematic refers to vertical velocity calculated from the divergence of the horizontal wind. Diabatic refers to vertical velocity calculated from the thermodynamic equation

of Eulerian advection schemes. This allows Schoeberl et al. to address, directly, whether or not the spurious mixing revealed in the Douglass et al. paper is related to the advection scheme. In this figure the results from two completely independent assimilation systems are used. The two assimilation systems are labelled UKMO (United Kingdom Meteorological Office, now, the Met Office) and DAO (Data Assimilation Office, now, the Global Modeling and Assimilation Office). The DAO system uses the finite-volume dynamical core and the panel labeled GCM (general circulation model) uses the finite-volume GCM. Vertical winds are calculated two ways. They are calculated diabatically using the heating rate information from the assimilation system and they are calculated kinematically, through continuity, using the horizontal winds from the assimilation.

The figure shows, first, the impact of the method of calculating the vertical wind using the diabatic information. When the diabatic information is used there is much less transport in the vertical. While this is, indeed, generally in better agreement with observations and theory, the diabatic winds no longer satisfy mass continuity with the horizontal winds. This result points to a self-limiting aspect of using diabatic winds in Eulerian calculations such as the ones of Douglass et al. (2003). Second, the Schoeberl et al. calculations show that even with the diabatic vertical winds, there is, still, significant horizontal mixing, which is compressed along isentropic surfaces. Third, the final panel shows that for the simulation, the free-running model, there is much less dispersion, which is in better agreement with both observations and theory. Schoeberl et al. attribute the excess dispersion in the assimilation systems to noise that is introduced by the data insertion.

These two studies point to the fact that the data insertion impacts the physics that maintains the balances in the conservation equations of momentum, heat, and mass. Both bias and the generation of noise have an impact. Both problems are difficult to address, with the problem of bias having fundamental issues of tractability. Again, while the data assimilation system does indeed provide better estimates of the primary variables, as the impact of data insertion is adjusted through the physics represented in the model, the derived parameters are often degraded. While there may be greater discrepancies in the absolute, day-to-day representation of constituents with free-running models, the consistent representation of the underlying physics allows more robust study of transport mechanisms and those features in the constituent data which are directly related to dynamics. The ultimate success of data assimilation for climate applications will be to preserve the physical consistency of the underlying model simulation in the presence of the insertion of observational information.

Note that recently, Pawson et al. (2007) have shown using NASA's Goddard Earth Observing System version 4 (GEOS-4) that the use of 6-h averaged wind fields instead of instantaneous analyses can substantially reduce problems in stratospheric transport associated with excessive mixing and an overstrong residual circulation. Also the ERA-Interim reanalysis significantly improves the dispersion of stratospheric tracers and calculations of the age of air (D. Dee, personal communication).

## 4 Summary

There is no doubt that reanalysis datasets play a central role in the modern practice of the scientific investigation of climate. Reanalyses are also used as lateral boundary conditions for regional climate models and dynamical downscaling experiments. There are thousands of references to the publications that describe the reanalysis datasets. In fact, the prominent use of "observations" that are, actually, a melding of model and observational information is a subject of interest to historians (P.N. Edwards, personal communication).

One reason the reanalysis datasets are widely used is that they provide an ordered and complete representation of the atmosphere that is nearly continuous in time. Reanalyses compile more observations from disparate spatial and temporal scales than individual researchers could accomplish. Furthermore, the data assimilation provides additional quality checking of those observations. Assimilation based analyses interpolate and extrapolate observational information using the physical principles of fluid dynamics to transport information. The success of the reanalysis datasets to represent atmosphere winds and temperatures in middle latitudes is remarkable. With this information it is possible to estimate dynamical variability and to bring observations scattered in space and time to a common framework. The successes are greatest for middle latitude problems and for problems with the intrinsic time-scales of weather forecasting – days.

For problems of longer and shorter time-scales, for problems in the tropics and the poles, for problems that rely upon the subscale physical parametrizations in the model, a set of deficiencies is revealed in the reanalysis datasets. Many of these deficiencies are related to bias in the assimilation system. There are tractable strategies for addressing some sources of bias. For other forms of bias, it is not clear that they can be fully eliminated. For this reason it is required that scientists maintain a critical scrutiny of reanalysis datasets in applications that require the calculation of mass, momentum, and energy budgets or the identification of temporal trends. Of special note, the ability of reanalysis datasets to provide robust geophysical information will vary by region and season. The propagation of biased information through the reanalysis system means that reanalysis datasets are not geophysically consistent.

The reanalysis datasets reflect with exquisite sensitivity the heterogeneity of the observation network. The act of performing a reanalysis does not eliminate the granularity of the observing system or relegate the granularity to being small enough to ignore. In fact, the sensitivity to granularity in the observing system is another factor motivating the development of a calibrated climate observing system (see, for example, Trenberth et al. 2002). Of course, we do not have the luxury of building a climate observing system for the past, and climate science requires long time series of observations. Reanalysis systems have the ability to extend information from modern observing systems to the past; they can contribute to the calibration of observing systems. This requires scientific investigation to optimize the use of subsets of the observations; this is a research path that is only beginning to be followed.

Following the summary of Bromwich et al. (2007), Bengtsson et al. (2004, 2007), Santer et al. (2004), Trenberth et al. (2008b) and many others, reanalyses are a powerful tool for climate studies, which must be used with a critical eye that

recognizes their limitations. The newest reanalyses, MERRA and ERA-Interim (see links at the end of the chapter), are just becoming available. They were designed to address many of the problems addressed here, and early indications are that they are an important step forward. However, it is not likely that these will eliminate all of the uncertainties in the systems. The development of reanalysis systems and techniques to address climate issues are an ongoing process, as models and data quality improve.

## 5 Web Resources

3rd WCRP International Conference on Reanalyses, Jan 26-Feb 1, 2008: http://jra.kishou.go.jp/3rac_en.html

United States Climate Change Science Program Synthesis and Assessment Product 1.3: Re-analyses of historical climate data for key atmospheric features. Implications for attribution of causes of observed change: http://www.climatescience.gov/Library/sap/sap1-3/default.php

NCEP/NCAR Reanalysis: http://www.cdc.noaa.gov/cdc/reanalysis/reanalysis.shtml

NCEP/DOE Reanalysis 2: http://www.cdc.noaa.gov/cdc/data.ncep.reanalysis2.html

ERA-40: European Centre for Medium-Range Weather Forecasts: http://www.ecmwf.int/

ECMWF Interim Reanalysis: European Centre for Medium-Range Weather Forecasts: http://www.ecmwf.int/products/data/archive/descriptions/ei/index.html

http://www.ecmwf.int/publications/newsletters/pdf/115.pdf

http://www.ecmwf.int/publications/newsletters/pdf/111.pdf

JRA-25: Japan Meteorological Agency (JMA): http://www.jreap.org/

NARR: NOAA North American Regional Reanalysis: http://wwwt.emc.ncep.noaa.gov/mmb/rreanl/index.html

MERRA: NASA Modern Era Retrospective-Analysis for Research and Applications http://gmao.gsfc.nasa.gov/merra/

ASR: Arctic System Reanalysis: http://polarmet.mps.ohio-state.edu/PolarMet/ASR.html

Ocean Reanalyses: http://www.clivar.org/data/synthesis/directory.php

**Acknowledgments** We thank L. Bengtsson, D. Dee, and K. Trenberth for reviewing the chapter and providing many useful comments.

## References

Adler, R.F., G.J. Huffman, A. Chang, et al., 2003. The version-2 global precipitation climatology project (GPCP) monthly precipitation analysis (1979-present). *J. Hydrometeor.*, **4**, 1147–1167.

Bengtsson, L., P. Arkin, P. Berrisford, et al., 2007. The need for dynamical climate assimilation. *Bull. Amer. Meteorol. Soc.*, **88**, 495–501.

Bengtsson, L., S. Hagemann and K.I. Hodges, 2004. Can climate trends be calculated from reanalysis data? *J. Geophys. Res.*, **109**, D11111, doi:10.1029/2004JD004536.

Bengtsson, L. and J. Shukla, 1988. Integration of space and in situ observations to study global climate change, *Bull. Amer. Meteorol. Soc*., **69**, 1130–1143.

Betts, A.K., 2004. Understanding hydrometeorology using global models. *Bull. Amer. Meteorol. Soc*., **85**, 1673–1688.

Betts, A.K. and M.G. Bosilovich, 2008. Comparison of MERRA with ERA-40 on river basin scales. *Session on Advances in Atmospheric Reanalyses*, American Meteorological Society Annual Meeting, New Orleans, LA, January 23, 2008.

Betts, A.K., M. Zhao, P.A. Dirmeyer and A.C.M. Beljaars, 2006. Comparison of ERA40 and NCEP/DOE near-surface datasets with other ISLSCP-II datasets. *J. Geophys. Res.*, **111**, D22S04, doi:10.1029/2006JD007174.

Bey, I., D.J. Jacob, R.M. Yantosca, et al., 2001. Global modeling of tropospheric chemistry with assimilated meteorology: Model description and evaluation. *J. Geophys. Res.*, **106**, 23073–23095.

Bosilovich, M.G., J. Chen, F.R. Robertson and R.F. Adler, 2008. Evaluation of global precipitation in reanalyses. *J. Appl. Meteor. and Climat*., **47**, 2279–2299.

Bosilovich, M.G., D. Mocko, J.O. Roads and A. Ruane, 2009. A multi-model analysis for the Coordinated Enhanced Observing Period (CEOP). *J. Hydrometeorol*., **10**, 912–934.

Bosilovich, M.G. and S.D. Schubert, 2001. Precipitation recycling over the central United States as diagnosed from the GEOS1 Data Assimilation System. *J. Hydrometeorol.*, **2**, 26–35.

Bromwich, D.H., A.N. Rogers, P. Kållberg, et al., 2000. ECMWF analyses and reanalyses depiction of ENSO signal in Antarctic precipitation. *J. Climate*, **13**, 1406–1420.

Bromwich, D.H., R.L. Fogt, K.I. Hodges and J.E. Walsh, 2007. A tropospheric assessment of the ERA-40, NCEP, and JRA-25 global reanalyses in the polar regions. *J. Geophys. Res*., **122**, D10111, doi: 10.1029/2006JD007859.

Chen J., A.D. Del Genio, B.E. Carlson and M.G. Bosilovich, 2008a. The spatiotemporal structure of 20th century climate variations in observations and reanalyses. Part I: Long-term trend. *J. Climate*, **21**, 2611–2633.

Chen J., A.D. Del Genio, B.E. Carlson and M.G. Bosilovich, 2008b. The spatiotemporal structure of 20th century climate variations in observations and reanalyses. Part II: Pacific pan-decadal variability. *J. Climate*, **21**, 2634–2650.

Compo, G.P., J.S. Whitaker and P.D. Sardeshmukh, 2006. The feasibility of a 100-year reanalysis using only surface pressure data. *Bull. Amer. Meteorol. Soc*, **87**, 175–190.

Cullather R.I., D.H. Bromwich and M.L. Van Woert, 1998. Spatial and temporal variability of Antarctic precipitation from atmospheric methods. *J. Climate*, **11**, 334–367.

Dee, D.P., 2005. Bias and data assimilation. *Q. J. R. Meteorol. Soc*., **131**, 3323–3342.

Dee, D.P. and A. da Silva, 1998. Data assimilation in the presence of forecast bias. *Q. J. R. Meteorol. Soc*., **124**, 269–295.

Douglass, A.R., M.R. Schoeberl, R.B. Rood and S. Pawson, 2003. Evaluation of transport in the Lower Tropical Stratosphere in a global chemistry and transport model. *J. Geophys. Res.*, **108**, Art. No. 4259.

Fekete, B.M., C.J. Vorosmarty, J.O. Roads and C.J. Willmott, 2004. Uncertainties in Precipitation and their impacts on runoff estimates. *J. Climate*, **17**, 294–302.

Gibson, J.K., P. Kållberg, S. Uppala, et al., 1997. *ERA Description, ECMWF Re-analysis Final Report Series,* 1.

Hagemann, S. and L.D. Gates, 2001. Validation of the hydrological cycle of ECMWF and NCEP reanalyses using the MPI hydrological discharge model. *J. Geophys. Res.*, **106**, 1503–1510.

Haimberger, L., 2006. Homogenization of radiosonde temperature time series using innovation statistics. *J. Climate*, **20**, 1377–1403.

Hou, A.Y., S.Q. Zhang, A.M. da Silva, et al., 2001. Improving global analysis and short-range forecast using rainfall and moisture observations derived from TRMM and SSM/I passive microwave sensors. *Bull. Amer. Meteorol. Soc.*, **81**, 659–679.

Hou, A.Y., S.Q. Zhang and O. Reale, 2002. Variational continuous assimilation of TMI and SSM/I rain rates: Impact on GEOS-3 analysis and forecasts. *Mon. Weather Rev*., **132**, 2094–2109.

Kalnay, E. and R. Jenne, 1991. Summary of the NMC/NCAR Reanalysis Workshop of April 1991. *Bull. Amer. Meteorol. Soc*., **72**, 1897–1904.

Kalnay E., M. Kanamitsu, R. Kistler, et al., 1996. The NCEP/NCAR 40-year reanalysis project. *Bull. Amer. Meteorol. Soc*., **77**, 437–471.

Kanamitsu, M., W. Ebisuzaki, J. Woollen, et al., 2002. NCEP-DOE AMIP-II reanalysis (R-2). *Bull. Amer. Meteorol. Soc*., **83**, 1631–1643.

Kistler R., E. Kalnay, W. Collins, et al., 2001. The NCEP/NCAR 50-year Reanalysis: Monthly means CD-ROM and documentation. *Bull. Amer. Meteorol. Soc*., **82**, 247–267.

Lait, L.R., 2002. Systematic differences between radiosonde measurements. *Geophys. Res. Lett*., **29**, doi:10.1029/2001GL014337.

Lin, S.J., 2004. A "vertically Lagrangian" finite-volume dynamical core for global models. *Mon. Weather Rev*, **132**, 2293–2307.

Mesinger, F., G. DiMego, E. Kalnay, et al., 2006. North American regional reanalysis. *Bull. Amer. Meteorol. Soc*., **87**, 343–360.

Newman M., P.D. Sardeshmukh and J.W. Bergman, 2000. An assessment of the NCEP, NASA, and ECMWF reanalyses over the tropical west Pacific warm pool. *Bull. Amer. Meteorol. Soc*., **81**, 41–48.

Newson, R.,1998. Results of the WCRP First International Conference on Reanalysis, *GEWEX News*, **8**, 3–4.

Onogi, K., 2000. ERA-40 Project Report Series 2. The long-term performance of the radiosonde observing system to be used in ERA-40, European Centre for Medium-Range Weather Forecasts, August 2000, 77pp.

Onogi, K., J. Tsutsui, H. Koide, et al., 2007. The JRA-25 reanalysis. *J. Meteor. Soc. Jpn*., **85**, 369–432.

Ozone Trends Panel, 1988. *WMO Report of the International Ozone Trends Panel*. World Meteorological Organization Global Ozone Research and Monitoring Project, Report No. 18.

Pavelsky, T.M. and L.C. Smith, 2006. Intercomparison of four global precipitation datasets and their correlation with increased Eurasian river discharge to the Arctic Ocean. *J. Geophys. Res.*, **111**, D21112, doi:10.1029/2006JD007230.

Pawson, S., I. Štajner, S.R. Kawa, H. Hayashi, W.-W. Tan, J.E. Nielsen, Z. Zhu, L.-P. Chang and N.J. Livesey, 2007. Stratospheric transport using 6-h-averaged winds from a data assimilation system. *J. Geophys. Res.*, **112**, D23103, doi:10.1029/2006JD007673.

Redder, C.R., J.K. Luers and R.E. Eskridge, 2004. Unexplained discontinuity in the U.S. radiosonde temperature data, Part II: Stratosphere. *J. Atmos. Oceanic Tech.*, **21**, 1133–1144.

Roads J. and A.K. Betts, 2000. NCEP–NCAR and ECMWF Reanalysis surface water and energy budgets for the Mississippi River Basin. *J. Hydrometeorol.*, **1**, 88–94.

Rood, R.B., 2003. Reanalysis. In *Data Assimilation for the Earth System*. NATO Science Series: IV. Earth and Environmental Sciences 26, Swinbank, R., V. Shutyaev and W.A. Lahoz (eds.), Kluwer Academic Publishers, Dordrecht, The Netherlands, pp 361–372, 378pp.

Rood, R.B., D.J. Allen, W. Baker, et al., 1989. The use of assimilated stratospheric data in constituent transport calculations. *J. Atmos. Sci.,* **46**, 687–701.

Santer, B.D., T.M.L. Wigley, A.J. Simmons, et al., 2004. Identification of anthropogenic climate change using a second-generation reanalysis. *J. Geophys. Res*., **109**, D21104, doi:10.1029/2004JD005075.

Schoeberl, M.R., A.R. Douglass, Z. Zhu and S. Pawson, 2003. A comparison of the lower stratospheric age-spectra derived from a general circulation model and two data assimilation systems. *J. Geophys. Res*., **108**, Art. No. 4113.

Schubert, S.D. and Y. Chang, 1996. An objective method for inferring sources of model error. *Mon. Weather Rev.*, **124**, 325–340.

Schubert S.D., R.B. Rood and J. Pfaendtner, 1993. An assimilated dataset for earth-science applications. *Bull. Amer. Meteorol. Soc*., **74**, 2331–2342.

Serreze, M.C., M.P. Clark and D.H. Bromwich, 2003. Monitoring precipitation over the Arctic terrestrial drainage system: Data requirements, shortcomings, and applications of atmospheric reanalysis. *J. Hydrometeorol.*, **4**, 387–407.

Serreze, M.C. and C.M. Hurst, 2000. Representation of Arctic precipitation from NCEP–NCAR and ERA Reanalyses. *J. Climate*, **13**, 182–201.

Simmons, A.J., P.D. Jones, V. da Costa Bechtold, et al., 2004. Comparison of trends and low-frequency variability in CRU, ERA-40, and NCEP/NCAR analyses of surface air temperature. *J. Geophys. Res*., **109**, D24115, doi:10.1029/2004JD005306.

Štajner, I., N. Winslow, R.B. Rood and S. Pawson, 2004. Monitoring of observation errors in the assimilation of satellite ozone data, *J. Geophys. Res*., **109**, D06309, doi:10.1029/2003JD004118.

Stohl, A., O.R. Cooper and P. James, 2004. A cautionary note on the use of meteorological analysis fields for quantifying atmospheric mixing. *J. Atmos. Sci*., **61**, 1446–1453.

Tan, W.-W., M.A. Geller, S. Pawson and A. da Silva, 2004. A case study of excessive subtropical transport in the stratosphere of a data assimilation system. *J. Geophys. Res.*, **109**, Art. No. D11102.

Trenberth, K.E., J.T. Fasullo and J. Kiehl, 2008a. Earth's global energy budget. *Bull. Amer. Meteorol. Soc.*, doi:10.1175/2008BAMS2634.1.

Trenberth, K.E., T.R. Karl. and T.W. Spence, 2002. The need for a systems approach to climate observations. *Bull. Amer. Meteorol.*, Soc, **83**, 1593–1602.

Trenberth, K.E., T. Koike and K. Onogi, 2008b. Progress and prospects for reanalysis for weather and climate, *Eos, Trans. Amer. Geophys. Union*, **89**, 234–235.

Trenberth, K.E. and J.G. Olson, 1988. An evaluation and intercomparison of global analyses from NMC and ECMWF. *Bull. Amer. Meteorol. Soc.,* **69**, 1047–1057.

Trenberth, K.E. and L. Smith, 2008. Atmospheric energy budgets in the Japanese Reanalysis: Evaluation and variability. *J. Meteor. Soc. Jpn.*, **86**, 579–592.

Trenberth, K.E. and L. Smith, 2009. The three dimensional structure of the atmospheric energy budget: Methodology and evaluation. *Climate Dyn.*, **32**, doi:10.1007/s00382-008-0389-3.

Uppala, S.M., P.W. Kållberg, A.J. Simmons, et al., 2005. The ERA-40 re-analysis. *Q. J. R. Meteorol. Soc.*, **131**, 2961–3012.

Viterbo, P. and A.K. Betts, 1999. Impact of the ECMWF reanalysis soil water on forecasts of the July 1993 Mississippi flood. *J. Geophys. Res.*, **104**, 19361–19366.

WCRP, 1998. *Proceedings of the 1st WCRP International Conference on Reanalyses.* Silver Spring, MD, USA, 27–31 October, 1997, WMO/TD-N 876.

WCRP, 2000. *Proceedings of the 2nd WCRP International Conference on Reanalyses.* Wokefield Park, nr. Reading, UK, 23–27 August 1999, WCRP-109, WMO/TD-N 985.

# Observing System Simulation Experiments

**Michiko Masutani, Thomas W. Schlatter, Ronald M. Errico, Ad Stoffelen, Erik Andersson, William Lahoz, John S. Woollen, G. David Emmitt, Lars-Peter Riishøjgaard and Stephen J. Lord**

## 1 Definition and Motivation of OSSEs

Observing System Simulation Experiments (OSSEs) are typically designed to use data assimilation ideas (see chapter *Mathematical Concepts in Data Assimilation*, Nichols) to investigate the potential impacts of prospective observing systems (observation types and deployments). They may also be used to investigate current observational and data assimilation systems by testing the impact of new observations on them. The information obtained from OSSEs is generally difficult, or in some contexts impossible, to obtain in any other way.

In an OSSE, simulated rather than real observations are the input to a data assimilation system (DAS for short). Simulated observational values are drawn from some appropriate source (several possibilities have been considered; see Sect. 3). These values are generally augmented by implicitly or explicitly estimating respective values of observational errors to make them more realistic (see Sect. 4). The resulting values are then ingested into a DAS (that may be as complex as an operational one) just as corresponding real observations would be. Simulations of both analyses and subsequent forecasts are then produced for several experiments, with each considering a distinct envisioned observing system; i.e., a distinct set of observation types and characteristics. The analysis and forecast products are then compared to evaluate the impacts of the various systems considered.

OSSEs are closely related to Observing System Experiments (OSEs). For an observing system in operational use, the OSE methodology consists of:

- A control run in which all observational data currently used for every-day operations are included;
- A perturbation run from which the observation type under evaluation is *excluded* while all other data are kept as for the control;
- A comparison of forecast skill between the control and perturbation runs.

M. Masutani (✉)
NOAA/NWS/NCEP/EMC, Camp Springs, MD, USA; Wyle Information Systems, El Segundo, CA, USA
e-mail: Michiko.Masutani@noaa.gov

W. Lahoz et al. (eds.), *Data Assimilation*, DOI 10.1007/978-3-540-74703-1_24,

OSEs are effectively Data-Denial Experiments (DDEs, discussed in Sect. 7.1). They reveal specifically what happens when a DAS is degraded by removing particular subsets of observations and thus measure the impacts of those observations.

The structure of an OSSE is formally similar to that of an OSE with one important difference: OSSEs are assessment tools for *new* data, i.e., data obtained by hypothetical observing systems that do not yet exist. The methodology of an OSSE consists of:

- Generation of reference atmospheric states for the entire OSSE period. This is usually done with a good-quality, realistic atmospheric model in a free-running mode without data assimilation. This is often called the Nature Run (NR for short), providing the proxy "truth," from which observations are simulated and against which subsequent OSSE assimilation experiments are verified;
- The generation of simulated observations, including realistic errors, for all existing observing systems and for the hypothetical future observing system;
- A control run (or experiment) in which all the data representing the current operational observational data stream are included;
- A perturbation run (or experiment) in which the simulated candidate observations under evaluation are added;
- A comparison of forecast skill between the control and perturbation runs.

The most common motivation for OSSEs regards estimating the potential impact of proposed new observation types. Although a new type may be highly accurate and robust, it does not provide complete, instantaneous global coverage with perfect accuracy. All new observation types therefore will be used in conjunction with other, mostly already existing, observation types and a background derived from a short-term model forecast. Since data assimilation is a blending of all such useful information, the impact of a new type can only be estimated by considering it in the context of all the other useful types. It is therefore necessary to investigate potential impacts in a complete and realistic DAS context.

New observation types that do not yet exist cannot provide observational values to be assimilated. If a prototype does exist but is not already deployed as envisioned, impacts that can be currently measured may be unrepresentative of future potential impacts or not statistically significant. The latter is always an issue with data assimilation because the data analysis problem is fundamentally statistical due to unknown aspects of observational and modelling errors. Under these conditions, the only way of estimating the potential impact of new observations is by appropriately simulating them; i.e., performing an OSSE of some kind.

Besides estimating the impact, and therefore the value, of an augmentation to the observing system, an OSSE can be used to compare the effectiveness of competing observation designs or deployment options. What is the cost to benefit ratio, for example, between using a nadir-looking versus a side-scanning instrument on a satellite? Or, for a lidar, what are the relative benefits of using various power

settings for the beams? An OSSE can aid the design before putting an instrument in production. Thus, well-conducted OSSEs can be invaluable for deciding trade-offs between competing instrument proposals or designs: the cost of an OSSE is a tiny fraction of the cost of developing and deploying almost any new observing system.

Furthermore, by running OSSEs, current operational data assimilation systems can be tested, and upgraded to handle new data types and volume, thus accelerating use of future instruments and observing systems. Additionally, OSSEs can hasten database development, data processing (including formatting) and quality control software. Recent OSSEs show that some basic tuning strategies can be developed before the actual data become available. All of this accelerates the operational use of new observing systems. Through OSSEs future observing systems can be designed to optimize the use of data assimilation and forecast systems to improve weather forecasts, thus giving maximum societal and economic impact (Arnold and Dey 1986; Lord et al. 1997; Atlas 1997).

There is another motivation for OSSEs that has been less often discussed. It exploits the existence of a known "truth" in the context of an OSSE. For a variety of purposes, including validating or improving an existing DAS or designing perturbations for predictability studies or ensemble forecasting, it is useful to characterize critical aspects of analysis errors. Evidence to guide such characterization is generally elusive since the DAS-produced analyses themselves are often the best estimates of the atmospheric state (by design) and, therefore, there is no independent dataset for determining errors. All observations have presumably been used, accounting optimally (to some degree) for their error statistics and accounting for their mutual relationships in time (using a forecast model for extrapolation or interpolation) or in space (e.g. quasi-geostrophy and spatial correlations) and thus robust independent datasets for verification are usually absent (although, e.g., research data such as ozonesondes and ozone from some instruments are not commonly assimilated, and thus are available for independent verification). While some information about DAS errors can be derived from existing data sources, it necessarily is incomplete and imperfect. Although any OSSE is necessarily also an imperfect simulation of reality, the analysis and forecast errors can be completely and accurately computed and thus fully characterized within the simulated context.

The fact that they are widely used and relied upon does not mean that OSSEs, or the experimental results created by them, are free of controversy. Because of the wide-ranging consequences of decisions on major Earth Observing Systems, any OSSE results on which these decisions are based will have to withstand intense scrutiny and criticism. One goal of this chapter is to suggest ways in which OSSEs can be made robust and credible.

In this chapter we present the basic guidelines for conducting OSSEs. A historical review is provided, and experiences from OSSEs conducted at the National Centers for Environmental Prediction (NCEP OSSE) are presented; finally, conclusions and the way forward are outlined.

## 2 Historical Summary of OSSEs

The OSSE approach was first adopted in the meteorological community to assess the impact of prospective observations, i.e., not available from current instruments, in order to test potential improvements in numerical weather prediction, NWP (Nitta 1975; Atlas 1997; Lord et al. 1997; Atlas et al. 2003a). In a review paper, Arnold and Dey (1986) summarize the early history of OSSEs and present a description of the OSSE methodology, its capabilities and limitations, and considerations for the design of future experiments. Meanwhile, OSSEs have been performed to assess trade-offs in the design of observing networks and to test new observing systems (e.g. Stoffelen et al. 2006).

In early OSSE studies, the same model used to generate the "Nature Run" or truth was used to assimilate the synthetic data, and to run forecasts (Halem and Dlouhy 1984). In these so-called "identical twin" OSSEs the physical parametrizations and discretized dynamical processes in the assimilating model exactly represent those in the surrogate atmosphere. Model errors due to parametrization and numerical implementation are thus neglected and a free model forecast run from given initial conditions would provide identical results for the Nature Run and the DAS model. Consequently, forecast errors arising from deficiencies in the forecast model representation of the real atmosphere are not accounted for; only forecast errors due to errors in the initial conditions are represented. This limitation has been noted to lead to overly optimistic forecast skill in the OSSE DAS.

Another effect of the neglected model errors is that the differences between observations, both existing and future ones, and background (i.e., forecast), O-B, tend to be smaller in case of an identical twin OSSE than in operational practice (Atlas 1997; Stoffelen et al. 2006). As a result, both the observation minus analysis (O-A) differences and analysis impact of the observations, A-B (analysis less background), tend to be smaller than expected. Several ways exist to test the reduced observation impact and overly optimistic forecast skill: e.g., by comparing the O-B and O-A distributions, single observing system impacts, and forecast skill metrics in the OSSE and operational practice (calibration). The chapter *Evaluation of Assimilation Algorithms* (Talagrand) provides details of methods used to evaluate the assimilation process.

Since the DAS background model error space in identical twin OSSEs is limited with respect to an operational model's error space, fewer observations are needed to correct the model state in the analysis step. In fact, the simulated observation set, unlike the real observations, has systematic characteristics consistent with the model formulation (e.g. scales of motion, mass-wind balance). Therefore, just a few observations could potentially correct the initial state errors and provide improved forecasts in an identical twin OSSE. On the other hand, as Atlas et al. (1985) point out, due to the simplified error space, observation "saturation" in the DAS will tend to occur at lower data volumes in an identical twin OSSE than in the case of assimilation of the real observations. This saturation may lead to underestimation of the impact of observing systems with extensive coverage (e.g. satellite systems). Moreover, observing systems that tend to correct errors due to numerical truncation

of the dynamics or due to physical parametrization, may be undervalued. This potential non-linear effect of sampling on identical twin OSSE forecast scores, makes the above-mentioned calibration tests (involving, e.g., O-A and O-B distributions) on the OSSE data assimilation system increasingly relevant.

Arnold and Dey (1986) recommend "fraternal twin" OSSEs as a way to address the shortcomings of "identical twin" OSSEs. In fraternal twin OSSEs, the NWP model used to simulate the observations is different from the forecast model in the OSSE data assimilation system, but not as different as the true atmosphere is from an operational forecast model. Examples can be found in Rohaly and Krishnamurti (1993), Keil (2004) and Lahoz et al. (2005). It is clear that the problems noted above with identical twin experiments will be reduced, but not absent for fraternal twin experiments. Stoffelen et al. (2006) test the absence of unrealistic observation impact in a fraternal twin OSSE. To avoid potential fraternal twin problems, the Nature Run and atmospheric data base may be produced at one NWP centre (Becker et al. 1996), while the impact experiments are run by another independent NWP centre (Masutani et al. 2006, 2010).

Another reported measure to reduce identical twin effects is to produce the Nature Run at high resolution and run the OSSE data assimilation system at lower spatial resolution. While useful for some studies, a potential disadvantage is that the observing system impact of a prospective system is tested at a resolution which is obsolete by the time the new observing system will be operationally implemented.

Atlas et al. (1985) report on the exaggerated OSSE impact of satellite-derived temperature soundings. At that time, the fraternal twin problem was raised as one cause, although these satellite soundings are rather abundant (see above). Other, and with hindsight perhaps more plausible, noted causes are:

- Simplified observation error characteristics. Observing systems can have complicated relationships (geophysical, spatial, and temporal) with the forecast model's atmospheric state and special care is needed to simulate them;
- The simulated observation coverage is over-optimistic. For example, the degree of cloud contamination of the measurements may be underestimated (e.g. Masutani et al. 1999);
- The simplifying assumption, usually made in OSSEs, that the distribution of observation errors is perfectly known;
- Temperature data are both simulated and assimilated, with no error from the Radiative Transfer Model (RTM) involved.

Again, comparison of observation impact and forecast skill, e.g., by comparing the O-B and O-A distributions; single observing system impacts; and forecast skill metrics in the OSSE and operational practice involving OSSE calibration, should reveal such problems.

Various simulation experiments have been attempted which use real data for existing instruments and only simulate future instruments. These methods do not require a Nature Run and allow experimentation on a specific (extreme) weather

event. Observing System Replacement Experiments (OSREs) could, for example, be used to test the impact of existing wind profile observations over Northern Hemisphere land and how these may be replaced by another observing system (Cress and Wergen 2001). Although an OSRE indicates how one could replace existing observing systems, it is, however, not a priori clear how to extrapolate these results to faithfully test new observation capabilities, e.g., like the DWL (Doppler Wind Lidar) capability to resolve the incomplete wind profile coverage over the oceans. To test new observation capabilities, Marseille et al. (2008a, b, c) developed a method called the Sensitivity Observing System Experiment (SOSE). In a SOSE, adjoint sensitivity structures are used to define a pseudo-true atmospheric state for the simulation of the prospective observing system. In a SOSE the forecast error is projected back onto the initial state, thereby setting the maximum achievable forecast improvement. An alternative method, the Analysis Ensemble System (AES) (Tan et al. 2007) uses the spread in the ensemble as a proxy for the analysis and background uncertainty based on arguments of error growth (Fisher 2003). Since the background, analysis and observation errors are larger in the AES than in the real DAS, it is not clear whether the same set of observations in both systems remain optimal for reducing the background uncertainty. In order to test the realism of the OSRE, SOSE and AES, both the analysis and forecast impacts need to be carefully calibrated, just as in an OSSE.

In this chapter, the term OSSE (sometimes *full* OSSE to distinguish from other simulation experiments) refers to a simulation experiment with a Nature Run model significantly different from the NWP model used for data assimilation. This provides a truth independent of the data assimilation system NWP model and of the Global Observing System (GOS) data coverage and quality. In an OSSE, all observations used for the DAS have to be simulated from the Nature Run. In a SOSE, OSRE or AES, only the future observations are simulated from analysis or forecast fields. These fields used may have limitations in comparison with a Nature Run in terms of biases and temporal consistency due to the GOS, DAS and NWP (adjoint) model involved. It is considered that simulation of all observations is a significant initial investment for an OSSE, but that interpolating observations is part of a DAS. In OSSEs, all the usual analysis and forecast verification metrics can be used to evaluate data impact, and the simulated data can be tested with several different data assimilation systems with minor modification to the operational systems. The data impact for OSSEs (and their variants) often varies with verification metric and DAS used. Note, however, that a truth is available for further verification of the DAS characteristics. Although a SOSE, OSRE or AES allow quick study of real extreme events, the SOSE requires an adjoint model to generate the new observations and the AES requires an established ensemble system. Calibration and interpretation of the results is complicated and needs to be tested carefully for the SOSE, OSRE and AES. Full OSSEs with a long Nature Run allow quantitative assessment of the analysis and forecast impact. Note, however, that there are many OSSEs conducted without calibration. Furthermore, during the early years of OSSEs, identical twin OSSEs or fraternal twin OSSEs were often conducted due to the lack of variety in state-of-the-art NWP models.

To conclude, although initial investment is required for a full OSSE, it is today the most reliable strategy to use full OSSEs for impact assessment of prospective observing systems.

## 3 The Nature Run

The Nature Run is a long, uninterrupted forecast by a NWP model whose statistical behaviour matches that of the real atmosphere. The ideal Nature Run would be a coupled atmosphere-ocean-cryosphere model with a fully interactive lower boundary. However, it is still customary to supply the lower boundary conditions (sea surface temperature, SST, and ice cover) appropriate for the span of time being simulated. Meteorological science is approaching this ideal, but such coupled systems are not yet mature enough to be used for Nature Runs. Although fully coupled systems are available, their usefulness and accuracy for OSSEs is unknown. Preliminary tests, however, suggest that coupled systems may be good enough for operational NWP in the near future (Saha et al. 2006; Kistler et al. 2008).

The advantage of using a long, free-running forecast to simulate the Nature Run is that the simulated atmospheric system evolves continuously in a dynamically consistent way. One can extract atmospheric states at any time. Because the real atmosphere is a chaotic system governed mainly by conditions at its lower boundary, it diverges from the real atmosphere a few weeks after the simulation begins. This does not matter *provided that* the climatological statistics of the Nature Run match those of the real atmosphere. A Nature Run should be a separate universe, ultimately independent from but with very similar characteristics to the real atmosphere.

### *3.1 Characteristics of the Nature Run*

One of the challenges for an OSSE is to demonstrate that the Nature Run does have the same statistical behaviour as the real atmosphere in every aspect relevant to the observing system under scrutiny. For example, an OSSE for a wind-finding lidar on board a satellite requires a Nature Run with realistic cloud climatology because lidars operate at wavelengths for which thick clouds are opaque. The cloud distribution thus determines the location and number of observations.

The Nature Run is central to an OSSE. It defines the *true* atmospheric state against which forecasts using simulated observations will be evaluated. This concept deserves more explanation. In 1986, Andrew Lorenc suggested the following definition of the “truth”: the projection of the true state of the atmosphere onto the model basis. As an example, if a spectral model produces a Nature Run, the true atmospheric state might be represented by spectral coefficients corresponding to triangular truncation at total wave number $n$ ($Tn$) on $L$ vertical levels. Atmospheric features too small to be captured by the model resolution are not incorporated in this truth.

The Nature Run is also the source of simulated observations. For each observing system, existing or future, a set of realistic observing times and locations is developed along with a list of observed parameters. An interpolation algorithm looks at the accumulated output of the Nature Run, goes to the proper time and location and then extracts the value of the observed parameter. If the Nature Run does not explicitly provide an observed parameter, the parameter is estimated from related variables that the model does provide. Because observations extracted from the Nature Run are the same as the defined truth (they are "perfect"), various sources of error must also be simulated and added to form observations with realistic accuracy with respect to the Nature Run itself.

Some OSSEs have used a succession of atmospheric analyses as a substitute for a Nature Run (Keil 2004; Lahoz et al. 2005). A succession of analyses is a collection of snapshots of the real atmosphere. For example, in the case of four-dimensional variational assimilation (4D-Var, see chapter *Variational Assimilation*, Talagrand), although the analyses may each be a realizable model state, they all lie on different model trajectories. The background (first guess) lies on the same model trajectory as the previous analysis because, in 4D-Var the analysis is a realizable model state (it does not require separate initialization or balancing). Once this background is adjusted by new data in 4D-Var, the model lies on a new trajectory, which may be close to the old one (the one that the background was on) but is nonetheless different. Each analysis marks a discontinuity in model trajectory, determined by the information content extracted by a DAS from the existing global observing systems (see chapter *The Global Observing System*, Thépaut and Andersson). Furthermore, residual systematic effects due to the spatially non-uniform and often biased observations, the DAS or the model state, may either favourably or unfavourably affect the potential of new observing systems to improve the forecasts. Thus, considering a succession of analyses as truth seriously compromises the attempt to conduct a "clean" experiment.

## *3.2 Evaluation and Potential Adjustment of the Nature Run*

No Nature Run is perfect and its shortcomings need to be investigated by comparison with real-world climatology and, if the shortcomings can compromise a particular OSSE, adjustments to the Nature Run may be needed.

Several NCEP OSSEs (Masutani et al. 2006, 2010) have used the Nature Run with *T213* horizontal resolution and 31 vertical levels (*T213* NR) provided by the European Centre for Medium-Range Weather Forecasts (ECMWF) and described in Becker et al. (1996). For the *T213* NR, quadratic grids with 60 km horizontal resolution were used to compute the physics. Note the corresponding linear grid space would be 90 km, which is more representative of the scale resolved by the *T213* NR. A 1-month model run starting on 5 February 1993 was saved every 6 h.

It is important that the Nature Run contain realistic clouds for evaluation of Doppler Wind Lidar (DWL) and cloud motion vector (CMV) data and simulation of radiances. Doppler Wind Lidar data can be retrieved only if the DWL

shots hit the target. Clouds are important targets for a DWL but they also interfere with the DWL shots at lower atmospheric levels. Therefore, large differences in the Nature Run cloud amount will affect the sampling of simulated data. Realistic clouds are also necessary for generating realistic cloud track winds from geostationary platforms. Clouds moreover affect the sampling and simulation of radiance data.

The observed estimates for total cloud cover come from three different sources: the USAF Real-Time Nephanalysis (RTNEPH; Hamill 1992; Henderson-Sellers 1986); the International Satellite Cloud Climatology Project (ISCCP); and the NESDIS experimental product, Clouds from the Advanced Very high Resolution Radiometer (CLAVR-phase; http://cimss.ssec.wisc.edu/clavr). The differences between the total cloud cover (TCC) in the three observational sources and the Nature Run are within the variability of the observations. In the *T213* NR, the High-level Cloud Cover (HCC) amount seems larger than the satellite observed estimate across all areas of the globe. The amount of Low-level Cloud Cover (LCC) in the *T213* NR over the ocean is less than observed and the amount of LCC over snow is too high. After careful investigation, it was found that, due to the lack of reliable observations, there is no strong evidence for an over-estimation of HCC and polar cloud by the *T213* NR. However, the under-estimation of low level stratocumulus clouds over the oceans and its over-estimation over snow was clearly evident, and adjustments were consequently applied (see Masutani et al. 1999).

Although the OSSE using the *T213* NR produced many valuable results, it also had limitations. First of all, due to advances in model development, it is neither realistic nor suitable to use a Nature Run produced by a NWP model more than 10 years old to test a current DAS. Second, since there is a significant drift from analyses in the tropics during the first several weeks of the Nature Run, the 1-month long Nature Run cannot be used to evaluate data impact in the tropics. The *T213* NR employed fixed SSTs; although fixed SSTs were not found to jeopardize the OSSEs, this is still a serious limitation of the *T213* NR. Note that a more recent *T511* Nature Run produced by ECMWF (Reale et al. 2007) showed a reduction of tropical convective rainfall during the first few weeks of the Nature Run period. This may mean that the Nature Run has much less convective rainfall compared to the real atmosphere, or that the analysis has too much convective rainfall compared to the real atmosphere. For the Nature Run to be useful, its statistics must lie within the climatological variability in the real analyses.

Producing accurate tropical forcing is a challenge for current NWP models. Nevertheless, the recent *T511* NR produced by ECMWF (see above) faithfully reproduces many aspects of the tropical atmosphere, at least in a statistical sense (Reale et al. 2007). For example, it reproduces the African Easterly Jet and African Easterly Waves in good agreement with observations.

There is great interest in OSSEs for studying forecasts of tropical waves and tropical cyclones (TCs). A prerequisite for such studies is a Nature Run that generates realistic tropical disturbances, e.g., hurricanes with well defined warm cores and realistic tracks. However, there are still significant differences between model produced tropical cyclones and observations, and the interaction of TCs with

SSTs requires further study (Tsutsui and Kasahara 1996). Finally, the properties of tropical cyclones relevant to the evaluation of a DAS are still to be investigated.

Mid latitude cyclone statistics in the Nature Run must also be realistic. The basic measures commonly used to compile mid latitude cyclone statistics are:

- Distribution of cyclone strength across a wide range of pressures;
- Cyclone lifespan;
- Cyclone deepening;
- Regions of cyclogenesis and cyclolysis;
- Distribution of cyclone speed and direction.

### *3.3 Requirements for a Future Nature Run*

The preparation of the Nature Run and the simulation of data from it consume significant resources. It is of practical importance to have one or two good-quality Nature Runs shared by many OSSEs. OSSEs with different Nature Runs are difficult to compare but OSSEs using different data assimilation systems and the same Nature Run can provide valuable cross-validation of data impact results. If Nature Runs are widely accessible, the Nature Runs and simulated data ought to be shared between many of the institutes carrying out the actual OSSEs.

The primary specifications of a Nature Run based on past experience of OSSEs are:

a. Employ a NWP model with demonstrated forecast skill;
b. Simulation span: since the data impact depends on the season, it is important that future Nature Runs cover long periods, preferably a whole year to allow selection of interesting subperiods for closer study;
c. Simulation sample: a temporal resolution higher than the OSSE analysis cycle. If more than one DAS is involved, this would ideally be a resolution higher than that of all participating data assimilation systems;
d. Simulation should resolve scales compatible with the main observing systems;
e. It is desirable that they should be based on an atmosphere-ocean coupled model; or at least, the Nature Run must be forced by an analysis incorporating frequently updated SST and sea ice;
f. Data archiving should be user-friendly and shareable with the community;
g. Simulation should agree with the real analyses in a statistical sense;
h. Chemistry and aerosol information which affect the data should be evaluated;
i. There should be a trade-off between the resolution and the complexity of the model;

The set of archived Nature Run variables should be enhanced to accommodate the need for OSSEs. For example, geopotential height at model levels is very desirable. Archiving of this variable will help the simulation of observations based on

height coordinates, such as those from DWL and profilers. Low resolution pressure level data and isentropic level data output on a standard grid are also very useful for OSSEs, as they can be used for verification of the experiments. However, producing these verification datasets can take up significant resources at the initial stages of setting up an OSSE.

A main requirement of full OSSE experiments is to avoid the identical or fraternal twin ("incest") problem, as discussed in Sect. 2. If the model from which hypothetical observations are extracted is the same as the assimilating model, the OSSE results will show unrealistic observation impact and overly optimistic forecast skill (Arnold and Dey 1986; Stoffelen et al. 2006). Thus the forecast model used for the Nature Run should *not* be used later on for DAS experiments in the full OSSE.

## 4 Assignment of Realistic Observation Errors

The following definitions concern data assimilation in general – see also chapter *Mathematical Concepts of Data Assimilation* (Nichols). In the definitions below, $\mathbf{x}$, $\mathbf{y}$ and $\varepsilon$ are vectors, and $\mathcal{H}$ is a non-linear operator.

(a) *The observation*: $\mathbf{y} = \mathbf{y}_t + \varepsilon_m$
$\mathbf{y}$ is the observed value, measured by some instrument, and $\boldsymbol{\varepsilon}_m$ is the observation error. The subscript $t$ refers to the true atmospheric value. We define the true value as the weighted average of the true atmospheric values within the volume sampled by the instrument. Petersen (1968) defined the "true" observation in this way, but quantitatively by means of an integral. Different instruments sample different volumes so that the true value of temperature appropriate for a radiosonde may not match the true value appropriate for the AMDAR (Aircraft Meteorological DAta Relay) system aboard a commercial jet, even if the two observations are assigned to the same location and time. Thus, the observed "truth" is very much scale-dependent, but defining it in this way is consistent with the definition of truth with respect to model resolution as proposed by Lorenc (1986) and discussed immediately below.
$\varepsilon_m$ refers to errors incurred during measurement or subsequent data processing. The errors can be random or systematic (i.e., biased).

(b) *The model state*: $\mathbf{x} = \mathbf{x}_t + \varepsilon_f$
$\varepsilon_f$ is the model state error. The state of a DAS model is defined by a set of parameters stored at the points of a model grid, or, alternatively, by a set of spectral coefficients. As noted above, we follow Lorenc (1986), in defining the true model state $\mathbf{x}_t$ as the true atmospheric state containing all scales from long waves down to cloud microphysics, but spectrally truncated at the model grid. Scales of motion that cannot be captured by the model grid (or the spectral truncation) are not included in the definition of the true state. The numerical model forecasts the state $\mathbf{x}$, but the forecast is subject to error $\varepsilon_f$, which is the

result of truncation associated with finite differencing, imperfect dynamics, and errors in the representation of physical processes, whether parametrized or not.

(c) *The forward model*: $\mathcal{H}(\mathbf{x})$

Forecasts are usually verified against observations (sometimes against an analysis). Because observations hardly ever coincide with model grid points, it is necessary to map the model forecast to the observations in order to make a direct comparison. The forward model $\mathcal{H}$ does this. Another name for $\mathcal{H}$ is the *observation operator,* because $\mathcal{H}$ operates on the model grid to generate a pseudo-observation, a best estimate of the observed value. It relies on the parameters computed by the model on the model grid in order to make a best estimate of the observed value. Sometimes the calculation is as simple as 3-D linear interpolation, but if the observed quantity does not match one of the predicted quantities, then $\mathcal{H}$ will also involve a transformation of variables. For example, the model may predict relative humidity, but the observed quantity is column integrated water vapour. In this case, in addition to interpolation, the forward model has to convert the predicted relative humidity and temperature to a specific humidity and integrate the specific humidity vertically from the surface to the top of the model atmosphere.

(d) *Representativeness*: $\mathbf{y}_t = \mathcal{H}(\mathbf{x}_t) + \varepsilon_r$

If the forward model $\mathcal{H}$ could be applied to the true values $\mathbf{x}_t$ (unknown in practice) on the model grid, we would have an observation that still lacks a representativeness error $\varepsilon_r$. The representativeness error has two causes:

(1) The model grid volume does not match the atmospheric volume that is the object of measurement. If the observed volume is small compared to the model grid volume, the measurement will represent scales of motion that the model grid cannot resolve. From the model's point of view, the observation contains subgrid scale noise, and this will contribute to the value of $\varepsilon_r$. In other words, because the representation of $\mathbf{x}_t$ is spectrally truncated, the projection $\mathcal{H}(\mathbf{x}_t)$ does not capture the subgrid scale atmospheric variance inherent in the observation. If the observed volume is larger than the model grid volume (e.g. a measurement of radiance in the microwave portion of the electromagnetic spectrum could involve a volume of atmosphere larger than the model grid volume), then the forward model will be an averaging operator rather than an interpolation operator. From the model's point of view, the observation is too smooth and $\varepsilon_r$ will relate to how well the model average spatially and temporally represents $\mathbf{y}$.

(2) If a transformation of variables is included in $\mathcal{H}$, the relationship is imperfectly known or it is approximated in order to minimize the number of computations, e.g., in case of radiance observations. This also contributes to $\varepsilon_r$. In fact, any operation incorporated in $\mathcal{H}$ may contribute an error component to $\varepsilon_r$.

To summarize, representativeness error arises from the mismatch between the DAS model grid volume and the volume sampled by the instrument, and also from a mismatch between the observed and predicted variables.

Some aspects of the representativeness error are not random but systematic. Even if we exclude subgrid effects that may be small if the model resolution is high enough, a computationally fast radiative transfer model applied to an atmospheric profile will generally yield imperfect radiances compared to the real atmosphere. This error will be almost identical whenever the atmospheric profiles are the same, since the model and physics remain unchanged. If such imperfections are complex functions of the atmospheric state, they may appear as random errors when computed from collections of states although they are in fact systematic. Modelling the representativeness error as though it were random may therefore introduce unrealistic effects if some aspect of the systematic nature of the error is important.

In all aspects of the data assimilation problem, representativeness error appears combined with instrument error. The combined error is called the "observation error", and its covariance (denoted by **R**) is a key statistic that determines the analysis error covariance. If the instrument and representativeness errors are uncorrelated, then $\mathbf{R} = \mathbf{E} + \mathbf{F}$, where **E** and **F** are the covariances of instrument and representativeness errors, respectively:

$$\mathbf{E} = \mathcal{E}[(\varepsilon_m - \langle\varepsilon_m\rangle)(\varepsilon_m - \langle\varepsilon_m\rangle)^T] \text{ and}$$
$$\mathbf{F} = \mathcal{E}[(\varepsilon_r - \langle\varepsilon_r\rangle)(\varepsilon_r - \langle\varepsilon_r\rangle)^T]$$

where $\langle\ \rangle$ denotes an average, $\mathcal{E}$ [.] denotes expectation value, and the superscript $T$ means vector transpose. The **R**, rather than **E** or **F**, is actually specified in the DAS. **E**, **F** and **R** are matrices.

Techniques to estimate **R** are imperfect. A poor specification will yield a suboptimal system; i.e., one with larger analysis error variance than otherwise. Generally, some further tuning of the error estimates is conducted so that the **R** incorporated in the system experiments appears close to optimal. These are, therefore, generally the values that must be duplicated as the observational errors in the OSSE if the responses in the real and simulated systems are to appear similar.

(e) *Application to OSSEs*:

In practice, real observations come with only an instrument error; they are inherently *representative* of the volume of atmosphere sampled. The representativeness error arises from the forward operator and has the two components mentioned above. We account for instrument error and, to be rigorous, also for the representativeness error, when we specify the observation error covariance in the DAS penalty function that is part of the variational analysis. In practice, we compute $\mathcal{H}(\mathbf{x})$, not $\mathcal{H}(\mathbf{x}_t)$.

By contrast, in an OSSE, one uses a forward model to *generate* an observation. After the forward model is applied to the grid point values of the Nature Run, we must add a random contribution $\varepsilon_r$ to the forward model output. The finer the resolution of the Nature Run and the more accurate the forward model, the smaller the

representativeness error will be. Finally, we must also add an appropriate instrument error to improve realism. In summary, we must compute:

$$\mathbf{y} = \mathbf{y}_t + \varepsilon_m = \mathcal{H}(\mathbf{x}_t) + \varepsilon_r + \varepsilon_m.$$

The random contribution $\varepsilon_r$ accounts for the missing subgrid scale variance, say ${\varepsilon_r}^{\mathrm{s}}$ and any error associated with a transformation in the forward model, say ${\varepsilon_r}^{\mathrm{H}} = \mathcal{H}_t(\mathbf{x}_t) - \mathcal{H}(\mathbf{x}_t)$ (so $\varepsilon_r = {\varepsilon_r}^{\mathrm{s}} + {\varepsilon_r}^{\mathrm{H}}$), where $\mathcal{H}_t$ is a hypothetical perfect forward model that operates on the NWP model defined truth. We find that:

$$\mathbf{y} = \mathcal{H}(\mathbf{x}_t) + \varepsilon_r^S - \mathcal{H}(\mathbf{x}_t) + \mathcal{H}_t(\mathbf{x}_t) + \varepsilon_m = \mathcal{H}_t(\mathbf{x}_t) + \varepsilon_r^S + \varepsilon_m$$

Note that $\mathbf{y}$ represents the Nature Run transformed through the hypothetical perfect forward model. Additionally, one should consider whether the difference $\mathcal{H}_t(\mathbf{x}_t) - \mathcal{H}(\mathbf{x}_t)$ might have a systematic component (i.e., a bias), since a normal random error distribution is assumed above in $\varepsilon_r$.

## 5 Simulation of Observations

### *5.1 Basic Guidelines*

Although a particular OSSE may be motivated by evaluation of a single instrument, it is still generally necessary to simulate all observations that are expected to be used along with it. Even a poor observing system will be better than none at all since the atmosphere is chaotic. Irrespective of how close to the real atmosphere a data assimilation experiment begins, without the constraint of further observations, after 15 days or so it will diverge to states expected to be as dissimilar to the atmosphere as two states randomly selected for the same month but different years. Thus, using a single observation type in an OSSE with other observations excluded results in a very large impact compared with no assimilation at all, but a much smaller and more realistic impact if other observations *are* considered.

Current observations quite effectively constrain the atmospheric analysis. In many places, the expected error variance of the analysis is less than that of most observations that have been employed in the analysis of the DAS model state (Note that many observations contain more information about the local atmosphere than the analysis; however, in the truncated model domain, the errors of these observations are larger, due to the representativeness error.). The analysis is better because it has used all nearby observations, including those implied by the background, accounting for the error statistics of each, at least in a crude but still useful way. The weighting of a new observation within the DAS will be determined by the presence of other observations. The impact of any additional observation essentially competes with that of all others. When the impacts of any single observation type

are measured, therefore, the improvements to the analysis or forecasts are generally quite small. Progress occurs when innovative instruments are added to those already used, but by small steps rather than great leaps.

Once the Nature Run is sufficiently validated, observations may be simulated. To do so, it is necessary to understand the relationship between the observations and the atmosphere, both the real atmosphere and the one represented by the Nature Run. Furthermore, at the next step in preparing the OSSE, simulated errors are generated to add to the corresponding simulated observations. The accuracy with which the DAS can reproduce the Nature Run in the OSSE will depend strongly on the characteristics of the errors associated with the observations. Prior to selecting a method for simulating the observations, it is therefore prudent to also understand the nature of all the types of error realistically associated with them.

Various observing instruments are designed to respond to differing atmospheric characteristics. Here, two such instruments will be contrasted: (i) a radiosonde, and (ii) a satellite instrument that measures infrared radiances. Together they represent several of the various aspects that must be considered when simulating observations and their errors.

The radiosonde is a comparatively simple instrument with a thermo-resistor used to measure temperature as the balloon ascends. The measurement is made along short segments of the trajectory of the balloon, with their length determined by the response and reporting times of the instrument. Compared with the much coarser resolution represented by the Nature Run, these may be considered as (almost) point values that are affected by all spatial scales. A function must therefore be developed to relate the observed value to the atmosphere as represented by the assimilating model (i.e., a function for the spatial representativeness error).

The other instrument is on board a satellite designed to measure infrared radiances coming from the Earth and atmosphere below. The satellite actually measures the energy of photons over some range of electromagnetic wavelengths collected on an antenna (see chapter *Research Satellites*, Lahoz). For the purpose of NWP as opposed to climate monitoring, data assimilation is mainly concerned with atmospheric fields: temperature, wind, pressure and constituents (e.g. water vapour, ozone, and perhaps minor species and aerosols). The observed radiances must be related to these fields if they are to be useful. Presumably an appropriate relationship exists; otherwise the observation would not be used for this purpose. The antenna collects radiances emitted from a possibly large volume of the atmosphere and is therefore most accurately related to some kind of average (with spatial weights determined by the viewing characteristics of the antenna and the orbiting satellite). This average will not in general correspond to that defining a grid volume average in a model data representation. Thus, some spatial interpolation or integrating relationship must also be defined.

For any observation types already used within a DAS, a useful relationship between what is observed and the representation of fields being analysed necessarily already exists. In the standard notation used for atmospheric data assimilation, this is the operator $\mathcal{H}$ that acts on the background field during the assimilation cycle

(see Sect. 4). For a new instrument not yet used, this operator needs to be developed. Development can be either empirical or physically based.

In general, $\mathcal{H}$ can be expanded into a sequence of one to several distinct functions acting on the state $\mathbf{x}$; e.g., as:

$$\mathcal{H}(\mathbf{x}) = \mathcal{S}(\mathcal{F}(\mathcal{I}(\mathbf{x})))$$

The function $\mathcal{I}$ denotes a possible interpolation from grid-point (or other discrete data representation) to observation locations; $\mathcal{F}$ denotes a possible physical (or other) relationship such as radiative transfer relating temperature and moisture to satellite-observed radiances; $\mathcal{S}$ denotes a possible integration of values, such as along a line of sight or within an antenna footprint. Any of $\mathcal{S}$, $\mathcal{F}$, or $\mathcal{I}$ may be absent for a particular observation type, and some types may be better described by a different sequence of operators or the employment of additional ones. The equation should therefore be considered as schematic, although for some observation types the presentation may be precise.

The more realistic the relationship between values representing the model state and the observed quantity, the more useful the real observation will be to the DAS and, correspondingly, the more realistic the simulated observation will be in the OSSE. A problem is that the time to develop the most accurate relationships may be prohibitive, and the benefits may be tiny compared to other shortcomings in the system. A relationship must be designed to be "good enough" for the intended purpose. Results must be carefully interpreted mindful of these criteria. The way these choices are evaluated will depend on the purpose. Inaccuracies in the results when compared to the "true" physical relationship can be handled to some degree by the statistical approach to representing errors in the DAS.

The $\mathcal{H}$ is designed with speed as well as accuracy in mind, especially if the DAS solves a large variational problem. In that case, a tangent linear version of $\mathcal{H}$ and its adjoint (see chapter *Variational Assimilation*, Talagrand) are generally applied to every iteration of the analysis increments (i.e., the difference between the analysis and the forecast). Thus, some compromises may be made that are not necessary when speed is not an issue. An example of this latter case is the generation of simulated observations from the Nature Run; these need only be produced once to be used in all subsequent relevant OSSEs. Thus, the simulation of observations from the Nature Run need not be done in the same way as the assimilation model. In fact, there are good reasons for selecting a different algorithm. These and other considerations are described in the next section.

## *5.2 Specific Issues Related to Different Observational Types*

Standard and simple forward models are used for extracting observed quantities from the "true" (i.e., Nature Run) background fields as the basis for the simulation of observations for use in OSSE experiments. This procedure will inevitably omit some

fraction of the error (from instrument variability and lack of model representativeness) to be found in real observations. Thus, simulation of observations for OSSE work is usually thought of as the synthesis of a signal from the background truth field (often referred to as a "perfect" observation), and some appropriate amount of noise, or "error." If the noise or error is indeed appropriate, then the impact of simulated observations on an OSSE will be similar to the impact that real-world observations have on operational assimilation. Although the instrument errors are in most cases fairly well defined, the derivation of the total error levels appropriate for application to perfect observations is a complex subject. This section describes some of the issues surrounding the creation of the perfect part of simulated observations.

### 5.2.1 Simulation of Conventional Observations

In order to create perfect observations, it is only necessary to locate the observation type to be simulated in the space and time coordinates of the background field. The most straightforward approach to this problem, for the case of simulating existing data sources, is just to use the locations of real observations for any given time and place. In the case of conventional observation sources (for example, TEMP, PILOT, SYNOP, AIREP, SHIP, BUOY, SATOB; see chapter *The Global Observing System*, Thépaut and Andersson) real world data patterns are readily available, and the specification of realistic simulated data patterns for these data types is simple. For the purposes of many OSSE experiments already conducted, this technique of locating conventional observing patterns is sufficient. However, in the set of simulated observations, the effects of observation circumstances and the expected evolution of the observing system should also be taken into account. Below we discuss several examples.

Radiosonde launch points can be located from existing real world datasets, but the balloon ascent and drift will depend on the atmosphere being sampled. The track of each radiosonde can be calculated using relatively simple transport models. For maximum realism, the calculation should be stepped at intervals sufficiently small to obtain information from the full vertical resolution of the Nature Run true fields. The resulting simulated profiles might be used without change in OSSE experiments, but would more likely be transformed into the more recognizable pattern of mandatory and significant vertical levels as presented to an operational DAS.

Surface land observations (for example, SYNOPs, METAR) present several issues to be considered for achieving realistic simulations. The question of location involves mainly the surface elevation and the measuring height. Although most real-world analogues contain some measure of the observation height, it may be advantageous in some cases to use a very high resolution digital elevation model and tables of particular instrument measuring heights to locate these data. There is also a need to interpolate surface values from the Nature Run background fields on a smoothed topography to a realistic topography of simulated observation points.

Commercial aircraft, the source of most aircraft observations, fly routes which use wind patterns to save fuel cost and avoid turbulence. Ideally, flight tracks for the OSSE should be formulated for simulated aircraft in the same way as they are

for real cases. However, the location of jets and turbulence can be very different for the Nature Run and the real world; the flight planning software is complicated, proprietary and even unique to individual airlines. It may be possible and worthwhile to develop a simplified generalized approach to formulating simulated flight track planning based on some general principles, in lieu of using the actual software employed by the airlines.

Cloud-tracked wind observations, and their unique observing errors, will depend on the specification and perception of cloud fields from the Nature Run. Satellite-borne instruments and observations of all types have unique relationships with various types of clouds, so this is a very important aspect for realistic simulation of satellite-based observations.

In general, it seems desirable to make use of synoptic features from the background truth fields to determine realistic locations for all simulated observations, at least to the extent this can be accomplished without exerting undue effort, or employing unrealistic assumptions. Many more OSSE experiments will need to be designed, conducted, and carefully examined in order to determine how important a realistic distribution of simulated observation locations is.

### 5.2.2 Simulation of Radiance Data

For the NCEP OSSE (see Sect. 9), the use of different Radiative Transfer Models (RTMs) for simulation and assimilation helps understand the errors associated with RTMs. Radiative transfer models used for simulation have been generally based on the RTTOV-6 (Radiative Transfer for TOVS) algorithm (Saunders et al. 1999). At NCEP, the OPTRAN model developed by NESDIS was used in the assimilation (Kleespies et al. 2004). Brightness temperatures were simulated and level-1B radiances synthesized with correlated measurement errors; the impact of clouds was also considered (Kleespies and Cosby 2001). Currently, the Community Radiative Transfer Model (CRTM) (Han et al. 2006; Weng 2007) and RTTOV are widely used in operational data assimilation systems. The SARTA (Stand-alone AIRS Radiative Transfer Algorithm) model (Strow et al. 1998) is also available and has been routinely used to simulate radiance data. These models allow the implementation of OSSEs using different RTMs for simulation and assimilation.

The simulation of radiances involves many procedures: simulation of orbits, evaluation of cloudiness, and assignment of surface conditions. Various properties such as surface emissivity and spectral response function have to be evaluated for each instrument. The characteristics of the instruments can change after launch, requiring a different set of coefficients at each stage. Ideally, the radiance data would be simulated as the Nature Run is produced. However, it is safer to save the Nature Run output frequently and simulate the radiance data afterwards, since radiances have to be simulated repeatedly with various conditions and error assignments.

If only clear-sky radiance data are used, a subgrid-scale sampling algorithm has to be developed when the radiances are simulated. If the footprint sizes are smaller than the Nature Run grid spacing, clear radiance data through small holes within the

cloudy grid have to be simulated. Using a probabilistic procedure to simulate cloud porosity is a possible way to produce the correct statistics. A functional relationship between clear sky probability and cloud fraction profile has to be derived to obtain a reasonable distribution (e.g. Marseille and Stoffelen 2003). If the cloud cover is used simply as a cut-off criterion for clear sky radiances, much of the clear sky radiance data from the porous areas of cumulus clouds are eliminated and large amounts of radiance data from above the clouds will be eliminated. Note that there are many stratospheric channels which are never affected by cloud.

Although both the OPTRAN and RTTOV models can simulate cloudy radiances, cloudy radiances have not been used in data assimilation systems (McNally et al. 2000). Further development of RTMs will include cloudy radiances in data assimilation systems (Liu and Weng 2006a, b). Cloudy radiances allow the simulation of imagery and moisture channels. While most of these channels may not be used for data assimilation, imagery and moisture channels can be used with observations to evaluate the Nature Run as well as the RTM itself. Note that since the Nature Run does not resolve cloud scales, even when radiances are modelled through cloud fraction, subgrid-scale clouds still need to be represented appropriately (e.g. in a statistical sense). Modelling the subgrid-scale cloud remains important to simulate cloudy radiances and for assimilation of radiance data. Testing RTMs with clouds is an important area for OSSEs.

Calibration of the radiance data includes a sampling algorithm which produces a similar distribution of observations as the real data. The adjoint technique (Zhu and Gelaro 2008) is especially useful in the calibration of radiance data, as it allows the skill of an individual channel to be assessed. The skill has to be evaluated for various conditions, as real errors are likely to be a function of geography, local atmospheric flow, season, and viewing angle. These errors are also likely to be correlated. The bias, variance, error correlation, and distribution function for the errors have to be modelled to be used by any data assimilation system. Bias correction is now a part of data assimilation systems (see chapter *Bias Estimation*, Ménard). As a result, one can bias correct the Nature Run radiances or implement the bias correction in the DAS itself.

### 5.2.3 Simulation of Doppler Wind Lidar (DWL) Data

As noted in the introduction (Sect. 1), one of the primary uses of OSSEs is to investigate and quantify the potential impact of a new observing system or combination of observing systems not currently being used together. No other instrument has been subjected to OSSE evaluation more than the Doppler Wind Lidar (DWL). With only radiosondes and a few radar wind profilers providing complete vertical profiles of the horizontal wind vector, gaining insight into the impact of a new wind profiler, especially over oceans and sparsely populated land areas, requires simulating the performance of the sounder without the benefit of a heritage instrument. Issues of observation errors including measurement errors and error of representativeness must be addressed. The DWL instrument is critically affected by both clouds

and aerosols. While clouds are represented reasonably well by current numerical models, aerosols are not.

In the United States, NASA and the Department of Defense (DoD) have supported the development of a Doppler Lidar Simulation Model, DLSM (Wood et al. 2000; Emmitt and Wood 2001). The DLSM was designed specifically to operate with the Nature Runs generated for OSSEs. Much attention has been given to incorporating cloud effects on the scale of the lidar beams (~100 m) and representing subgrid-scale turbulence that would affect the precision of the DWL line-of-sight (LOS) measurement (Emmitt and Wood 1989, 1991a).

A major role for OSSEs in preparing for a space-based DWL mission has been the generation of data requirements and subsequently derived instrument design specifications (Atlas et al. 2003b). Instrument designers have used the DLSM to conduct NWP impact trade-off studies related to orbit, instrument wavelengths, laser pulse energies, and signal processing strategies (Emmitt and Wood 1991b). NASA and NOAA have conducted numerous OSSEs using DWL observations simulated by the DLSM (Atlas and Emmitt 1995; Lord et al. 2002; Masutani et al. 2003; Riishøjgaard et al. 2003; Woollen et al. 2008).

In Europe, a similar Doppler Lidar In-space Performance Atmospheric Simulator (LIPAS) has been developed (Marseille and Stoffelen 2003) in support of the ADM-Aeolus mission to fly a space-borne DWL in 2011 (Stoffelen et al. 2005) – see chapter *Research Satellites* (Lahoz). LIPAS has been used to conduct OSSEs (Stoffelen et al. 2006) and simulates aerosol variability, vertical overlap of clouds and all relevant instrument performance characteristics.

The usual OSSE process involves a team composed of representatives of the operational weather forecasting community, instrument specialists and data stakeholders. The availability of models such as the DLSM and LIPAS allows the optimistic perspective of the instrument proposers and the more cautious expectations from the NWP communities to be explored over a range of assumed instrument performance within a realistic model and data assimilation environment. In the case of the DWL, the competition with other sources of wind information (including wind information contained in the background state) leads to an integrated impact which is usually more modest than that expected by the technologists. On the other hand, synergies with other sources of wind information (e.g. scatterometers and cloud motion vectors) are illuminated in ways not easily quantified without the OSSE.

## 6 Initial Conditions and Spin-Up Period

### *6.1 Initial Conditions*

The initial conditions for an OSSE must be generated carefully to reduce noise due to the difference between the Nature Run and the NWP model used for OSSEs. If an appropriate initial condition is not used, the OSSE will be contaminated by noise from the initial conditions and it will be hard to assess the data impact.

When starting a limited-period OSSE at some point within the Nature Run, initial conditions have to be generated carefully. If the initial conditions are generated from a different model, large biases between the models have to be removed, and some model variables may have to be estimated. Possible strategies to generate initial conditions include:

(i) *Generate the initial conditions by interpolation from the Nature Run*: It is possible to interpolate the initial conditions from the Nature Run to an OSSE model grid and use this as the initial conditions. As there is a large amount of noise produced from inconsistent initial conditions, it usually takes a few weeks for the OSSE to settle down. This procedure requires careful development of the interpolation procedure. Both the differences between the model variables and the bias between the Nature Run and the OSSE data assimilation system have to be carefully handled.
(ii) *Take the initial conditions from a precursor analysis*: In this approach one generates a precursor analysis starting from the same time and date as the Nature Run and uses the analysis as the initial conditions with the same DAS used for the OSSEs. The precursor analysis does not have to be of a high resolution but should be provided at the lowest resolution used for the OSSE. Not all operational data have to be included, but there should be enough data over the ocean to provide a reasonable description of large scale features, particularly in the Southern Hemisphere.

If the DAS used for the precursor run is the same as the OSSE data assimilation system, but has a higher resolution than the precursor analysis, the transition from the precursor analysis to the OSSE will be smooth. However, it takes a few time steps for the OSSE system to show the full resolution features.

If the OSSE DAS is different from the DAS used for the precursor run, an interpolation has to be performed. Exchanging analyses between different DAS is routinely done in real operational forecasting. This process can also be evaluated by OSSEs.

## *6.2 Spin-Up Period*

A real analysis is used for the initial conditions of the Nature Run. During the first 2–3 weeks, a drift occurs from the real atmosphere to the model atmosphere, particularly in the tropics. This period (called the spin-up period) should not be used for an OSSE because it lies within the limit of predictability (at least for the largest scales) and still contains traces of the real atmospheric conditions.

The Nature Run NWP model and initial analysis have errors that depend on the real atmospheric state due to data distribution, and DAS and NWP model specification. When the Nature Run state has evolved to one which is unrelated to the real atmosphere, these errors can be assumed to have disappeared. One can use trends in the O-B (observation minus background) and O-A (observation minus analysis)

differences to determine whether the Nature Run errors are independent of those of the real atmosphere. Depending on the type of experiment, the time for error independence to occur could be less than 2–3 weeks (see above).

## 7 Evaluation of OSSE Results

The data impact in an analysis and forecast could be very different. For example, if the model is not performing well, large differences between the background (forecast) and observations will create a large analysis impact; however, that improvement will not be maintained in the forecast skill. On the other hand, a small analysis impact may become a large forecast improvement in areas where the model is performing well. The areas showing data impact in the analysis and forecast may not be the same. Improvements can also propagate between regions: e.g., improvements in upper level wind will propagate towards lower levels in the forecast.

Data impact varies with spatial and time scales. For example, the impact in the mass fields could be very different from the impact in the wind fields. Below we discuss various aspects of data impact.

### *7.1 Data Denial (or Adding) Experiments (DDEs)*

The most common method used to test the impact of specific data is to compare the analysis and forecast skill with and without the specific data. Many diagnostic methods used to evaluate the Nature Run can also be used to evaluate the forecast and analysis. With real data the impact is measured as the forecast skill without the specific data compared against the best analysis or fit to observations. Usually, the analysis with the most data is considered to be the best and used as the control (defined in Sect. 1). Various skill scores for simulated experiments can be evaluated against either the control experiment or the Nature Run itself, while experiments with real data can be evaluated only against the control.

There are many evaluation methods, but it is important to produce a consistent evaluation for all experiments when the results are compared. Many diagnostic techniques used to evaluate the Nature Run can also be used to evaluate the results. Examples are given below.

(1) Root Mean Square Error (RMSE). Root mean square error does not require climatology; therefore, this is the easiest evaluation that can be performed, and is often the first evaluation to be implemented. In a real system, RMSE is computed as the departure from the control experiment, which is usually the analysis with the most observations. For simulated experiments, RMSE can be computed from the departure from the Nature Run. The RMSE can be evaluated with the zonal mean or the time mean removed;

(2) Anomaly correlation (AC). Anomaly correlation is affected by the climatology used, so it is important to use the same climatology for all skill comparisons. It is better to use a less than perfect climatology than to use different climatologies in skill comparisons. Traditionally, the AC of the 500 hPa geopotential height has been used, but Masutani et al. (2006, 2010) showed that other levels and variables need to be evaluated. Calculating ACs for different spatial scales is also crucial;
(3) Storm track and intensity. Evaluations are done to determine the improvement in the storm track for selected events;
(4) Fit to observations. This requires a forward model (see Sect. 4). For the NCEP OSSEs, an evaluation against Nature Run will replace this method. It is still important to compare the fit to observations during the calibration process, i.e., test the realism of the O-B and the O-A distributions (see chapter *Evaluation of Assimilation Algorithms*, Talagrand);
(5) Evaluation of the realism of a Nature Run by assessing the likelihood of extremes lying outside the normal range of analysed or measured values;
(6) Amplitude, wavelength and propagation speed (or phase) of waves;
(7) Comparisons which may shed light on the realism of disturbances in the model and identify possibly unrealistic or spurious scales of motion;
(8) Evaluating the analysis and forecast of precipitation using, e.g., threat scores *TS* ($TS = AC/(AF + AO - AC)$, where $AC =$ area correct, $AF =$ area forecast, $AO =$ area observed);
(9) The statistics of analysis increments. Errico et al. (2007) showed that the spectral decomposition of analysis increments reveals the performance of a DAS.

## *7.2 Adjoint–Based Techniques*

An adjoint–based technique (ADJ) to estimate the impact of observations on NWP analyses has been developed and is described in detail in Langland and Baker (2004) – see also chapter *Variational Assimilation* (Talagrand). This is a powerful method that describes the contributions from different observations. This technique allows detection of impact, be it positive or negative, from any observation. There are advantages and disadvantages compared with Data Denial Experiments (DDEs) (Zhu and Gelaro 2008; Gelaro and Zhu 2009):

- The ADJ measures the impacts of observations in the context of all other observations present in the assimilation system, while the observing system is modified in the DDE (i.e., gain matrix differs for each DDE member);
- The ADJ measures the impact of observations separately at every analysis cycle versus the background, while the DDE measures the total impact of removing data information accumulated in both the background and analysis;

- The ADJ measures the response of a single forecast metric to all perturbations of the observing system, while the OSE measures the effect of a single perturbation on all forecast metrics;
- The ADJ is restricted by the tangent linear assumption (valid ~1–3 days), while the DDE is not;
- The ADJ and DDE techniques produce a similar qualitative pattern on the short-term forecast with some exceptions;
- The ADJ may help our understanding in the interactions and redundancies among various observing systems.

## 8 Calibration of OSSEs

Calibration of OSSEs verifies the simulated data impact by comparing it to real data impact. In order to conduct an OSSE calibration, the data impact of existing instruments has to be compared to their impact in the OSSE.

The simulated impact experiments should mimic the equivalent real experiments. In any case, the observation-minus-background (i.e., forecast) difference is the sum of three terms: the measurement error, the representativeness error, and a background error transformed by $\mathcal{H}$. Realistic estimates of the variances and spatial covariance of these errors must be made for an effective OSSE. One way to ensure that measurement errors, representativeness errors, and forecast (background) errors are all properly specified is to compare the statistical properties of $\mathbf{y} - \mathcal{H}(\mathbf{x})$ (the innovation) of the OSSE with those of the real world assimilation $\mathbf{y} - \mathcal{H}(\mathbf{x})$ for each observing system; they should match. Similarly, the statistical properties of the analysis increments for the OSSE and the real world assimilation should match. Thus, distributions of observation minus background (O-B) differences and observation minus analysis (O-A) differences for each observation type in the simulation should be similar to the statistics in an equivalent experiment with real data. In effect, the simulated observations should force the OSSE model state toward the Nature Run in the same way that real observations force the operational model state toward the projected true atmospheric state.

One way of calibrating an OSSE is to use a DDE (see Sect. 7.1) to find out whether the assimilation of a specific type of observation has the same statistical effect on a forecast within the simulation as it does in the real world. For example, if automated aircraft reports are withheld from an operational data assimilation system, will the statistical measures of forecast degradation be the same as they would be in a system where all observation types are simulated and the Nature Run provides truth? An alternative method of calibration is to use the ADJ (see Sect. 7.2) to adjust the observational error so as to achieve a similar data impact with real observations.

When calibrating the OSSE, similarity in the amount of impact from existing data in the real and simulated atmospheres needs to be achieved. If the impacts are different this needs to be explained. For example, synoptic systems in the Nature Run and the real world are different, and that will cause differences in the data impact. If the

differences are caused by the procedure used in simulating the data, the simulation of the data has to be repeated until a satisfactory agreement is achieved.

Ideally, a complete calibration would be performed every time the DAS changes. However, we would spend our entire resources on calibration if we try for perfection. Of course, we will never reach the perfect calibration. Thus, we need to select test sets of experiments to use for calibration and for verification.

## 9 Experiences from the NCEP OSSE

### *9.1 Background of the NCEP OSSE*

Various types of OSSEs have been performed (see Sect. 2); however, to our knowledge, the OSSE performed at the National Centers for Environmental Prediction (NCEP) is the most extensive one so far, and one where calibrations have been performed and presented in a regular manner. The calibration of data impact has been performed by comparing the data impact with both real and simulated data. Without calibration, the simulated data impact cannot be related to the real data impact. The NCEP OSSE is also the first OSSE where radiance data from satellites were simulated and assimilated. A forecast run with a version of the ECMWF model was used to produce the Nature Run, instead of using an analysis or using the same NWP model used for the assimilation (see Sects. 1, 2 and 3).

Since the DWL is one of the most costly instruments, various simulation experiments have been funded and performed. In the NCEP OSSE, instead of evaluating a specific instrument, four representative types of DWL were evaluated (see Sect. 9.3 below for details). The results show a potentially powerful impact from DWL, but also show that without a careful design of the observing system and a significant effort in developing the data assimilation system, DWL will not be utilized to its best potential.

### *9.2 Calibration Performed for NCEP OSSE*

The calibrations were performed on existing instruments, such as the denial of RAOB (radiosonde observations) wind, RAOB temperature, and TIROS Operational Vertical Sounder (TOVS) radiances in various combinations. The geographical distribution of time-averaged Root Mean Square Error (RMSE) shows generally satisfactory agreement between real and simulated impacts. In both the real and simulated analysis, a large analysis impact in the tropics is seen to decrease in the forecast fields. In the Northern Hemisphere mid latitudes, the RMSE distribution of forecasts shows similar spatial patterns in the real and simulated analyses.

Figure 1 shows anomaly correlation (AC) skill in the 72-h 500 hPa geopotential height forecasts verified against the analysis from control experiments. The analysis

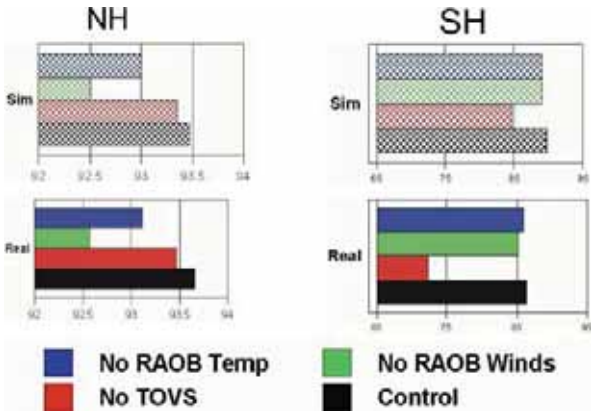



**Fig. 1** 500 hPa height anomaly correlation, time averaged between February 13 and 28. Seventy two-hour forecast fields are verified against the control analysis. Control runs include all conventional data and TOVS radiances. For each run the RAOB winds, RAOB temperatures and TOVS radiance are withdrawn in turn (experiments NWIN, NTMP, NTV, respectively). The *left two panels* are for the Northern Hemisphere and the *right two panels* for the Southern Hemisphere. The *top two panels* are for simulation experiments and the *bottom two* are for real experiments. With permission from Masutani et al. (2010)

of the control experiments (CTL) includes conventional observations and TOVS. TOVS (NTV experiment), RAOB wind (NWIN experiment), and RAOB temperature (NTMP experiment) are withdrawn, and the real and simulated data impacts are compared.

In both real and simulated experiments, the RAOB wind has the most impact: overall for the Northern Hemisphere; very slightly more than RAOB temperature for the Southern Hemisphere. Its impact and the magnitude and spatial pattern of the impact are in good agreement for real and simulated experiments. However, the effect of withholding TOVS data in the Southern Hemisphere is much greater in reality than in simulation. Note that a time-varying real SST was used in the assimilation and a constant SST in the simulation. In order to investigate the cause of this inconsistent result, eight experiments were compared: real or simulated analysis, constant or real SSTs, and with or without TOVS data. The consistency in response between the simulated and real atmosphere to the two different SSTs was confirmed. These results suggest that if the SST has a large temporal variability, the impact of TOVS data becomes more important. When TOVS data are used, the analyses with the two different SSTs become closer because TOVS data contain information about SSTs. Although using constant SSTs to generate the Nature Run is not desirable, we conclude that the data impact of slowly varying SSTs in the Southern Hemisphere can be tested with the *T213* NR (see Sect. 3.2).

These results suggest that a realistically variable SST is required for a more reliable OSSE. Ideally, an ocean-atmosphere coupled model is desirable for producing a better Nature Run, but this may require further development of current coupled models.

## 9.3 Evaluation of DWL Impact Using the NCEP OSSE

In the NCEP OSSE, instead of evaluating a specific type of the DWL instrument, four representative types of DWL are evaluated. The data impact from a specific type of DWL is expected to be estimated from the data impact of these four types of DWL. After these idealized experiments, a more realistic DWL will be simulated and evaluated. The four types of DWL are as follows:

- DWL with scanning, while sampling is from all vertical levels;
- DWL without scanning, while sampling is from all vertical levels and in only one direction;
- DWL with scanning, while sampling is from upper levels;
- DWL with scanning, while sampling is from lower levels and clouds.

Upper and lower level sampling represent DWL measurements of molecular and aerosol particle returns, respectively.

First, many experiments were done to illustrate the impact of conventional and DWL data over the first few days of the period under investigation. Then, selected sets of experiments, including model forecasts, were extended to the whole Nature Run period. The impact of DWL was assessed using AC (anomaly correlation) for 500 hPa geopotential heights; then the results of time-averaged geographical distributions and a time series of RMSE were also studied. Traditionally, the AC for 500 hPa geopotential height is used to evaluate the data impact, but it was soon evident in this study that the impact on 500 hPa geopotential height is very limited.

The meridional wind ($v$) is mainly used to assess the performance of the DWL. Note that the evolution of atmospheric phenomena at the shorter time and smaller spatial scales is dominated by the wind field, while for longer time and larger spatial scales the mass (temperature) field is dominant (Stoffelen et al. 2005; Kalnay 1985). In the Northern Hemisphere, excellent skill at the global scale is mostly achieved by existing data (conventional and TOVS). Therefore, the impact of DWL is expected at the synoptic scales. The skill to predict temperature ($T$) comes mainly from planetary scale events, while the skill to predict $v$ comes mainly from the synoptic scale. The zonal wind ($u$) and meridional wind ($v$) contain the information about relative vorticity at the synoptic scale, while $u$ and $T$ contain information about the wave guide (Hoskins and Ambrizzi 1993). Therefore, $v$ depicts information about relative vorticity. The large scale $u$ component can be inferred from temperature, $T$, observations in the extratropics, while DWL wind observations mainly define the synoptic scale wave which is represented in relative vorticity and the meridional wind, $v$.

The data impact further depends on the resolution of the DAS. There are many reasons to expect that the data impact might be reduced with higher resolution models (or better forecast models), because they can provide much better background (forecast) fields and there is less room for data to improve the analysis. On the other hand, a higher resolution model will be able to effectively utilize data in finer detail, and that may lead to a higher data impact. Moreover, the smaller scales evolve faster

than the larger scales, so their evolution needs to be analysed more often with new observations.

Masutani et al. (2006, 2010) showed that improvements in AC (anomaly correlation) scores caused by the insertion of DWL winds are less in a higher-resolution DAS than for a lower-resolution DAS because the first guess field from a higher resolution forecast model is more accurate and leaves less room for improvement. At the larger spatial scales, improvements in the model are more important, but wind data clearly improve the analysis and forecast at the smaller spatial scales. The results very much depend on the spatial scale considered. The NCEP OSSE also showed that the data impact depends on the DAS. Therefore, OSSEs performed using various DAS will be needed to establish confidence in the evaluation of future instruments.

Finally, data impact is also tested using various thinning strategies. For example, data are thinned to 10% in various ways:

- Uniformly;
- 10 min on followed by 90 min off;
- Targeted to areas with large analysis error;
- Targeted to data void areas;
- Comparison between thinning and increase in observational error.

The NCEP OSSE results show that OSSEs are a very powerful tool for assessing the effect of data distribution.

## 10 Summary and Concluding Remarks for OSSEs

Credible OSSEs may be performed that realistically evaluate the impact of prospective observations. The challenges of OSSEs, such as differences in character between the Nature Run and real atmosphere, the process of simulating data and the estimation of observational errors all affect the results. Evaluation metrics moreover affect the conclusions. Thus, consistency in results is important. Some results may be optimistic and some pessimistic. However, it is important to be able to evaluate the sources of errors and uncertainties. As more information is gathered, we can perform more credible OSSEs. If the results are inconsistent, the cause of the inconsistency needs to be investigated carefully. Only when the inconsistencies are explained, interpretation of the results becomes credible.

The NCEP OSSEs (Masutani et al. 2006, 2010) have demonstrated that carefully conducted OSSEs are able to provide useful recommendations which influence the design of future observing systems. Based on this work, OSSEs can be used to investigate:

- The effective design of orbit and configuration of an observing system;
- The effective horizontal and vertical data density;

- The evolution of data impact with forecasts;
- The balance between model improvement and improvements in data density and quality;
- The combined impacts of mass (temperature) data and wind data;
- The development of bias correction strategies.

As models improve, there is less improvement in the forecast due to the observations. Sometimes the improvement in forecasts due to model improvements can be larger than the improvement due to observations. However, even in the Northern Hemisphere, forecasts at the subsynoptic scales require much better observations. In the tropics, models need to be improved to retain the analysis improvement for more than a few days of the forecast (Žagar et al. 2008). OSSEs will be a powerful tool for providing guidelines for future development in these areas.

(i) *Value of OSSEs*: Operational centres are busy getting the best possible value out of existing instruments. We expect that carefully designed OSSEs will enable scientists to make strong and important contributions to the decision making process for future observing systems. Time will be saved in using the new data when compared to the work required to use observing systems that were built without any guidance from OSSEs. However, there is a serious dilemma in spending resources on OSSEs. If a NWP centre devotes resources to getting the greatest benefit out of existing data sources, it misses the opportunity to assess critical future observing systems, with the result that it must live with whatever new observing systems appear in the future rather than influence their development. If it devotes its resources entirely to OSSEs, it may not be paying enough attention to today's valuable data.

(ii) *Challenges of OSSEs*: OSSEs are a challenge to weather services. OSSEs require strong leaders with a clear vision, because many of the efforts offer long-term rather than short-term benefits. Although operational systems should benefit from carefully executed OSSEs through lower cost of implementation, there are immediate costs to OSSEs.

OSSEs are very labour intensive. The Nature Run has to be produced using state-of-the-art NWP models at the highest resolution. Simulating data from a Nature Run requires large computational resources, and simulations and assimilations have to be repeated with various configurations. OSSEs also require extensive knowledge of many aspects of the NWP system. Expert knowledge is also required for each instrument. Efficient collaborations are thus essential for producing timely and reliable results.

(iii) *Role of stakeholders*: OSSEs will be conducted by various scientists with different interests. Some will want to promote particular instruments. Others may want to aid in the design of the global observing system. Specific interests may introduce bias into OSSEs but they may also introduce strong motivations.

Operational centres will perform the role of finding a balance among conflicting interests to seek an actual improvement in weather predictions. They may be regarded as unbiased and thus be best placed for this role; on the other hand, difficulties in finding resources may hamper their effort.

(iv) *Recommendations*: Ideally, all new instruments should be tested by OSSEs before they are selected for construction and deployment. OSSEs will also be important in influencing the design of the instruments and the configuration of the global observing system (see chapter *The Global Observing System*, Thépaut and Andersson). While the instruments are being built, OSSEs will help prepare the DAS for the new instruments. Developing a DAS to assimilate a new type of data is a significant task. However, this effort has traditionally been made only after the data became available. The OSSE effort demands that this same work be completed earlier; this will speed up the actual use of the new data and proper testing, increasing the exploitation lifetime of an innovative satellite mission.

From the experience of performing OSSEs during recent decades, we realize that using the same Nature Run is essential for conducting OSSEs to deliver reliable results in a timely manner. The simulation of observations requires access to the complete model data and a large amount of resources; thus it is important that the simulated data from many institutes be shared among all the OSSEs. By sharing the Nature Run and simulated data, multiple participants in OSSEs will be able to produce results which can be compared; this will enhance the credibility of the results.

(v) *Final word*: NCEP's experience with OSSEs demonstrates that they often produce unexpected results. Theoretical predictions of the data impact and theoretical backup of the OSSE results are very important as they provide guidance on what to expect. On the other hand, unexpected OSSE results will stimulate further theoretical investigations. When all efforts come together, OSSEs will help with timely and reliable recommendations for future observing systems.

## References

Arnold, C.P., Jr. and C.H. Dey, 1986. Observing-systems simulation experiments: Past, present, and future. *Bull. Amer. Meteorol. Soc*., **67**, 687–695.

Atlas, R. 1997. Atmospheric observation and experiments to assess their usefulness in data assimilation. *J. Meteor. Soc. Jpn*., **75**, 111–130.

Atlas, R. and G.D. Emmitt, 1995. Simulation studies of the impact of space-based wind profiles on global climate studies. *Proceedings American Meteorological Society's 6th Symposium on Global Change Studies*, January, Dallas, TX.

Atlas, R., G.D. Emmitt, J. Terry, E. Brin, J. Ardizzone, J.C. Jusem and D. Bungato, 2003a. Recent observing system simulation experiments at the NASA DAO. Preprints, *7th Symposium on Integrated Observing Systems*, 9–13 February 2003, Long Beach, California, American Meteorological Society.

Atlas, R., G.D. Emmitt, J. Terry, E. Brin, J. Ardizzone, J.C. Jusem and D. Bungato, 2003b. OSSEs to determine the requirements for space-based lidar winds for weather prediction. *SPIE's Laser Radar Technology and Applications VIII Conference*, April, Orlando, FL.

Atlas, R., E. Kalnay, J. Susskind, W.E. Baker and M. Halem, 1985. Simulation studies of the impact of future observing systems on weather prediction. *Proceedings of the 7th Conference on NWP*, pp145–151.

Becker, B.D., H. Roquet and A. Stoffelen 1996. A simulated future atmospheric observation database including ATOVS, ASCAT, and DWL. *Bull. Amer. Meteorol. Soc.*, **10**, 2279–2294.

Cress, A. and Wergen, W. 2001. Impact of profile observations on the German Weather Service's NWP system. *Meteor. Zeitschrift*, **10**, 91–101.

Emmitt, G.D. and S.A. Wood, 1989. Simulation of a space-based Doppler lidar wind sounder – sampling errors in the vicinity of wind and aerosol inhomogeneities. *Fifth Conference on Coherent Laser Radar*, June, Munich, Federal Republic of Germany.

Emmitt, G.D. and S.A. Wood, 1991a. Simulating thin cirrus clouds in observing system simulations experiments (OSSE) for LAWS. *Proceedings American Meteorological Society's 7th Symposium on Meteorological Observations and Instrumentation, Special Session on Laser Atmospheric Studies*, January 14–18, New Orleans, LA, pp460–462.

Emmitt, G.D. and S.A. Wood, 1991b. Simulated wind measurements with a low power/high PRF space-based Doppler lidar. *Optical Remote Sensing of the Atmosphere, 5th Topical Meeting*, November 18–21, Williamsburg, VA.

Emmitt, G.D. and S.A. Wood, 2001. Simulating space-based lidar performance using global and regional scale atmospheric numerical models. *Optical Remote Sensing Topical Meeting*, February, Coeur d'Alene, ID.

Errico, R.M., R. Yang, M. Masutani and J. Woollen, 2007. Estimation of some characteristics of analysis error inferred from an observation system simulation experiment. *Meteor. Zeitschrift*, **16**, 695–708.

Fisher, M., 2003. Background error covariance modelling. *Proceedings of ECMWF Seminar, Recent Developments in Data Assimilation for Atmosphere and Ocean*, 8–12 September 2003, Reading, UK, pp45–64.

Gelaro, R. and Y. Zhu, 2009. Examination of observation impacts derived from observing system experiments (OSEs) and adjoint models. *Tellus*, **61A**, 179–193.

Halem, M. and R. Dlouhy, 1984. Observing system simulation experiments related to space-borne lidar wind profiling. Part 1: Forecast impact of highly idealized observing systems. Preprints, *Conference on Satellite Meteorology/Remote Sensing and Applications*, June 25–29, 1984, American Meteorological Society, Clearwater, FL, pp272–279.

Hamill, T.M., R.P. d'Entrement and J.T. Bunting, 1992. A description of the air force real-time, nephanalysis model. *Weather Forecasting*, **7**, 288–306.

Han, Y., P. van Delst, Q. Liu, F. Weng, B. Yan, R. Treadon and J. Derber, 2006. JCSDA Community Radiative Transfer Model (CRTM) – Version 1, *NOAA Tech Report 122*.

Henderson-Sellers, A., 1986. Layer cloud amount for January and July 1979 from 3D-Nephanalysis. *J. Climate Appl. Meteor.*, **24**, 118–132.

Hoskins, B.J. and T. Ambrizzi, 1993. Rossby wave propagation on a realistic longitudinally varying flow . *J. Atmos. Sci*., **50,** 1661–1671.

Kalnay, E, J.C. Jusem and J. Pfaendtner, 1985. The relative importance of mass and wind data in the FGGE observing system. *Proceedings of the NASA Symposium on Global Wind Measurements*, Columbia, MD, NASA, 1–5.

Keil, M., 2004. Assimilating data from a simulated global constellation of stratospheric balloons. *Q. J. R. Meteorol. Soc*., **130**, 2475–2493.

Kistler, R., NCEP Staff, Contractors and Visiting Scientists, Past and Present, 2008. Reanalysis at NCEP: Past, Present and Future. *Third WCRP International Conference on Reanalysis*, February, 2008, Tokyo, Japan.

Kleespies, T.J. and D. Crosby 2001. Correlated noise modelling for satellite radiance simulation. *AMS preprint volume for the 11th Conference on Satellite Meteorology and Oceanography*, October 2001, Madison Wisconsin, pp604–605.

Kleespies, T.J., P. van Delst, L.M. McMillin and J. Derber, 2004. Atmospheric Transmittance of an Absorbing Gas. 6. An OPTRAN Status Report and Introduction to the NESDIS/NCEP Community Radiative Transfer Model. *Appl. Opt.*, **43**, 3103–3109.

Lahoz, W.A., R. Brugge, D.R. Jackson, S. Migliorini, R. Swinbank, D. Lary and A. Lee 2005. An observing system simulation experiment to evaluate the scientific merit of wind and ozone measurements from the future SWIFT instrument. *Q. J. R. Meteorol. Soc*., **131**, 503–523.

Langland, R.H. and N.L. Baker, 2004. Estimation of observation impact using the NRL atmospheric variational data assimilation adjoint system. *Tellus*, **56A**, 189–203.

Liu, Q. and F. Weng, 2006a. Radiance assimilation in studying Hurricane Katrina. *Geophys. Res. Lett.*, **33**, L22811, doi:10.1029/2006GL027543.

Liu, Q. and F. Weng, 2006b. Detecting warm core of Hurricane from the special sensor microwave imager sounder. *Geophys. Res. Lett*., **33**, L06817.

Lord, S.J., E. Kalnay, R. Daley, G.D. Emmitt and R. Atlas 1997. Using OSSEs in the design of the future generation of integrated observing systems. Preprints, *1st Symposium on Integrated Observing Systems*, 2–7 February 1997, American Meteorological Society, Long Beach, CA.

Lord, S.J., M. Masutani, J.S. Woollen, J.C. Derber, G.D. Emmitt, S.A. Wood, S. Greco, R. Atlas, J. Terry and T.J. Kleespies, 2002. Impact assessment of a Doppler wind lidar for NPOESS/OSSE. *American Meteorological Society's 6th Symposium on Integrated Observing Systems*, January, Orlando, FL.

Lorenc, A.C., 1986. Analysis methods for numerical weather prediction. *Q. J. R. Meteorol. Soc.*, **112**, 1177–1194.

Marseille, G.J. and A. Stoffelen, 2003. Simulation of wind profiles from a space-borne Doppler wind lidar. *Q. J. R. Meteorol. Soc.*, **129**, 3079–3098.

Marseille, G.J., A. Stoffelen and J. Barkmeijer, 2008a. Sensitivity Observing System Experiment (SOSE) – A new effective NWP-based tool in designing the global observing system. *Tellus A*, **60**, 216–233, doi:10.1111/j.1600-0870.2007.00288.x.

Marseille, G.J., A. Stoffelen and J. Barkmeijer, 2008b. Impact assessment of prospective space-borne Doppler wind lidar observation scenarios. *Tellus A*, **60**, 234–248, doi:10.1111/j.1600-0870.2007.00289.x.

Marseille, G.J., A. Stoffelen and J. Barkmeijer, 2008c. A cycled sensitivity observing system experiment on simulated Doppler wind lidar data during the 1999 Christmas storm "Martin". *Tellus A*, **60**, 249–260, doi:10.1111/j.1600-0870.2007.00290.x.

Masutani, M., K. Campana, S. Lord and S.-K. Yang, 1999. Note on cloud cover of the ECMWF nature run used for OSSE/NPOESS project. *NCEP Office Note* No. 427.

Masutani, M., J.S. Woollen, S.J. Lord, G.D. Emmitt, T.J. Kleespies, S.A. Wood, S. Greco, H. Sun, J. Terry, V. Kapoor, R. Treadon and K.A. Campana, 2010. Observing system simulation experiments at the national centers for environmental prediction. *J. Geophys. Res*., **115**, doi:10.1029/2009JD012528.

Masutani, M., J.S. Woollen, S.J. Lord, G.D. Emmitt, S. Wood, S. Greco, T.J. Kleespies, H. Sun, J. Terry, J.C. Derber, R.E. Kistler, R.M. Atlas, M.D. Goldberg and W. Wolf, 2003. Observing system simulation experiments for NPOESS – assessment of Doppler wind lidar and AIRS. *American Meteorological Society's The Simpson Symposium*, February, Long Beach, CA.

Masutani, M., J.S. Woollen, S.J. Lord, T.J. Kleespies, G.D. Emmitt, H. Sun, S.A. Wood, S. Greco, J. Terry and K. Campana, 2006. Observing System Simulation Experiments at NCEP. *NCEP Office Note* No. 451.

McNally, A.P., J.C. Derber, W.-S. Wu and B.B. Katz, 2000. The use of TOVS level-1 radiances in the NCEP SSI analysis system. *Q. J. R. Meteorol. Soc*., **129**, 689–724.

Nitta, T., 1975. Some analyses of observing systems simulation experiments in relation to First GARP Global Experiment. *GARP Working Group on Numerical Experimentation, Report No. 10*, US GARP Plan, pp1–35. [Available from the National Academy of Sciences, 2101 Constitution Ave. N.W., Washington, DC 20418.]

Petersen, D.P., 1968. On the concept and implementation of sequential analysis for linear random fields. *Tellus*, **20**, 673–686.

Reale, O., J. Terry, M. Masutani, E. Andersson, L.P. Riishøjgaard and J.C. Jusem, 2007. Preliminary evaluation of the European Centre for Medium-Range Weather Forecast (ECMWF)

nature run over the tropical Atlantic and African monsoon region. *Geophys. Res. Lett*., **34**, L22810, doi:10.1029/2007GL031640.

Riishøjgaard, L.P., R. Atlas and G.D. Emmitt, 2003. Analysis of simulated observations from a Doppler wind lidar. *American Meteorological Society's 12th Conference on Satellite Meteorology*, February, Long Beach, CA.

Rohaly, G.D. and T.N. Krishnamurti, 1993. An observing system simulation experiment for the laser atmospheric wind sounder (LAWS). *J. Appl. Meteor.,* **32**, 1453–1471.

Saha, S., S. Nadiga, C. Thiaw, J. Wang, W. Wang, Q. Zhang, H.M. Van den Dool, H.-L. Pan, S. Moorthi, D. Behringer, D. Stokes, M. Peña, S. Lord, G. White, W. Ebisuzaki, P. Peng and P. Xie, 2006. The NCEP climate forecast system. *J. Climate*, **16**, 3483–3517.

Saunders R.W., M. Matricardi and P. Brunel, 1999. An improved fast radiative transfer model for assimilation of satellite radiance observations. *Q. J. R. Meteorol. Soc*., **125**, 1407–1425.

Stoffelen, A., G.J. Marseille, F. Bouttier, D. Vasiljevic, S. De Haan and C. Cardinali, 2006. ADM-Aeolus Doppler wind lidar observing system simulation experiment. *Q. J. R. Meteorol. Soc*., **619**, 1927–1948.

Stoffelen, A., J. Pailleux, E. Källén, J.M. Vaughan, L. Isaksen, P. Flamant, W. Wergen, E. Andersson, H. Schyberg, A. Culoma, R. Meynart, M. Endemann and P. Ingmann, 2005. The atmospheric dynamic mission for global wind fields measurement. *Bull. Amer. Meteorol. Soc.*, **86**, 73–87.

Strow, L.L., H.E. Motteler, R.G. Benson, S.E. Hannon and S. De Souza-Machado, 1998. Fast computation of monochromatic infrared atmospheric transmittances using compressed lookup tables. *J. Quant. Spectrosc. Radiat. Transfer*, **59**, 481–493.

Tan, D.G.H., E. Andersson, M. Fisher and L. Isaksen 2007. Observing system impact assessment using a data assimilation ensemble technique: Application to the ADM-Aeolus wind profiling mission. *Q. J. R. Meteorol. Soc.*, **133**, 381–390.

Tsutsui, J. and A. Kasahara 1996. Simulated tropical cyclones using the National enter for Atmospheric Research community climate model. *J. Geophys. Res.*, **101**, 15013–15032.

Weng, F., 2007. Advances in radiative transfer modelling in support of satellite data assimilation. *J. Atmos. Sci*., **64**, 3803–3811.

Wood, S.A., G.D. Emmitt and S. Greco, 2000. DLSM. A coherent and direct detection lidar simulation model for simulating space-based and aircraft-based lidar winds. *AeroSense 2000*, April, Orlando, FL.

Woollen, J.S., M. Masutani, H. Sun, Y. Song, G.D. Emmitt, Z. Toth, S.J. Lord and Y. Xie 2008. Observing systems simulation experiments at NCEP OSSEs for realistic adaptive targeted DWL Uniform observation and AIRS. *AMS preprint, Symposium on Recent Developments in Atmospheric Applications of Radar and Lidar*, New Orleans, LA, pp 20–24, January 2008.

Žagar, N., A. Stoffelen, G.J. Marseille, C. Accadia and P. Schlüssel, 2008. Impact assessment of simulated doppler wind lidars with a multivariate variational assimilation in the Tropics. *Mon. Weather Rev.,* **136**, 2443–2460, doi:10.1175/2007MWR2335.1.

Zhu, Y. and R. Gelaro, 2008. Observation sensitivity calculations using the adjoint of the Gridpoint Statistical Interpolation (GSI) analysis system. *Mon. Weather Rev.*, **136**, 335–351.

# Data Assimilation for Other Planets

**Stephen R. Lewis**

## 1 Introduction

The application of data assimilation methodology to terrestrial problems in meteorology, atmospheric physics and physical oceanography has already been described extensively within this book. Data assimilation, the combination of observations and numerical models which provide physical constraints, organize and propagate the observational information which is introduced, also offers significant potential advantages for the analysis of atmospheric data from other planets. The Solar System provides seven examples of thick neutral atmospheres in addition to that of the Earth: Mars, Venus and Saturn's moon Titan, which all have relatively large rocky cores surrounded by thinner atmospheres, like the Earth, and four largely gaseous Giant Planets, Jupiter, Saturn, Uranus and Neptune. In recent years satellites have been placed in orbit about Mars in particular, but also Venus, Jupiter and Saturn, in contrast to the relatively rapid fly-by missions in the initial stages of the exploration of the Solar System. These spacecraft provide the potential for long sequences of atmospheric observations. Together with the necessary advances in numerical modelling of planetary atmospheres, these new missions have provided an opportunity for the application of data assimilation techniques for the analysis of planetary observations. As described in this chapter, data assimilation has now been employed with some success in the context of the atmosphere of Mars and more ambitious studies are planned for the future. Assimilation in these unfamiliar and, compared to Earth, data-poor environments also provides valuable lessons for the development of terrestrial assimilation, especially in situations where it is vital to extract the maximum information from a limited observational record.

S.R. Lewis (✉)
Department of Physics and Astronomy, The Open University, Milton Keynes MK7 6AA, UK
e-mail: S.R.Lewis@open.ac.uk

W. Lahoz et al. (eds.), *Data Assimilation*, DOI 10.1007/978-3-540-74703-1_25,

## 2 Motivation for the Assimilation of Extra-Terrestrial Data

The motivation behind the application of data assimilation to atmospheric data from other planets is in principle very similar to the motivation for its use on the Earth. Typical spacecraft observations of radiances at various wavelengths, most commonly in the infrared, must be interpreted and used to constrain the thermodynamic and dynamic state of the atmosphere under observation in a systematic way. In the past, this has typically been done by the retrieval of individual temperature profiles, for example, and by mapping and interpolation in space and time of either observed radiances (or brightness temperatures) or sets of individual retrieved profiles to obtain global fields. Simple balance relationships, such as the gradient wind approximation, have been used to derive estimates of further quantities such as zonal winds from longitudinally-averaged temperatures. Such straightforward procedures are justified in the case of relatively sparse observations of a planetary atmosphere which may be much less well understood than that of the Earth. As seen earlier in the book, in recent years the terrestrial meteorological and oceanographic community have benefited greatly from the application of more sophisticated data assimilation techniques to the relatively large number of observations available to them (see chapters in Part II, *Observations*). It is natural that planetary scientists would propose similar analyses to maximize the valuable information that can be extracted from the relatively smaller data sets that are available to them, as was done by several teams for Mars in the 1990s (e.g. Banfield et al. 1995; Lewis and Read 1995; Lewis et al. 1996, 1997; Houben 1999; Kass 1999; Zhang et al. 2001).

Planetary scientists do not yet have the strong motivation provided by the regular requirement to provide initial states for near-future weather forecasts, which has provided much of the impetus behind the development of data assimilation techniques for Earth (see chapter *Numerical Weather Prediction*, Swinbank). As a consequence, the resources available for planetary modelling and data assimilation are much smaller and to date schemes have generally been developed by only a handful of individuals and small teams of researchers. Aside from this practical limitation, the atmospheres of other planets are simply much less well-understood than that of the Earth and in many cases no sufficiently realistic general circulation model (GCM) exists which may be constrained by observations. Observational and model error characteristics, error growth and inherent biases have all received very limited study, if they have been considered in the literature at all.

Data assimilation does, however, offer many potential benefits to planetary science, not least in offering the prospect of a systematic reanalysis of past and present spacecraft data. By using a physically self-consistent atmospheric model, data assimilation is also able to extract information about variables not directly observed, for example to provide a self-consistent set of global temperatures, winds and surface pressure even where only one or two of these atmospheric fields may be observed, or, more likely, the observations are in the form of radiances which require inversion to derive temperatures. In these cases, assimilation effectively offers a good "first guess" in the form of a model forecast of the atmospheric state which might be used with a forward model to predict radiances, or as the basis for a conventional atmospheric inversion.

As in terrestrial atmospheric science and physical oceanography, at the same time data assimilation provides a systematic method of testing and validating models, for example by the identification of regions or fields where the model predicts a consistent misfit with the observations (see chapter *The Role of the Model in the Data Assimilation System*, Rood). This is of particular value in planetary science where models are often at a quite early stage of development and it is not necessarily the case that experience of the Earth will carry over directly. Data assimilation also permits the intercomparison of observations made of different fields, or at different time and places by separate instruments, permitting the extraction of the maximum information by combining two different data sets in an objective way.

Having noted the use of assimilation for improving models, it is important that while the maximum information is extracted from the valuable and limited observational record for another planet, at the same time this record is not over-used. For example, a set of observations could be used to improve the model itself or to estimate the model state, but not both in a recursive fashion. It is possible that uncertain model parameters can be included formally in the model state and that both may be estimated at once. A practice in terrestrial numerical weather prediction is to accumulate records of model output statistics and to perform a linear regression between the prediction and subsequent verification as a means of improving the model based on very large numbers of observations (e.g. Kalnay 2003). This may not yet be possible for other planets, owing to the more limited observational record.

Although short-term weather forecasting for the near-surface meteorology of another planet is still a distant prospect, forecasts of some atmospheric properties and, perhaps most importantly, their likely variance are vital now for spacecraft and instrument design and planning. Uses include predictions of upper atmosphere density for satellite *aerobraking* and *aerocapture* (this is the use of the atmospheric friction around 100 km altitude and above to decelerate spacecraft to aid their capture into low planetary orbits), entry, descent and landing studies for atmospheric entry vehicles, and estimates of the range of surface conditions which will be experienced in the lifetime of landed spacecraft. Such forecasts are often made on the basis of past experience and climatology, but for other planets the latter can be unknown or involve unwarranted extrapolations from previous mission data relevant to different locations and times of year. Models are starting to be used as the basis for generating more comprehensive climatologies for Mars (Lewis et al. 1999; Justus et al. 2002), in particular for regions of the atmosphere, or under conditions which have not yet been observed in detail. Data assimilation will play an increasingly important role here as the means of constraining and improving these models at times when some observations are available.

## 3 Data Assimilation for the Atmosphere of Mars

The atmosphere of Mars is the most obvious first extra-terrestrial target for data assimilation, motivated both by its similarities to the atmosphere of the Earth and by the regular launch of spacecraft missions over the last decade, resulting in an

increased observational data set and an increased need to better understand the atmosphere for mission operations, in particular for aerobraking, aerocapture and entry descent and landing.

Like the Earth, Mars is a largely solid planet with a radius of 3,389 km, surface gravity of 3.72 $ms^{-2}$ and a solar day (*sol*) of 88,775 s, around 40 min longer than the day on Earth. The rotation axis of Mars is tilted at a similar angle to the plane of the ecliptic, 25.2° compared to 23.5° for Earth, and so Mars experiences a similar pattern of seasons over the year of 668.6 sols, almost twice as long as a year on Earth. The atmosphere is also largely transparent, but is composed of 95% carbon dioxide with a typical surface pressure of 610 Pa (the typical surface pressure on Earth is 101,300 Pa, or 1,013 hPa). Temperatures can reach above the freezing point of water on a warm, summer's afternoon, but can also fall to 145 K in polar night, at which point carbon dioxide freezes out around the Winter Pole forming a large seasonal ice cap containing up to a third of the total mass of the atmosphere. Despite the differences in atmospheric composition and mass, the atmospheric pressure scale height is only a little larger (roughly 10 km compared to 7.5 km on Earth) and the horizontal deformation radius is about 1,000 km in both cases; the lower gravity on Mars compared to Earth is compensated by the lower specific gas constant for the carbon dioxide rich atmosphere, resulting in a rather similar static stability for the lower atmosphere on both planets.

Transient, baroclinic weather systems are observed in martian mid latitudes, especially in the Northern Hemisphere (Barnes 1981, 1980; Collins et al. 1996; Wilson et al. 2002; Banfield et al. 2004), on a similar scale to those seen on Earth but with typically one to four high and low pressure systems around a latitude circle owing to the smaller planetary radius. Intriguingly, these travelling waves on Mars appear to be much more regular, and sometimes almost periodic, than typical terrestrial mid latitude weather systems (Barnes 1980, 1981; Collins et al. 1996; Read and Lewis 2004).

The similarities of martian atmospheric dynamics to that of the Earth have led to the development of several Mars GCMs from the late 1960s onwards, typically derived from terrestrial models (for reviews see, e.g., Zurek et al. 1992; Lewis 2003; Read and Lewis 2004 and references therein). The most advanced of these models are comparable in complexity with a terrestrial global model used for numerical weather prediction or for climate studies (see chapters *The Role of the Model in the Data Assimilation System*, Rood; *Reanalysis: Data Assimilation for Scientific Investigation of Climate*, Rood and Bosilovich).

Despite its similarities with the atmosphere of the Earth, at least two factors make that of Mars different from the perspective of data assimilation. Firstly, the lower atmospheric density, and hence lower heat capacity, on Mars means that the atmosphere responds very much more quickly to changes in radiative forcing. This is particularly true at times when the atmosphere of Mars contains large amounts of suspended dust, which absorbs visible radiation and heats the atmosphere. A typical radiative relaxation time scale for the lower martian atmosphere is around two sols (Goody and Belton 1967; Gierasch and Goody 1967, 1968), and may be as low as one sol when the atmosphere is dusty, an order of magnitude shorter than

radiative relaxation times for the Earth's atmosphere. This means that a Mars GCM will respond very quickly to its own radiative forcing scheme and, if this is not precisely correct together with an accurate spatial and temporal dust distribution, the GCM may rapidly "forget" information introduced by assimilation of past data where there is an absence of current observations. It should be noted that there are considerable uncertainties in the radiative properties and size distribution of martian dust and consequently dust heating parametrizations in GCMs are likely to be subject to substantial errors.

Secondly, the observation that errors grow roughly exponentially with time in accordance with deterministic chaos theory on Earth (e.g. Ehrendorfer 1997; Toth 2001) may not necessarily be true on Mars, at least at some times of year when model simulations indicate that error growth can decay with time and the atmosphere appears highly predictable (Newman et al. 2004). The implications of this potentially greater predictability on Mars are yet to be fully explored.

In addition to a realistic numerical model, data assimilation requires a stream of observations and several early assimilation efforts for the martian atmosphere were motivated by the launch in 1992 of the ill-fated NASA Mars Observer (MO) spacecraft (Cunningham et al. 1992), lost around the time of orbital insertion in 1993. Like several subsequent NASA missions, MO was intended for a two-hourly sun-synchronous low polar orbit, passing over the Equator at 2:00 am and 2:00 pm local time, and it was this regular, repetitive mapping of the atmosphere that made data assimilation for Mars an attractive option. MO was followed by Mars Global Surveyor (MGS), launched in 1996, which re-flew some of the MO instruments, including notably for atmospheric observations the Thermal Emission Spectrometer (TES) (Christensen et al. 1992), which is an infrared sounder operating mainly in nadir mode, though with some limb observations. TES has produced a spectacular dataset covering almost three complete martian years from 1999 to 2004 and is the subject of several current data assimilation studies. TES nadir soundings typically allow the retrieval of temperature profiles between the surface and about 40 km with a vertical resolution of one scale height (10 km) or greater and total column opacities of dust and water ice (Conrath et al. 2000, 2002; Smith et al. 2000, 2001; Smith 2004), as well as various surface properties.

A second instrument from MO, the limb-sounding Pressure Modulator InfraRed Radiometer (McCleese et al. 1992) was re-flown on a second unsuccessful mission, Mars Climate Orbiter in 1998, but a new version of the limb-sounding radiometer, Mars Climate Sounder (MCS) (McCleese et al. 2007) is presently in orbit about Mars aboard the Mars Reconnaissance Orbiter. MCS has been mapping the martian atmosphere over at least one seasonal cycle. The principal advantages MCS will offer over TES for atmospheric assimilation are routine limb-sounding, with coverage up to about 80 km and half-scale height, 5 km, vertical resolution, with the ability to differentiate between dust, condensates and water vapour and to profile each in the vertical. Several groups are preparing to assimilate MCS data in the coming years.

It should also be noted that two other Mars spacecraft, NASA's 2001 Mars Odyssey (Saunders et al. 2004) and ESA's 2003 Mars Express (Schmidt 2003) have

both provided fascinating remote sensing observations revealing much about the martian climate and surface, but to date observations from either mission have not been assimilated into a Mars GCM.

In direct contrast to Earth, at the time these orbital spacecraft have been operating there have been few, if any, surface-based in situ meteorological observations, with the notable exceptions of the instruments on the NASA Phoenix polar lander, operating 5 months in 2008, and the Mini-TES instruments on the Mars Exploration Rovers (Smith et al. 2006) which provide lower atmosphere profiles up to a few km in height. Crucially, there have been no systematic surface pressure measurements other than from Phoenix, from Mars Pathfinder or a few months in 1996 (Schofield et al. 1997) and the longer, multi-annual record from the Viking Landers in the late 1970s (Hess et al. 1980). This lack of pressure data makes it difficult to constrain the mass budget of the Mars GCMs and brings novel difficulties in data assimilation compared to the Earth, where surface pressure is a fundamental observation to be included in any meteorological analysis. Indeed, it might be argued that surface pressure is the most important observable quantity for constraining the whole troposphere in terrestrial atmospheric data assimilation (e.g. Anderson et al. 2005) and this emphasizes the need for more martian surface observations in future, although it is highly unlikely that there will be a sufficiently dense network of surface stations on Mars to constrain a model on their own in the foreseeable future.

## *3.1 Data Assimilation Schemes for Mars*

Two approaches have been taken in developing data assimilation schemes to work with martian observations. On one hand, new schemes have been developed tailored specifically to exploit the characteristics of the data which are expected; normally remotely sensed temperature profiles from a regular two-hourly, polar orbit. On the other hand, terrestrial schemes with heritage in the numerical weather prediction community have been adapted and re-tuned for martian conditions.

Banfield et al. (1995) exploited the repetitive nature of the polar orbit to propose a variant on the full sequential Kalman filter (Kalman 1960), which is made computationally economic by only calculating the gain matrix once, and then holding it steady in time (For a discussion on the Kalman filter see chapter *Mathematical Concepts of Data Assimilation*, Nichols). The steady-state gains are computed once at the start of each assimilation experiment by an iterative technique and then applied throughout to each observation, making the gains a function of relative longitude between observation and model points. This was shown to work well in a highly idealized, single-layer primitive equation model, observing the mass field but not the velocity field.

The steady-state Kalman filter was later applied to TES mapping phase data (Zhang et al. 2001) using the NASA Ames Mars GCM (Pollack et al. 1990; Haberle et al. 1993). This study only assimilated ten sols of MGS mapping phase TES data, but was able to show a small improvement in the agreement between model and

observations; although the success was only limited and the model response was degraded in south Polar Regions. There also seemed to be little evidence that the assimilation was converging sufficiently well to capture the transient waves. The problems experienced were attributed to problems with the assumed dust opacity and distribution in the GCM.

An alternative approach was taken by Houben (1999), who employed a Mars GCM with relatively low resolution, a spectral model with 17 Legendre modes in latitude, 7 waves in longitude and 16 vertical levels, and with highly simplified physical parametrizations including linear Newtonian cooling in place of a full radiation scheme. Houben was thus able to reduce the complexity of the model so that it could be constrained with the number of MGS observations available in one sol. Assimilation was accomplished by a four-dimensional variational technique, 4D-Var (Talagrand and Courtier 1987; Courtier and Talagrand 1987) (For a discussion on 4D-Var see chapter *Variational Assimilation*, Talagrand.). This study is notably complementary to other assimilation techniques which employ a full Mars GCM and assimilate data using a more empirical technique.

Kass (1999) used the NASA Ames Mars GCM, but with a form of assimilation based on optimal interpolation, OI (Bengtsson and Gustafsson 1971; Rutherford 1972). Kass assimilated TES temperature profiles over a 25-sol, 17-orbit period during MGS aerobraking using optimal interpolation. He found that the Winter Hemisphere jet was moved polewards and that the amplitude of waves became stronger compared to an independent experiment with the Mars GCM. He was also able to demonstrate that the transient component of the surface pressure field was modified in response to the assimilation of temperature data, in an interesting contrast to the terrestrial meteorological experience which suggests that surface pressure observations are crucial and tend to drive behaviour in the atmosphere above.

Another scheme which draws heavily on terrestrial experience with some success was developed by Lewis et al. (Lewis and Read 1995; Lewis et al. 1996, 1997) based closely on the analysis correction scheme (Lorenc et al. 1991), in operational use at the Meteorological Office (UK) at the time. This scheme is a form of the successive corrections method which has proved simple and robust in many trial studies with artificial data under martian conditions. Observations are spread in both space and time by the use of empirically-tuned functions and the relatively inexpensive data assimilation scheme is paired with a fully comprehensive Mars GCM (Forget et al. 1999; Lewis et al. 1999). Assimilation of the TES data using this technique during the MGS aerobraking hiatus has been described (Lewis et al. 2007), as has an analysis of the thermal tidal behaviour throughout the MGS mapping phase (Lewis and Barker 2005). The mapping phase assimilation has been validated by a cross-comparison of model temperature profiles sampled at the same time and place as profiles obtained by radio occultation, also using the MGS spacecraft (Montabone et al. 2006a). Focused studies have included investigations of martian dust storms (Montabone et al. 2005) and detailed reconstructions of the atmosphere at the time of recent entry probes (Montabone et al. 2006b). The results of this assimilation procedure are further validated by comparing the planetary waves in the assimilation

with those from direct synoptic mapping analyses of the TES retrievals. The output from this assimilation over the full three martian years observed by MGS/TES is being made freely available and is the subject of many ongoing atmospheric investigations. For example, Wilson et al. (2008) use the statistical differences between the output of the assimilation and independent model experiments to infer potential longitudinal-mean errors in the model radiation budgets, in this case ascribed to the absence of equatorial water ice clouds in the independent model experiments.

With new MCS data in prospect and with the vast TES data set not yet fully exploited, new martian data assimilation schemes are also under development. These include applications of both variational and Ensemble Kalman filter schemes, which are now widely used at leading terrestrial data assimilation centres (e.g. Rabier 2005; Houtekamer and Mitchell 2005) – see also chapter *Mathematical Concepts of Data Assimilation* (Nichols). Research is also ongoing into direct assimilation of observed radiances rather than using pre-retrieved temperature profiles as most of the martian studies described above have done. Removing the need for a separate retrieval and linking the observed infrared radiance directly to the atmospheric state has many attractions, but is a challenging prospect particularly with a limb-sounding radiometer such as MCS. Chapters *Assimilation of Operational Data* (Andersson and Thépaut) and *Constituent Assimilation* (Lahoz and Errera) discuss the direct assimilation of radiances from operational and research satellites on Earth.

## *3.2 Results from Martian Data Assimilation*

Some early results from martian data assimilation are illustrated in this section based on the assimilation of TES observations throughout the aerobraking hiatus and scientific and extended mapping phases, a period of almost three martian years, using the analysis correction scheme of Lewis et al. (2007). The full assimilation period is summarized by Fig. 1, which shows the assimilated dust optical depth in the visible, averaged over all longitudes and converted to an equivalent optical depth at a standard reference pressure of 610 Pa. The Mars Years (*MY*) are numbered here following an arbitrary scheme (following Clancy et al. 1995), and the time of year is indicated by *areocentric longitude*, $L_S$, an angle varying from 0° to 360°, where $L_S = 0°$ is spring equinox, $L_S = 90°$ is summer solstice, $L_S = 180°$ is autumn equinox and $L_S = 270°$ is winter solstice (seasons are for the Northern Hemisphere of Mars). Figure 1 includes the aerobraking hiatus period (*MY23*, $L_S = 190°$–$260°$), during which time the spacecraft orbital period was being reduced from 45 to 24 h and the configuration was more difficult for atmospheric assimilation owing to the long orbital period and irregular and intermittent observational coverage. The subsequent 2-h scientific mapping phase orbit provided more regular observations throughout almost three Martian years of operation (*MY24*, $L_S = 141°$ to *MY27*, $L_S = 72°$).

The impact of the assimilation on the zonal mean state of the atmosphere, even in the least optimal aerobraking hiatus period is illustrated by the zonal-mean temperature and zonal winds in the assimilation during the *Noachis* dust storm period,

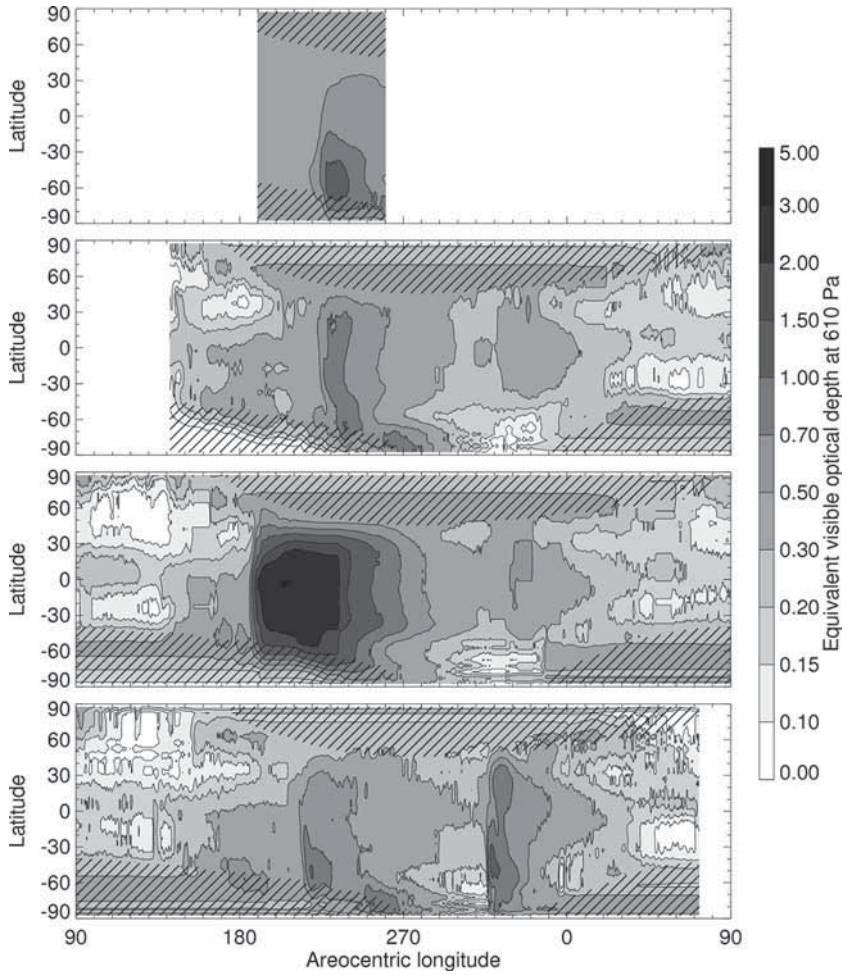



**Fig. 1** Assimilated dust optical depth at 610 Pa from the full period for which TES observations are available. Each panel shows one martian year, Mars Years (*MY*) 23–26 from *upper panel down*, with summer solstice at the *left-hand edge* (aerocentric longitude = 90°). The hatched regions indicate where there are few, if any, total opacity observations (the mean surface temperature is below 160 K and there is insufficient thermal contrast between the atmosphere and surface to retrieve total atmospheric opacities)

a regional, moderate dust storm that began around *MY23*, $L_S = 225°$ in the martian Southern Hemisphere. Figure 2 shows the zonal mean state in the assimilation and Fig. 3 the differences between this and a model run with a dust state which is a close match to the mean conditions before the dust storm. Enhanced warming throughout the middle atmosphere at most latitudes is apparent, as is a strong polar warming above the North Pole, and enhancement of the polar westerly jet, thanks to

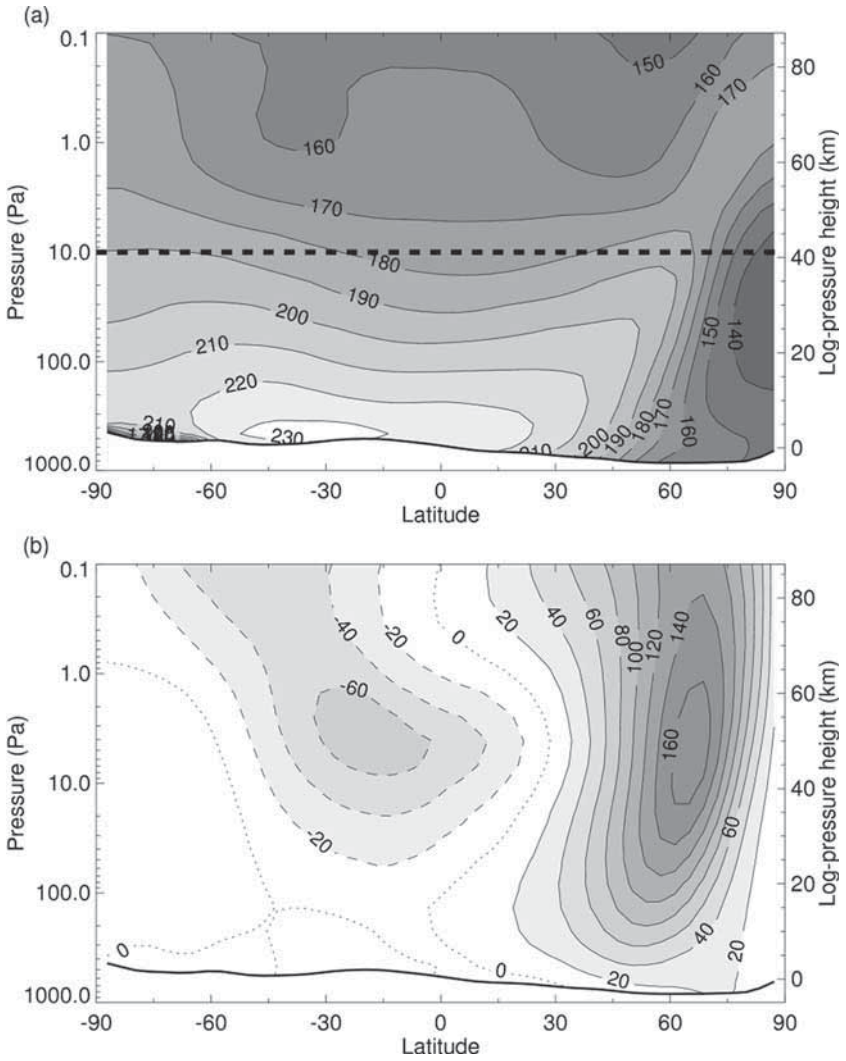



**Fig. 2** (**a**) Zonal-mean temperature and (**b**) zonal-mean zonal wind, time-averaged over the period $L_S = 225°–233°$. The *horizontal*, *dashed line* indicates the approximate level above which no temperature data from the nadir soundings was available. The zero contour is *dotted* and negative contours *dashed*. Log-pressure height is defined as $-10 \log(p/610\ \text{Pa})$ km as an approximate conversion from pressure $p$ to height above the 610 Pa level. Reprinted from Lewis et al. (2007), with permission from Elsevier

the enhanced meridional circulation in the model; no observations were available in this region at this time.

One major motivation for using assimilation techniques is in order to investigate the *transient wave* behaviour on Mars, which is difficult to interpret when

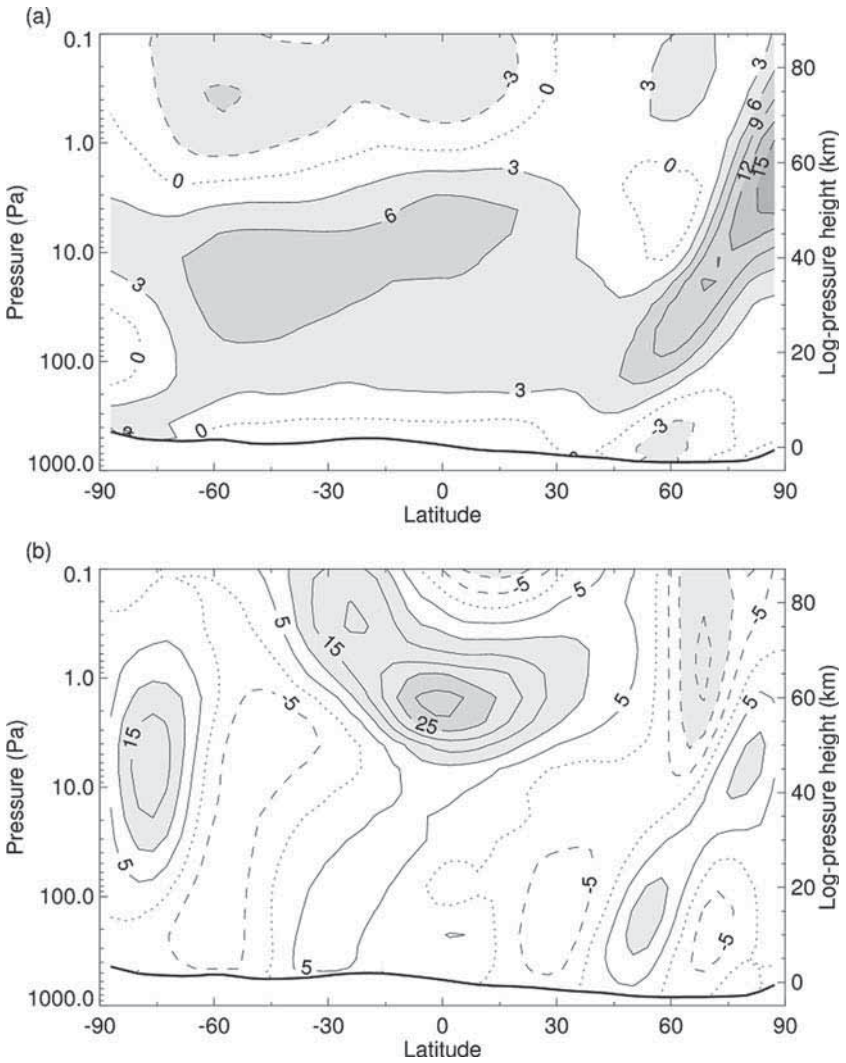



**Fig. 3** Differences, assimilation minus a model run with no dust storm, in (**a**) zonal-mean temperature and (**b**) zonal-mean zonal wind, averaged over the period shown in Fig. 2. Positive values indicate that the assimilation gives higher values than the model. Reprinted from Lewis et al. (2007), with permission from Elsevier

the observations are made asynchronously from a single orbiting spacecraft. Lewis and Barker (2005) described the atmospheric thermal tide behaviour, an analysis which is extended by Figs. 4 and 5 here to show the diurnal and semidiurnal tidal amplitudes respectively throughout the MGS mapping phase. These amplitudes are difficult to analyse from the data directly, since at low latitudes only two local times of day are observed, but the model responds to the changing dust optical depth and

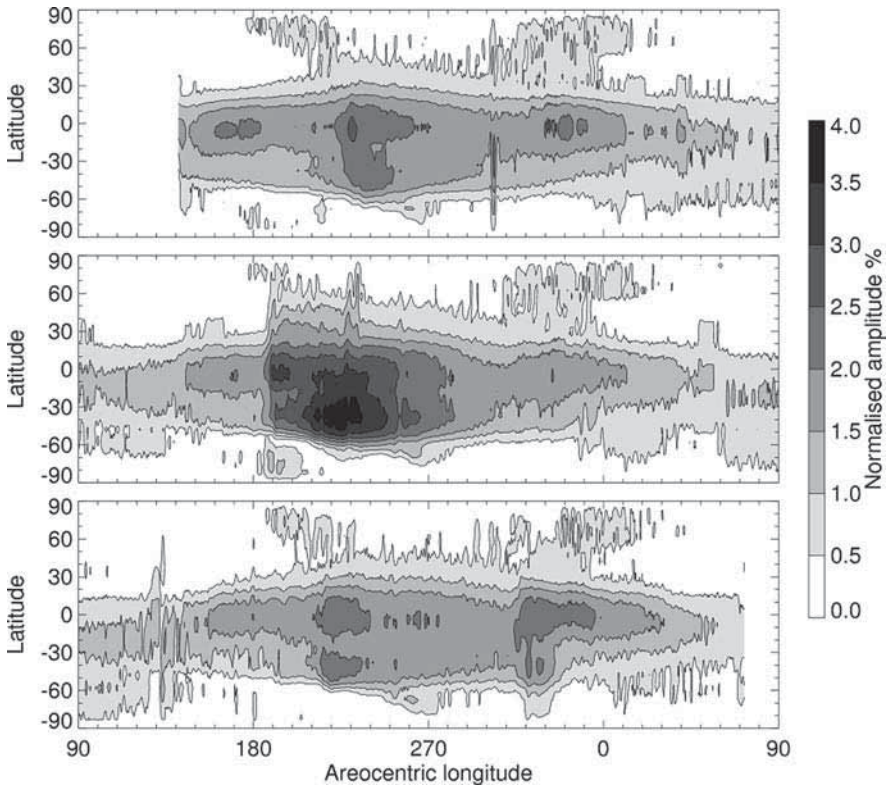



**Fig. 4** The normalized amplitude of the surface pressure signature of the diurnal tide, shown as a function of latitude and time during the MGS mapping phase, the same period as the *lower three panels* of Fig. 1 (Mars Year, *MY* 24, $L_S = 141°$ to Mars Year, *MY* 27, $L_S = 72°$)

exhibits very variable tidal behaviour compared to a model run with a steady, prescribed dust field. The correlation between the semidiurnal tide and optical depth (Fig. 1) is striking.

A principal advantage of data assimilation of data from a single, polar orbiting satellite is in its ability to reconstruct transient waves. The *Hovmoller diagrams* in Fig. 6 show (a) transient temperature on the 50 Pa pressure surface (~25 km altitude) and (b) transient pressure, corrected to the Mars reference datum to remove topographic signals. Both variables are shown at 62.5°N over the entire Northern Hemisphere winter period, $L_S = 180°$–$360°$, of *MY* 24, the 1st year of the MGS scientific mapping phase period. The temperature and pressure have been time-filtered to remove tides and quasi-stationary features.

Transient waves can be seen to propagate eastwards in both panels of Fig. 6. These waves have low zonal wavenumbers, primarily 1–3, with wavenumber 1 dominating throughout much of this period. Of interest is the period around $L_S = 220°$–$260°$, when the atmospheric temperature shows a strong, long-period

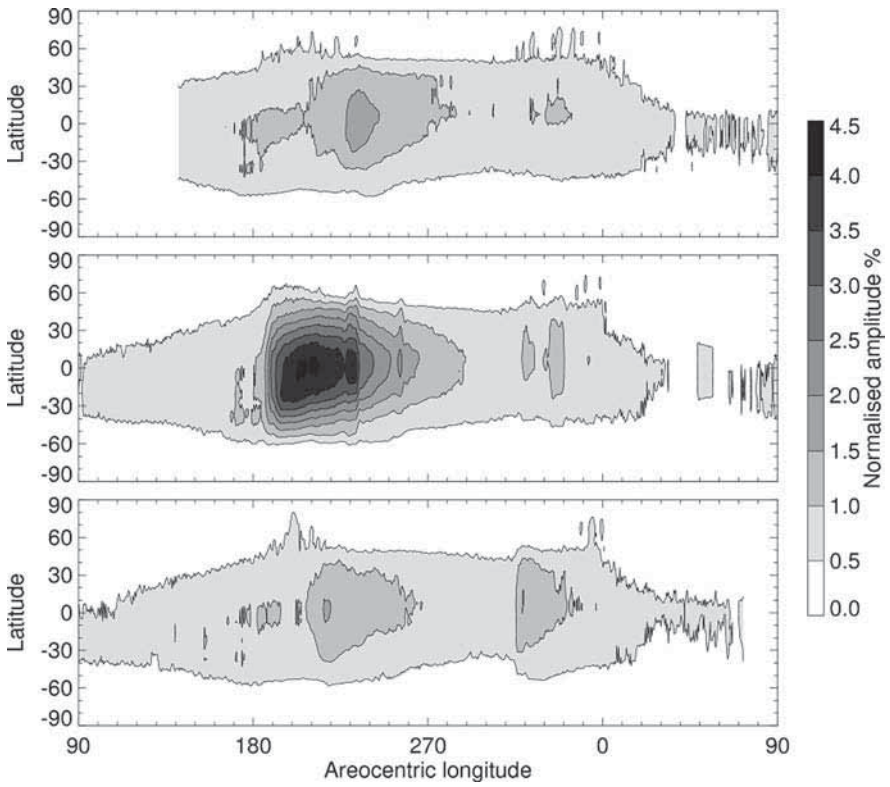



**Fig. 5** The normalized amplitude of the surface pressure signature of the semidiurnal tide, as Fig. 4

wavenumber 1 signal. This is equivalent to a wobble in the polar vortex, the region of strong westerly winds which circulate the Winter Pole (for the analogue on Earth, see chapter *General Concepts in Dynamics and Meteorology*, Charlton-Perez et al.). If this moves to a position not centred over the geographic pole, it will appear as a wavenumber 1 wave as seen at any mid to high latitude. The long-period wavenumber 1 signal is detached from the weaker, shorter period (2–10 days) waves seen near the surface in the pressure signal. A modulation from these waves can still be seen in the temperature signal. At other times the waves are broadly coherent over this altitude range.

It is also notable that the waves near the surface are stronger after the autumn equinox and before the spring equinox, whereas the 50 Pa temperature signal peaks around winter solstice. This solsticial pause in the near-surface waves is seen to recur in all 3 years analysed.

There is a strong topographic influence on the strength of the waves, with maxima being consistently seen at longitudes corresponding to lowlands in *Acidalia*, *Utopia* and *Arcadia Planitia*, which break the longitudinal symmetry of Mars into regions reminiscent of storm zones on Earth.

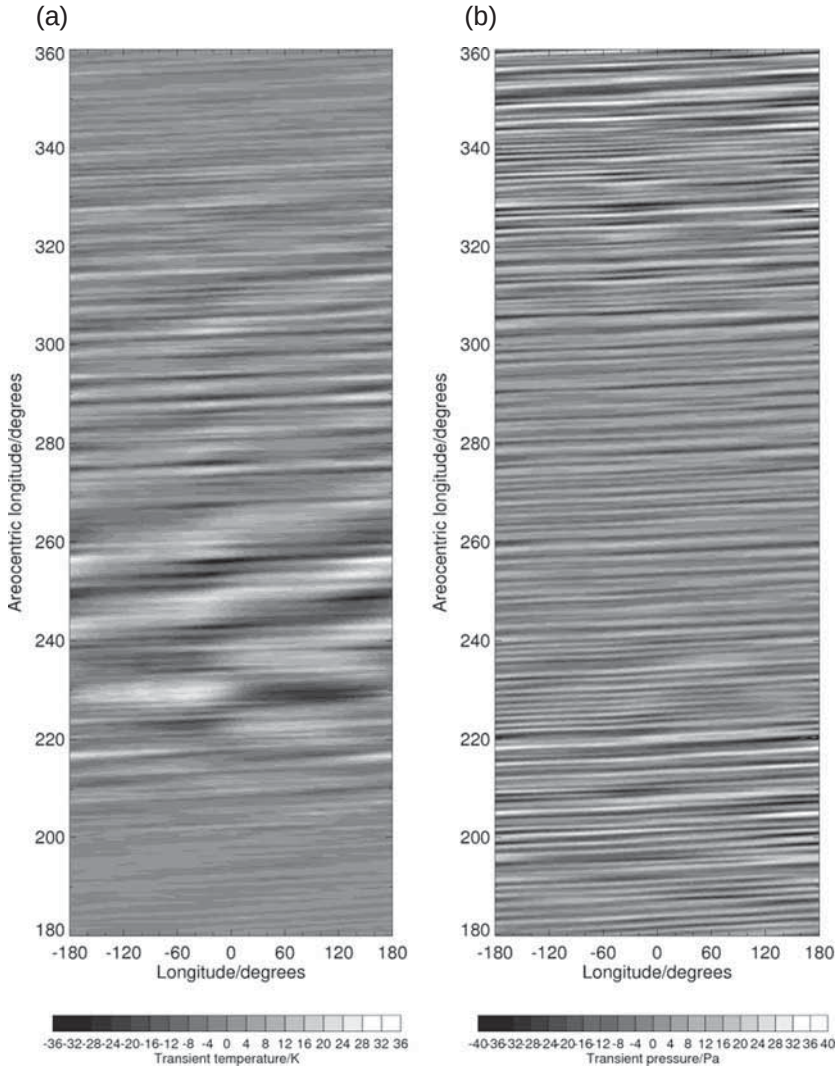



**Fig. 6** (**a**) Transient temperature (K) on the 50 Pa pressure surface (~25 km altitude) and (**b**) transient pressure (Pa), corrected to the Mars reference datum to remove topographic signals at 62.5°N over the period, $L_S = 180°–360°$, of *MY* 24

## 4 Future Prospects for Other Planets

To date, a formal process of data assimilation has not been attempted for any other planetary atmospheres, though the observations that are available are naturally used to inform and to constrain models. The most likely next application beyond Mars is to the atmosphere of Venus, with new ESA Venus Express observations now

available (Titov et al. 2006) and recent advances in Venus GCMs (Yamamoto and Takahashi 2003, 2006; Lee et al. 2005, 2007; Hollingsworth et al. 2007; Lebonnois et al. 2010). The Venus models still have simplified physical parametrizations compared with the Mars and terrestrial GCMs and are not yet able to make accurate, quantitative predictions of, for example, equatorial winds. Venus, with its very long radiative relaxation timescale (varying from around an Earth day at the cloud tops to many years in the lower atmosphere) and slow rotation rate, which means that assumptions about *geostrophic balance* will not apply (see chapter *General Concepts in Meteorology and Dynamics*, Charlton-Perez et al.), will offer new challenges for data assimilation.

Data assimilation for the atmospheres of the Giant Planets is also a challenging prospect. Again, there are few complete global atmospheric models (Dowling et al. 1998, 2006), but there is now a substantial observational data set from the Galileo mission to Jupiter (Young 1998, 2000) and the Cassini mission to Saturn (Mitchell 2007). These missions have not provided the very repetitive coverage of the satellites observing Mars and Venus, but they have made multiple orbits and observed some atmospheric features repeatedly, so limited-area assimilations may become feasible at some point in the future.

## 5 Implications for Terrestrial Data Assimilation

Although data assimilation for other planets is presently only in a nascent state, advances are happening rapidly. Planetary data assimilation has examples of both building on older, established terrestrial techniques and developing new ideas tailored to specific problems. In particular it demonstrates assimilation scheme performance in a very data-poor environment, perhaps more analogous to physical oceanographic applications than to terrestrial numerical weather forecasting. There are particular challenges in a planetary context, often with only poor knowledge of the dominant physical processes and with highly simplified models without a well-known climatology. A notable feature of the martian data assimilation studies outlined in this chapter, and most likely future planetary data assimilation studies, is their reliance on remotely-sensed observations from a single satellite at any one time with the lack of contemporaneous in situ measurements. That there has been at least some limited success reported is interesting with regard to the terrestrial problem. On Earth, surface and near-surface measurements are clearly of great importance for determining the state of the lower atmosphere and satellite observations have been introduced to terrestrial data assimilation schemes for numerical weather prediction at a later stage. The planetary problem is almost being tackled in reverse, with models now incorporating satellite data assimilation being used to form climate databases to assist in the entry, descent and landing process for spacecraft which will hopefully make the surface and in situ atmospheric measurements in future.

**Acknowledgments** The author is grateful to W. Gregory Lawson for his insightful comments on the first draft of this chapter.

## References

Anderson, J.L., B. Wyman, S. Zhang and T. Hoar, 2005. Assimilation of surface pressure observations using an ensemble filter in an idealized global atmospheric prediction system. *J. Atmos. Sci*., **62**, 2925–2938.

Banfield, D., B.J. Conrath, P.J. Gierasch, R.J. Wilson and M.D. Smith, 2004. Traveling waves in the martian atmosphere from MGS TES nadir data. *Icarus*, **170**, 365–403.

Banfield, D., A.P. Ingersoll and C.L. Keppenne, 1995. A steady-state Kalman filter for assimilating data from a single polar orbiting satellite. *J. Atmos. Sci*., **52**, 737–753.

Barnes, J.R., 1980. Time spectral-analysis of mid-latitude disturbances in the martian atmosphere. *J. Atmos. Sci*., **37**, 2002–2015.

Barnes, J.R., 1981. Mid-latitude disturbances in the martian atmosphere – a 2nd Mars year. *J. Atmos. Sci*., **38**, 225–234.

Bengtsson, L. and N. Gustafsson, 1971. Experiment in assimilation of data in dynamical analysis. *Tellus*, **23,** 328–336.

Christensen, P.R., D.L. Anderson, S.C. Chase, R.N. Clark, H.H. Kieffer, M.C. Malin, J.C. Pearl, J. Carpenter, N. Bandiera, F.G. Brown and S. Silverman, 1992. Thermal emission spectrometer experiment – Mars observer mission. *J. Geophys. Res*., **97**, 7719–7734.

Clancy, R.T., S.W. Lee, G.R. Gladstone, W.W. McMillan and T. Rousch, 1995. A new model for Mars atmospheric dust based upon analysis of ultraviolet through infrared observations from Mariner 9, Viking, and Phobos. *J. Geophys. Res*., **100**, 5251–5263.

Collins, M., S.R. Lewis, P.L. Read and F. Hourdin, 1996. Baroclinic wave transitions in the martian atmosphere. *Icarus*, **120**, 344–357.

Conrath, B.J., J.C. Pearl, M.D. Smith and P.R. Christensen, 2002. MGS TES results: Atmospheric structure, aerosols, and dynamics. *Highlights Astron*., **12**, 638–641.

Conrath, B.J., J.C. Pearl, M.D. Smith, W.C. Maguire, P.R. Christensen, S. Dason and M.S. Kaelberer, 2000. Mars Global Surveyor Thermal Emission Spectrometer (TES) observations: Atmospheric temperatures during aerobraking and science phasing. *J. Geophys. Res*., **105**, 9509–9519.

Courtier, P. and O. Talagrand, 1987. Variational assimilation of meteorological observations with the adjoint vorticity equation. 2. Numerical results. *Q. J. R. Meteorol. Soc*., **113**, 1329–1347.

Cunningham, G.E., A.L. Albee and T.E. Thorpe, 1992. Mars Observer as a precursor to intensive exploration of Mars. *Acta Astronautica*, **28**, 259–275.

Dowling, T.E., M.E. Bradley, E. Colon, J. Kramer, R.P. LeBeau, G.C.H. Lee, T.I. Mattox, R. Morales-Juberias, C.J. Palotai, V.K. Parimi and A.P. Showman, 2006. The EPIC atmospheric model with an isentropic/terrain-following hybrid vertical coordinate. *Icarus*, **182**, 259–273.

Dowling, T.E., A.S. Fischer, P.J. Gierasch, J. Harrington, R.P. LeBeau and C.M. Santori, 1998. The explicit planetary isentropic-coordinate (EPIC) atmospheric model. *Icarus*, **132**, 221–238.

Ehrendorfer, M., 1997. Predicting the uncertainty of numerical weather forecasts: A review. *Meteorol. Z*., **6**, 147–183.

Forget, F., F. Hourdin, R. Fournier, C. Hourdin, O. Talagrand, M. Collins, S.R. Lewis, P.L. Read and J.P. Huot, 1999. Improved general circulation models of the martian atmosphere from the surface to above 80 km. *J. Geophys. Res*., **104**, 24155–24175.

Gierasch, P. and R. Goody, 1967. An approximate calculation of radiative heating and radiative equilibrium in the martian atmosphere. *Planet. Space Sci*., **15**, 1465–1477.

Gierasch, P. and R. Goody, 1968. A study of the thermal and dynamical structure of the martian lower atmosphere. *Planet. Space Sci*., **16**, 615–636.

Goody, R. and M.J.S. Belton, 1967. Radiative relaxation times for Mars – a discussion of martian atmospheric dynamics. *Planet. Space Sci*., **15**, 247–256.

Haberle, R.M., J.B. Pollack, J.R. Barnes, R.W. Zurek, C.B. Leovy, J.R. Murphy, H. Lee and J. Schaeffer, 1993. Mars atmospheric dynamics as simulated by the Nasa Ames general-circulation model .1. The zonal-mean circulation. *J. Geophys. Res*., **98**, 3093–3123.

Hess, S.L., J.A. Ryan, J.E. Tillman, R.M. Henry and C.B. Leovy, 1980. The annual cycle of pressure on Mars measured by Viking-Lander-1 and Viking-Lander-2. *Geophys. Res. Lett*., **7**, 197–200.

Hollingsworth, J.L., R.E. Young, G. Schubert, C. Covey and A.S. Grossman, 2007. A simple-physics global circulation model for Venus: Sensitivity assessments of atmospheric superrotation. *Geophys. Res. Lett*., **34**, L05202.

Houben, H., 1999. Assimilation of Mars global surveyor meteorological data. *Adv. Space Res*., **23**, 1899–1902.

Houtekamer, P.L. and H.L. Mitchell, 2005. Ensemble Kalman filtering. *Q. J. R. Meteorol. Soc*., **131**, 3269–3289.

Justus, C.G., B.F. James, S.W. Bougher, A.F.C. Bridger, R.M. Haberle, J.R. Murphy and S. Engel, 2002. Mars-GRAM 2000: A Mars atmospheric model for engineering applications. *Adv. Space Res*., **29**, 193–202.

Kalman, R.E., 1960. A new approach to linear filtering and prediction problems. *Trans. ASME J. Basic Eng*., **82D**, 35–45.

Kalnay, E., 2003. *Atmospheric Modeling, Data Assimilation and Predictability*, Cambridge University Press, New York, 341 pp.

Kass, D.M., 1999. *Change in the Martian Atmosphere,* Ph.D. Thesis, Planetary Science, California Institute of Technology, Pasadena, CA.

Lebonnois, S., F. Hourdin, V. Eymet, A. Crespin, R. Fournier and F. Forget, 2010. Superrotation of Venus' atmosphere analysed with a full General Circulation Model. *J. Geophys. Res.,* (accepted).

Lee, C., S.R. Lewis and P.L. Read, 2005. A numerical model of the atmosphere of Venus. *Planet. Atmos., Ionospheres Magnetospheres*, **36**, 2142–2145.

Lee, C., S.R. Lewis and P.L. Read, 2007. Superrotation in a Venus general circulation model. *J. Geophys. Res*., **112**, E04S11.

Lewis, S.R., 2003. Modelling the martian atmosphere. *Astron. Geophys*., **44**, 6–14.

Lewis, S.R. and P.R. Barker, 2005. Atmospheric tides in a Mars general circulation model with data assimilation. *Adv. Space Res*., **36**, 2162–2168.

Lewis, S.R., M. Collins and P.L. Read, 1997. Data assimilation with a martian atmospheric GCM: An example using thermal data. *Adv. Space Res*., **19**, 1267–1270.

Lewis, S.R., M. Collins, P.L. Read, F. Forget, F. Hourdin, R. Fournier, C. Hourdin, O. Talagrand and J.P. Huot, 1999. A climate database for Mars. *J. Geophys. Res*., **104**, 24177–24194.

Lewis, S.R. and P.L. Read, 1995. An operational data assimilation scheme for the martian atmosphere. *Adv. Space Res*., **16**, 9–13.

Lewis, S.R., P.L. Read and M. Collins, 1996. Martian atmospheric data assimilation with a simplified general circulation model: Orbiter and lander networks. *Planet. Space Sci*., **44**, 1395–1409.

Lewis, S.R., P.L. Read, B.J. Conrath, J.C. Pearl and M.D. Smith, 2007. Assimilation of thermal emission spectrometer atmospheric data during the mars global surveyor aerobraking period. *Icarus*, **192**, 327–347.

Lorenc, A.C., R.S. Bell and B. Macpherson, 1991. The meteorological office analysis correction data assimilation scheme. *Q. J. R. Meteorol. Soc*., **117**, 59–89.

McCleese, D.J., R.D. Haskins, J.T. Schofield, R.W. Zurek, C.B. Leovy, D.A. Paige and F.W. Taylor, 1992. Atmosphere and climate studies of Mars using the Mars observer pressure modulator infrared radiometer. *J. Geophys. Res*., **97**, 7735–7757.

McCleese, D.J., J.T. Schofield, F.W. Taylor, S.B. Calcutt, M.C. Foote, D.M. Kass, C.B. Leovy, D.A. Paige, P.L. Read and R.W. Zurek, 2007. Mars climate sounder: An investigation of thermal and water vapor structure, dust and condensate distributions in the atmosphere, and energy balance of the polar regions. *J. Geophys. Res*., **112**, E05S06.

Mitchell, R.T., 2007. The Cassini mission at Saturn. *Acta Astronautica*, **61**, 37–43.

Montabone, L., S.R. Lewis and P.L. Read, 2005. Interannual variability of martian dust storms in assimilation of several years of Mars Global Surveyor observations. *Adv. Space Res*., **36**, 2146–2155.

Montabone, L., S.R. Lewis, P.L. Read and D.P. Hinson, 2006a. Validation of martian meteorological data assimilation for MGS/TES using radio occultation measurements. *Icarus*, **185**, 113–132.

Montabone, L., S.R. Lewis, P.L. Read and P. Withers, 2006b. Reconstructing the weather on Mars at the time of the MERs and Beagle 2 landings. *Geophys. Res. Lett*., **33**, L19202.

Newman, C.E., P.L. Read and S.R. Lewis, 2004. Investigating atmospheric predictability on Mars using breeding vectors in a general circulation model. *Q. J. R. Meteorol. Soc*., **130**, 2971–2989.

Pollack, J.B., R.M. Haberle, J. Schaeffer and H. Lee, 1990. Simulations of the general circulation of the martian atmosphere. 1. Polar processes. *J. Geophys. Res*., **95**, 1447–1473.

Rabier, F., 2005. Overview of global data assimilation developments in numerical weather-prediction centres. *Q. J. R. Meteorol. Soc*., **131**, 3215–3233.

Read, P.L. and S.R. Lewis, 2004. *The Martian Climate Revisited: Atmosphere and Environment of a Desert Planet,* Springer-Praxis Publisher, Berlin, New York, 402 pp.

Rutherford, I., 1972. Data assimilation by statistical interpolation of forecast error fields. *J Atmos. Sci*., **29**, 809–815.

Saunders, R.S., R.E. Arvidson, G.D. Badhwar, W.V. Boynton, P.R. Christensen, F.A Cucinotta, W.C. Feldman, R.G. Gibbs, C. Kloss, M.R. Landano, R.A. Mase, G.W. McSmith, M.A. Meyer, I.G. Mitrofanov, G.D. Pace, J.J. Plaut, W.P. Sidney, D.A. Spencer, T.W. Thompson and C.J. Zeitlin, 2004. 2001 Mars Odyssey mission summary. *Space Sci. Rev*., **110**, 1–36.

Schmidt, R., 2003. Mars Express – ESA's first mission to planet Mars. *Acta Astronautica*, **52**, 197–202.

Schofield, J.T., J.R. Barnes, D. Crisp, R.M. Haberle, S. Larsen, J.A. Magalhaes, J.R. Murphy, A. Seiff and G. Wilson, 1997. The Mars Pathfinder atmospheric structure investigation meteorology (ASI/MET) experiment. *Science*, **278**, 1752–1758.

Smith, M.D., 2004. Interannual variability in TES atmospheric observations of Mars during 1999–2003. *Icarus*, **167**, 148–165.

Smith, M.D., J.C. Pearl, B.J. Conrath and P.R. Christensen, 2000. Mars Global Surveyor Thermal Emission Spectrometer (TES) observations of dust opacity during aerobraking and science phasing. *J. Geophys. Res*., **105**, 9539–9552.

Smith, M.D., J.C. Pearl, B.J. Conrath and P.R. Christensen, 2001. Thermal Emission Spectrometer results: Mars atmospheric thermal structure and aerosol distribution. *J. Geophys. Res*., **106**, 23929–23945.

Smith, M.D., M.J. Wolff, N. Spanovich, A. Ghosh, D. Banfield, P.R. Christensen, G.A. Landis and S.W. Squyres, 2006. One martian year of atmospheric observations using MER Mini-TES. *J. Geophys. Res*., **111**, E12S13.

Talagrand, O. and P. Courtier, 1987. Variational assimilation of meteorological observations with the adjoint vorticity equation. 1. Theory. *Q. J. R. Meteorol. Soc*., **113**, 1311–1328.

Titov, D.V., H. Svedhem, D. McCoy, J.P. Lebreton, S. Barabash, J.L. Bertaux, P. Drossart, V. Formisano, B. Haeusler, O.I. Korablev, W. Markiewicz, D. Neveance, M. Petzold, G. Piccioni, T.L. Zhang, F.W. Taylor, E. Lellouch, D. Koschny, O. Witasse, M. Warhaut, A. Acomazzo, J. Rodrigues-Cannabal, J. Fabrega, T. Schirmann, A. Clochet and M. Coradini, 2006. Venus express: Scientific goals, instrumentation, and scenario of the mission. *Cosmic Res*., **44**, 334–348.

Toth, Z., 2001. Ensemble forecasting in WRF. *Bull. Amer. Meteorol. Soc*., **82**, 695–697.

Wilson, R.J., D. Banfield, B.J. Conrath and M.D. Smith, 2002. Traveling waves in the northern hemisphere of Mars. *Geophys. Res. Lett*., **29**, 1684, doi:10.1029/2002GL014866.

Wilson, R.J., S.R. Lewis, L. Montabone and M.D. Smith, 2008. Influence of water ice clouds on martian tropical atmospheric temperatures. *Geophys. Res. Lett*., **35**, L07202, doi: 10.1029/2007GL032405.

Yamamoto, M. and M. Takahashi, 2003. The fully developed superrotation simulated by a general circulation model of a Venus-like atmosphere. *J. Atmos. Sci*., **60**, 561–574.

Yamamoto, M. and M. Takahashi, 2006. An aerosol transport model based on a two-moment microphysical parameterization in the Venus middle atmosphere: Model description and preliminary experiments. *J. Geophys. Res*., **111**, E08002.

Young, R.E., 1998. The Galileo probe mission to Jupiter: Science overview. *J. Geophys. Res*., **103**, 22775–22790.

Young, R.E., 2000. Correction to "The Galileo probe mission to Jupiter: Science overview". *J. Geophys. Res*., **105**, 12093–12093.

Zhang, K.Q., A.P. Ingersoll, D.M. Kass, J.C. Pearl, M.D. Smith, B.J. Conrath and R.M. Haberle, 2001. Assimilation of Mars global surveyor atmospheric temperature data into a general circulation model. *J. Geophys. Res*., **106**, 32863–32877.

Zurek, R.W., J.R. Barnes, R.M. Haberle, J.B. Pollack, J.E. Tillman and C.B. Leovy, 1992. Dynamics of the atmosphere of Mars. In *Mars*, Matthews, M.S. (ed.), University of Arizona Press, Tucson, AZ.

# Appendix

## List of Acronyms

| | |
|---|---|
| AATSR: | Advanced Along Track Scanning Radiometer |
| ACE: | Atmospheric Chemistry Experiment |
| ACVT-MA: | Atmospheric Chemistry Validation Team – Modelling and Assimilation |
| ADEOS: | Advanced Earth Observing Satellite |
| ADM: | Atmospheric Dynamics Mission |
| AIRS: | Atmospheric InfraRed Sounder |
| AMSR: | Advanced Microwave Sounding Radiometer (AMSR-E on EOS Aqua, AMSR-2 on GCOM-W) |
| AMSU: | Advanced Microwave Sounding Unit |
| ASAR: | Advanced Synthetic Aperture Radar |
| A-SCOPE: | Advanced Space Carbon and climate Observation of Planet Earth |
| ASSET: | ASSimilation of Envisat daTa |
| ASTER: | Advanced Spaceborne Thermal Emission and reflection Radiometer |
| ATMOS: | Atmospheric Trace MOlecule Spectroscopy |
| ATOVS: | Advanced TOVS |
| BASCOE: | Belgian Assimilation System for Chemical ObsErvation (previously the Belgian Assimilation System for Chemical Observations from Envisat) |
| BIRA-IASB: | Belgisch Instituut voor Rüimte Aeronomie - Institut d'Aeronomie Spatiale de Belgique (Belgian Institute of Space Aeronomy) |
| BLUE: | Best Linear Unbiased Estimate (also Best Linear Unbiased Estimator) |
| CALIPSO: | Cloud Aerosol Lidar and Infrared Path finder Satellite Observation |
| CAMELOT: | Composition of the Atmospheric Mission concEpts and sentinel Observation Techniques |
| CAPACITY: | Composition of the Atmosphere: Progress to Applications in the user CommunITY |
| CCMVal: | Chemistry-Climate Model Validation |
| CERES: | Clouds and the Earth's Radiant Energy System |
| CLAES: | Cryogenic Limb Array Etalon Spectrometer |
| CMA: | China Meteorological Administration |
| CMAM: | Canadian Middle Atmospheer Model |
| CNES: | Centre National d'Études Spatiales |
| CONAE: | COmisión Nacional de Actividades Espaciales (National Space Activities Commission) – Argentina Space Agency |

W. Lahoz et al. (eds.), *Data Assimilation*, DOI 10.1007/978-3-540-74703-1,

| | |
|---|---|
| CoReH$_2$O: | COld REgions Hydrology high-resolution Observatory |
| CrIS: | Cross-track Infrared Sounder |
| CRISTA: | CRyogenic Infrared Spectrometers and Telescopes for the Atmosphere |
| CSA: | Canadian Space Agency |
| CTM: | Chemistry-Transport Model |
| DA: | Data Assimilation |
| DARC: | Data Assimilation Research Centre, UK |
| DLR: | Deutsches zentrum für Luft-und Raumfahrt, Germany |
| DMSP: | Defense Meterological Satellite Program |
| DORIS: | Doppler Orbitography and Radiopositioning Integrated by Satellite |
| DU: | Dobson Units |
| EarthCARE: | Earth Clouds And Radiation Explorer |
| EC: | European Commission |
| ECMWF: | European Centre for Medium-Range Weather Forecasts |
| ECV: | Essential Climate Variable |
| EKF: | Extended Kalman Filter |
| EnKF: | Ensemble Kalman Filter |
| EOS: | Earth Observing System |
| EOS MLS: | EOS Microwave Limb Sounder |
| EP: | Earth Probe |
| EPS: | EUMETSAT Polar System |
| ERA: | ECMWF ReAnalysis |
| ERS: | European Research Satellite |
| ESA: | European Space Agency |
| ESSA: | Environmental Survey Satellite |
| EU: | European Union |
| EUMETSAT: | EUropean organisation for the exploitation of METeorological SATellites |
| FCDR: | Fundamental Climate Data Record |
| FGAT: | First Guess at the Appropriate Time |
| FLEX: | FLuorescence Explorer |
| FTIR: | Fourier Transform InfraRed |
| GCM: | General Circulation Model |
| GCOM: | Global Change Observation Mission |
| GCOS: | Global Climate Observing System |
| GEMS: | Global Earth system Monitoring using Space and in-situ data |
| GEO: | Group on Earth Observations |
| GEOS: | Goddard Earth Observing System |
| GEOSS: | Global Earth Observing System of Systems |
| GERB: | Geostationary Earth Radiation Budget experiment |
| GHRSST: | Global Ocean Data Assimilation Experiment (GODAE) High Resolution SST project |
| GLI: | GLobal Imager |
| GMAO: | Global Modeling Assimilation Office (previously the Data Assimilation Office, DAO) |
| GMES: | Global Monitoring for Environment and Security |
| GOCE: | Gravity field and steady-state OCEan circulation |
| GODAE: | Global Ocean Data Assimilation Experiment |
| GOES: | Geostationary Operational Environmental Satellite |
| GOME and GOME-2: | Global Ozone Monitoring Experiment |
| GOMOS: | Global Ozone Monitoring by Occultation of Stars |
| GOS: | Global Observing System |
| GOSAT: | Greenhouse gas Observing SATellite |

| | |
|---|---|
| GSI: | Gridpoint Statistical Interpolation |
| HALOE: | HALogen Occultation Experiment |
| HIRDLS: | High Resolution Dynamics Limb Sounder |
| HIRS/4: | High resolution Infrared Radiation Sounder/4 |
| HRDI: | High Resolution Doppler Imager |
| HSB: | Humidity Sounder for Brazil |
| IASI: | Infrared Atmospheric Sounding Interferometer |
| IGACO: | Integrated Global Atmospheric Chemistry Observations |
| ILAS: | Improved Limb Atmospheric Spectrometer |
| IR: | InfraRed |
| ISAMS: | Improved Stratospheric And Mesospheric Sounder |
| JAXA: | Japan Aerospace space eXploration Agency |
| KF: | Kalman Filter |
| KNMI: | Koninklijk Nederlaands Meteorologisch Instituut (The Royal Dutch Meteorological Institute) |
| LEKF: | Local Ensemble Kalman Filter |
| LETKF: | Local Ensemble Transform Kalnam Filter |
| LIMS: | Limb Infrared Monitor of the Stratosphere |
| LRR: | Laser RetroReflector |
| MACC: | Monitoring Atmospheric Composition and Climate |
| MAESTRO: | Measurements of Aerosol Extinction in the Stratosphere and Troposphere Retrieved by Occultation |
| MERIS: | MEdium Resolution Imaging Spectrometer |
| MIPAS: | Michelson Interferometer for Passive Atmospheric Sounding |
| MISR: | Multi-angle Imaging SpectroRadiometer |
| MLS: | Microwave Limb Sounder |
| MODIS: | MODerate resolution Imaging Spectroradiometer |
| MOPITT: | Measurements Of Pollution in The Troposphere |
| MOZART: | Model of OZone And Related Tracers |
| MSC: | Met Service Canada |
| MSG: | Meteosat Second Generation |
| MTG: | Meteosat Third Generation |
| MWR: | MicroWave Radiometer |
| NASA: | National Aeronautics and Space Administration |
| NCAR: | National Center for Atmospheric Research |
| NCEP: | National Centers for Environmental Prediction |
| NCEP GFS: | NCEP Global Forecasting System |
| NIES: | Japanese National Institute for Environmental Studies |
| NILU: | Norsk Institutt for Luftforskning (Norwegian Institute for Air Research) |
| NMC: | National Meteorological Center |
| NMHCs: | Non-Methane HydroCarbons |
| NOAA: | National Oceanic and Atmospheric Administration |
| NPOESS: | National Polar-orbiting Operational Environmental Satellite System |
| NWP: | Numerical Weather Prediction |
| OCO: | Orbiting Carbon Observatory |
| OI: | Optimal Interpolation |
| OmA: | Observation minus Analysis |
| OmF: | Observation minus Forecast |
| OMI: | Ozone Monitoring Instrument |
| OMPS: | Ozone Mapping and Profiler Suite |
| OSIRIS: | Optical Spectrograph and InfraRed Imager System |
| OSE: | Observing System Experiment |
| OSSE: | Observing System Simulation Experiment |

| | |
|---|---|
| OSTIA: | Operational Sea Surface Temperature and Sea Ice Analysis |
| PARASOL: | Polarization and Anisotropy of Reflectances for Atmospheric Sciences coupled with Observations from a Lidar |
| PIRATA: | Prediction and Research Moored Array in the Tropical Atlantic (formerly the Pilot Research Moored Array in the Tropical Atlantic) |
| POAM: | Polar Ozone and Aerosol Measurement |
| POLDER: | POLarization and Directionality of the Earth's Reflectance |
| PREMIER: | PRocess Exploration through Measurements of Infrared and milli-metre wave Emitted Radiation |
| PROMOTE: | PROtocol for MOnitoring for The GMES service Element |
| PSAS: | Physical-space Statistical Analysis Scheme |
| PSC: | Polar Stratospheric Cloud |
| RA-2: | Radar Altimeter 2 |
| RH: | Relative Humidity |
| RT: | Radiative Transfer |
| RTM: | Radiative Transfer Model |
| SAR: | Synthetic Aperture Radar |
| SBUS: | Solar Backscatter Ultraviolet Sounder |
| SBUV/2: | Solar Backscatter Ultra-Violet/2 |
| SCIAMACHY: | Scanning Imaging Absorption spectrometer for Atmospheric CHartographY |
| SEVIRI: | Spinning Enhanced Visible and InfraRed Imager |
| SGLI: | Second generation GLI |
| SMAP: | Soil Moisture Active and Passive |
| SMOS: | Soil Moisture and Ocean Salinity |
| SMR: | Sub-Millimeter Radiometer |
| SNSB: | Swedish National Space Board |
| SPARC: | Stratospheric Processes And their Role in Climate |
| SPEEDY: | Simplified Parameterizations primitivE-Equation Dynamics model |
| SSM/I: | Special Sensor Microwave Imager |
| SSMIS: | Special Sensor Microwave Imager/Sounder |
| SWIFT: | Stratospheric Wind Interferometer For Transport studies |
| TES: | Tropospheric Emission Spectrometer |
| TIROS: | Television and InfraRed Observations Satellite |
| TMI: | TRMM Microwave Imager |
| TOMS: | Total Ozone Mapping Spectrometer |
| TOU: | Total Ozone Unit |
| TOVS: | TIROS Operational Vertical Sounder |
| TRAQ: | TRopospheric composition and Air Quality |
| TRMM: | Tropical Rainfall Measuring Mission |
| UARS: | Upper Atmosphere Research Satellite |
| UKMO: | UK Meteorological Office (now The Met Office) |
| UNFCCC: | United Nations Framework Convention on Climate Change |
| UTLS: | Upper Troposphere / Lower Stratosphere |
| UV: | UltraViolet |
| VAR: | VARiational |
| WMO: | World Meteorological Organization |
| WMO-GAW: | WMO – Global Atmospheric Watch |

# Index