VOLUME 112
ISSUE 4
15 APRIL 2008
Special Issue: Remote Sensing Data Assimilation
ELSEVIER
Remote Sensing of Environment
http://www.elsevier.com/locate/rse

Available online at www.sciencedirect.com

ScienceDirect

Remote Sensing of Environment 112 (2008) 1421 – 1433

Remote Sensing of Environment

www.elsevier.com/locate/rse

Ensemble member generation for sequential data assimilation

M.R.J. Turner [a,*], J.P. Walker [a], P.R. Oke [b]

[a] *Department of Civil and Environmental Engineering, The University of Melbourne, Parkville, Victoria 3010, Australia*
[b] *CSIRO Marine and Atmospheric Research and Wealth from Oceans Flagship Program, Hobart, Tasmania 7001, Australia*

Received 5 September 2006; received in revised form 30 January 2007; accepted 4 February 2007

Abstract

Using an ensemble of model forecasts to describe forecast error covariance extends linear sequential data assimilation schemes to nonlinear applications. This approach forms the basis of the Ensemble Kalman Filter and derivative filters such as the Ensemble Square Root Filter. While ensemble data assimilation approaches are commonly reported in the scientific literature, clear guidelines for effective ensemble member generation remain scarce. As the efficiency of the filter is reliant on the accurate determination of forecast error covariance from the ensemble, this paper describes an approach for the systematic determination of random error. Forecast error results from three factors: errors in initial condition, forcing data and model equations. The method outlined in this paper explicitly acknowledges each of these sources in the generation of an ensemble. The initial condition perturbation approach presented optimally spans the dynamic range of the model states and allows an appropriate ensemble size to be determined. The forcing data perturbation approach treats forcing observations differently according to their nature. While error from model physics is not dealt with in detail, discussion of some commonly used approaches and their limitations is provided. The paper concludes with an example application for a synthetic coastal hydrodynamic experiment assimilating sea surface temperature (SST) data, which shows better prediction capability when contrasted with standard approaches in the literature.

Keywords: Ensemble member generation; Sequential data assimilation; Ensemble filtering; Hydrodynamic modelling; Remote sensing; SST

1. Introduction

The Ensemble Kalman Filter (EnKF) was introduced by Evensen (1994) to ameliorate linearisation errors in model state analyses and error covariance estimates when applying the Extended Kalman Filter (EKF) to highly nonlinear problems. The EnKF has been widely applied in oceanography and meteorology (e.g., Evensen & van Leeuwen, 1996; Keppenne, 2000) and more recently in hydrology (Reichle et al., 2002a) with demonstrable success.

A significant issue in the application of ensemble forecast assimilation schemes is to achieve a realistic range of ensemble members from which the model error covariances are diagnosed, termed the forecast error here. Failure to achieve such a set of ensemble members will result in a suboptimal analysis, as the Kalman gain weighting matrix will place undue emphasis on either the observations or the modelled forecasts (depending on whether the forecast error is under or over estimated) and also affect the correlated states. While the need for appropriate forecast error is appreciated, it is often difficult in practice to determine the appropriate level of forecast error to be used in the assimilation. Consequently it is frequently not dealt with as rigorously as other aspects of ensemble data assimilation, such as the type of ensemble filter. The observation error covariances are assumed in this paper to be well understood.

There are three factors that contribute to error in a model forecast and these should be used to achieve variability in an ensemble forecast. These are i) initial conditions, ii) forcing data, and iii) model equations. This paper investigates the first two of these error sources. Model error as a result of equation choice, domain discretisation and parameter accuracy, is not investigated in this paper, but a limited discussion is included for completeness. The methods described in this paper for the

[*] Corresponding author.
E-mail addresses: mturner@civenv.unimelb.edu.au (M.R.J. Turner), j.walker@unimelb.edu.au (J.P. Walker), peter.oke@csiro.au (P.R. Oke).

0034-4257/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.rse.2007.02.042

generation and propagation of an ensemble are demonstrated by means of an observation system simulation experiment (OSSE) for the assimilation of sea surface temperature (SST) into a coastal hydrodynamic model.

2. The Ensemble Kalman Filter

The EnKF accounts for nonlinear models through an ensemble of model predictions which use the nonlinear model physics. The ensemble analysis is expressed as

$$\mathbf{X}^a = \mathbf{X}^f + \mathbf{P}_e\mathbf{H}^T[\mathbf{H}\mathbf{P}_e\mathbf{H}^T + \mathbf{R}]^{-1}[\mathbf{D} - \mathbf{H}\mathbf{X}^f], \tag{1}$$

where $\mathbf{X}$ is a matrix of n model state realisations $[\mathbf{x}^1, \mathbf{x}^2,\ldots, \mathbf{x}^n]$, $\mathbf{D}$ is a matrix of observation ensembles, subscript e denotes an ensemble approximation, and superscripts a and f denote analysis and forecast respectively. In the standard EnKF (Evensen, 2003), perturbations are added to the observations to generate a matrix consisting of an ensemble of observations $\mathbf{D}$, based on the observation error covariance $\mathbf{R}$. The forecast error covariance $\mathbf{P}$ is approximated from the ensemble of model predictions by

$$\mathbf{P}_e = \frac{\mathbf{X}'^f\mathbf{X}'^{f\mathrm{T}}}{n-1}, \tag{2}$$

where

$$\mathbf{X}'^f = \mathbf{X}^f - \overline{\mathbf{X}}^f, \tag{3}$$

denotes a matrix of ensemble deviations with the overbar denoting the ensemble mean, $\overline{\mathbf{X}}^f = \mathbf{X}^f\mathbf{1}_n$, and $\mathbf{1}_n$ is an $n \times n$ matrix in which each element has a value of $\frac{1}{n}$. The use of an ensemble approximation $\mathbf{P}_e$ is based on the assumption that in the limit of an infinite number of ensembles members

$$\lim_{n\to\infty} \mathbf{P}_e = \mathbf{P}. \tag{4}$$

If the spread of the ensemble forecast is too large, then the forecast error covariance $\mathbf{P}_e$ will be overestimated and the analysis will tend to overfit the observations. Conversely, if the spread of the ensemble forecast is too small, then the forecast error covariance $\mathbf{P}_e$ will be underestimated and the analysis will tend to under utilise the observations. In either case, an inaccurate ensemble representation of forecast error will result in a sub-optimal filter. For this reason it is necessary to generate and propagate the ensemble with realistic variability when using an ensemble sequential data assimilation technique. Furthermore, specification of ensemble correlations may be as important as the specification of the ensemble magnitude (spread), however only ensemble spread, and not spatial or cross-correlation of ensemble members, is discussed herein.

3. Ensemble initiation

The uncertainty of the initial state estimates is represented by the initial spread of the ensemble members. In the method outlined by Evensen (2003), ensemble members are generated by taking an initial best-guess of the states, and then adding perturbations in the form of random correlated fields to each ensemble member. Importantly, this approach includes a recommendation to “integrate the ensemble over a time interval covering a few characteristic time scales of the dynamical system” (Evensen, 2003) to ensure dynamic stability and correct multivariate correlations before commencing the assimilation. This approach is the basis of several papers (Houtekamer & Mitchell, 1998; Keppenne, 2000; Reichle et al., 2002a).

An improved sampling scheme has been proposed by Evensen (2004), based upon the work of Pham (2001). This method uses an ensemble of randomly generated, spatially correlated fields, and perturbation independence is sought by performing a Singular Value Decomposition (SVD). The first n singular vectors are then combined with another random orthogonal matrix and the singular values are adjusted appropriately. Zupanski et al. (2006) attempts to address the initialisation problem more explicitly, extending on the previous methods. However, a disadvantage of all the methods mentioned above is that they are applied prior to the assimilation period and require a spin up for dynamic stability, by which time the prescribed error distribution may have been altered by the model equations.

Two additional methods have been used in operational ensemble forecasting: the breeder method (Toth & Kalnay, 1993, 1997) and optimal perturbations (Molteni et al., 1996). The basis of these two methods is to generate a set of the fastest growing errors. The two methods have been investigated in a paper by Miller and Ehret (2002) which studied forecasting of multimodal systems with small ensemble sizes. They found that the optimal perturbations (also termed singular vectors) method performed well, especially for systems with small initial variance. In cases of larger initial variance the breeder method performed well, although there were occurrences when it failed to observe bimodal evolution.

While these two methods may be suitable for ensemble initialisation in certain circumstances, they are unsuitable for recommendation as a generic approach. The optimal perturbation method requires an adjoint model to generate the fastest growing errors, which is typically unavailable unless specifically developed. For this reason, in spite of the obvious benefits of the optimal perturbation method, its use is impractical. While the breeder method is simple to apply, its ability to accurately estimate forecast error variance is questionable. Because it is a random method and relies on the model to generate perturbations in the direction of the largest growing error, there is the possibility that all perturbations generated will cluster towards one direction, thus reducing the ensemble rank. Moreover, the method relies on inherent model nonlinearities to breed the perturbations, making it ineffective for weakly nonlinear models.

In addition to the ensemble spread, it is important that the matrix of ensemble deviations have a high rank. This allows for smaller ensemble sizes, and makes ensemble techniques more efficient. This can be seen by considering the underlying EnKF equations. As Evensen (2003) has shown, the EnKF analysis Eq. (1) can be written as a linear combination of the ensemble deviations, and therefore an analysis is more efficient if the ensemble deviations are independent. Producing ensembles with deviations that are linearly independent results in a more efficient assimilation filter.

A set of n initial state vectors $[\mathbf{x}^1, \mathbf{x}^2,\ldots, \mathbf{x}^n]$ should thus be generated by adding a set of n independent perturbation vectors $[\mathbf{x}'^1, \mathbf{x}'^2,\ldots, \mathbf{x}'^n]$ to the best-guess initial condition, $\mathbf{x}$. The initial state vectors $\mathbf{x}^i$ become the column vectors in $\mathbf{X}$, while the independent perturbation vectors $\mathbf{x}'^i$ become the column vectors of $\mathbf{X}'^f$.

A range of physically realistic deviations can be obtained by taking snapshots of state values from a long model run and removing the spatial mean from each at those instants in time. This gives a $m\times p$ matrix, $\mathbf{F}$, where m is the number of state variables at different points in space and p is the number of snapshots extracted. Each column is a vector representing physically realistic perturbations about a zero mean. Each element $\mathbf{F}_{y,k}$ can be expressed as

$$\mathbf{F}_{y,k} = [\mathbf{X}_k]_y - <\mathbf{X}_k>, \tag{5}$$

where $\mathbf{X}_{y,k}$ is the state value at position y and time k, and $<\mathbf{X}_k>=1/N.\sum_{y=1}^{N}[\mathbf{X}_k]_y$, with N the number of gird points: the spatial average at time step k.

By extracting snapshots at a time interval less than the smallest temporal scale and over a time period longer than the largest characteristic time scale, the full dynamic range of conditions of the field to be initialised will be covered, thus spanning a wide range of deviation possibilities. A set of n ensemble deviations $\mathbf{X}'^f$ is then taken from the first n spatial singular vectors of a SVD of $\mathbf{F}$ as described below.

The matrix $\mathbf{F}$ is decomposed using a SVD such that

$$\mathbf{F} = \mathbf{U\Sigma V}^T, \tag{6}$$

where $\mathbf{U}$ and $\mathbf{V}$ are square orthogonal column matrices of dimensions $m\times m$ and $p\times p$ respectively and $\mathbf{\Sigma}$ is a diagonal matrix with diagonal elements that are the singular values of $\mathbf{F}$ arranged in nonincreasing order. The singular values express the importance of their respective (spatial) singular vector, the columns of $\mathbf{U}$. As the columns of $\mathbf{U}$ are orthogonal, perturbation independence is assured. By using the singular vectors contained in the first n columns of $\mathbf{U}$, the range of dynamic states is objectively, and concisely, represented, as these vectors explain the most significant spatial variation in the model.

The singular vectors are then scaled so that their standard deviation is equal to an a *priori* assumption of the initial state uncertainty. If the initial variance of the ensemble is unknown, the average univariate vector standard deviation gives guidance for the initial spread of the ensemble members. Except for the last step of scaling the singular vectors to the a priori initial uncertainty, this method provides an objective means for initialising an ensemble, ensuring linear independence of each ensemble deviation.

4. Ensemble size

Determination of the number of ensemble members required for reliable forecast error estimation is an unresolved issue for sequential ensemble data assimilation methods. While many studies have focussed on the sensitivity of an ensemble forecast system to ensemble size (e.g. Houtekamer & Mitchell, 2001), there have been no recommendations made based on sound theoretical evidence. If the ensemble deviations are independent, a guide to the upper limit on the number of ensemble members required is the number of model state elements. A larger ensemble size would imply some level of perturbation dependency and the resulting analysis would be inefficient.

Setting the number of ensemble members equal to the number of model state elements is unrealistic for a large distributed model. For the OSSE case study presented later in this paper, this would entail an ensemble with between 20,000 and 100,000 members, depending on whether univariate or multivariate assimilation was pursued. A deterministic thirty-day model run takes about one hour of real time to compute [1] meaning that it would take over two years to compute the same forecast with 20,000 ensemble members with the same computing resources.

Such an exercise is also unnecessary in models that contain a high degree of state interdependence. Where model states are evolved by similar equations and forced by similar conditions, their values become highly correlated and consequently the errors are also highly correlated. When ensemble errors are highly correlated, fewer independent vectors are required to describe the range of ensemble perturbations and the ensemble size can be reduced accordingly.

The degree of state independence can be determined through the significance of singular values obtained from the SVD performed in the previous section. For instance, if 95% of the variance in the system is explained by the first 50 singular values, using 500 ensemble members would be excessive.

This approach considers an optimal case, which assumes that the ensemble deviations are independent. In reality, subject to similar forcing data and model equations, some degree of ensemble deviation dependence must develop over time, implying that more ensemble members are needed than the SVD suggests. For instance, even with one variable, a minimum number of ensemble members are needed to express the probability distribution of the error about the state. As such, ensemble assimilation methods benefit large-state models more than small-state models.

5. Ensemble propagation

Once the number of initial ensemble members has been determined and the ensemble members initiated, it is necessary to propagate them through time. Without a mechanism to continually introduce realistic forecast error into the ensemble propagation – errors from model physics, parameter uncertainties and model forcing error – the ensemble spread will tend to collapse for bounded, i.e. nondispersive and nonchaotic, models. This is due to the analysis step reducing the ensemble uncertainty or spread by

$$\mathbf{P}^a = (\mathbf{I} - \mathbf{KH})\mathbf{P}^f, \tag{7}$$

[1] Using a dedicated dual-processor sunfire v60x server.

each time observations are assimilated. Ensemble variation should therefore increase with time after an analysis step, in response to uncertain forcing fields and model errors from inadequate physics and uncertain parameters. To maintain appropriate spread in the ensemble forecast, noise must be added to the forecast. This can be through the addition of model error either as resulting from uncertainties in model physics and parameters therein, or through the use of an ensemble of forcing data. Innovation filters such as those discussed by Mitchell and Houtekamer (2000) and their statistical analysis are of use for diagnosing inaccurate model error covariance information but will not be covered specifically in this paper. The nonlinear forecast model can be explicitly represented as

$$x_{k+1} = f(\mathbf{x}_k, \mathbf{h}_k + \epsilon_k) + v_k, \tag{8}$$

where $\mathbf{h}_k$ is the forcing data at time k, $\boldsymbol{\epsilon}_k$ and v_k are zero mean random processes representing model error caused by forcing data error and model physics equation error, respectively.

5.1. Model physics error

Model predictions incorporate error even when forcing data are perfect due to the choices or limitations in regard to the model physics and parameters. This includes error associated with assumptions, theory and or conceptualisations within the underlying equations, errors due to the computational grid and its discretisation, numerical errors associated with the timestep or numerical methods used to solve the mathematical equations, and uncertainty associated with any model parameters adopted. In his review paper, Hamill (2002) lists three categories for adding model error: i) using stochastic equations, ii) adding noise to the forecast ensemble at the analysis time (without integrating noise in the model), and iii) using multimodel ensembles. The methods for dealing with model physics error are well described in the literature, even though some methods such as the addition of noise at analysis time can lead to the creation of physically unrealistic model states. Model physics error is not included in this paper and will not be discussed further.

5.2. Forcing error

The impact of forcing data as a source of model error has usually received less attention than the model structure and parameterisation. This apparent oversight is probably due to the mind set of the oceanic and atmospheric data assimilation community, where work is predominantly undertaken with large scale chaotic models. Although forcing error may be included in these models, it is generally of secondary importance and restricted to the ensemble initialisation; model error due to chaotic nonlinear equations dominates the forecast error. However, for other models the influence of forcing data may be more important, and in fact the dominant source of forecast error. For example, without external influences diffusive systems evolve to steady state spatially mean conditions. Errors in forcing data are associated with the measurement (or prediction) of forcing data and its spatial representation.

Establishing the uncertainty associated with forcing data is simpler than establishing model physics (or equation) uncertainty. This is because the uncertainty of recording instruments is typically well known, and as data are collected at various locations the spatial uncertainty can be reliably estimated.

While discussion of forcing error is rare in the literature, it has received some recent attention (Brusdal et al., 2003; Natvik & Evensen, 2003; Reichle et al., 2002b; Robert & Alves, 2003). These papers have generally included forcing error by adding Gaussian random noise to the forcing fields using a specified standard deviation, although the treatment of perturbed forcing appears to have been undertaken in a simplistic manner. For instance, Reichle et al. (2002b) selected the size of perturbations to be added based on simple order-of-magnitude considerations. As such, there remains considerable scope to deal with perturbed forcing data more rigorously. Henceforth, a theoretical framework for generating perturbed forcing data is developed here.

The aim of the approach is to avoid the addition of bias to the forcing data while adding perturbations that represent the forcing data uncertainty. Throughout the discussion the data are assumed to be point time series, which is appropriate for many data assimilation applications. However, there would be little difficulty in extending the techniques described here to spatially varying fields. As noted previously this paper is focused on determination of the correct ensemble spread and not on correlations between the ensembles.

A framework for generating perturbed forcing data for typical data types is as follows. Consider the vector $\mathbf{h}^o$ containing an observed time series of point forcing data h_k (scalar) with p records in time

$$\mathbf{h}^o = [h_1^o, h_2^o \ldots, h_k^o, \ldots h_p^o]^T \tag{9}$$

used to force a model with n ensemble members. If the forcing data are to be unbiased such that $\mathbf{E}\langle\boldsymbol{\epsilon}_k\rangle = 0$, then generation of an ensemble of n forcing data sets $\mathbf{h}^1$, $\mathbf{h}^2$,…, $\mathbf{h}^j$, $\mathbf{h}^n$, is required such that $\mathbf{E}\langle\mathbf{h}_k\rangle = \mathbf{h}_k^o$, with $\mathbf{h}_k$ the kth element (scalar) of the vector $\mathbf{h}$. This condition ensures that the ensemble of forcing data is unbiased relative to the original forcing data.

While various forms are possible, an error and offset form has been adopted to reflect that the data may suffer from both calibration and sampling errors. The jth ensemble member realisation for the scalar forcing variable at time step k is estimated by

$$\mathbf{h}_k^j = h_k^o + \zeta_k^j + \beta^j, \tag{10}$$

with ζ_k^j indicating the jth ensemble realisation of ζ_k with k as time index and β^j is the jth realisation of β. ζ_k is a time dependent error term of $N(0, \sigma_1)$, being a normally distributed random number with zero mean and a standard deviation of σ_1 applied to individual forcing values, and β is an $N(0, \sigma_2)$ offset and a single realisation β^j is applied to the entire jth ensemble member time series. While a single realisation of β is used, its standard deviation σ_2 may vary in time. Eq. (15) gives an example of this. Applying an offset β, in addition to the error term ζ_k, provides an additional mechanism for spreading the forcing ensemble members while also retaining the structure or temporal correlation in the original time series. This is useful for data to which the model is highly noise sensitive, and for data that has a high degree of structure in its time series. Without β and relying entirely on ζ_k may lead to excessive imposed noise that could generate numerical instability, as well as unrealistic data values.

An advantage of the adopted formulation is that it is simple and easily calculated in real time. Moreover, the parameters controlling spread (the standard deviation of β and ζ) can be assigned a physical meaning.

The magnitude and form of perturbations added to generate the ensemble forcing fields are controlled by two standard deviation terms σ_1 and σ_2. Realistic values for these parameters can be obtained by analysing the error in observed data and this allows control over the introduction of forcing error.

Based on the form of Eq. (10), the generation of three types of forcing data are considered: i) unrestricted, ii) semi-restricted, and iii) restricted. The notion of a restricted, or otherwise, data type relates to whether the data type has a fixed boundary outside of which values are not physically allowable.

The different data types are considered because the spatial error distribution varies with data type. In the example presented in this paper, spatial variability is used as a proxy for ensemble variability. The specification of error according to data type aids the generation of unbiased physically realistic data sets. Essentially, the proposed method is a generic means to generate a skewed probability distributions by removing the bias from a Gaussian distribution. For a particular variable where the (skewed-)distribution (Gamma, log-normal etc.) is known, it may be used in preference to the proposed method.

5.3. Unrestricted value fields

An unrestricted data type is not physically constrained over its normal range. An example of an unrestricted data field is air temperature. As the value of the data can range freely throughout the domain, the data error is independent of the data value, and the instrument error in measuring air temperature is assumed constant irrespective of the actual temperature. Unrestricted value fields therefore have the standard deviation of the error term specified as

$$\sigma_1 = \xi, \tag{11}$$

where ξ is constant in time. The standard deviation of the offset is given by

$$\sigma_2 = \chi, \tag{12}$$

where χ is also constant.

5.4. Semi-restricted value fields

A semi-restricted data type is physically constrained by an upper or lower limit. For the lower limited case the domain is $[h_{\min}, \infty)$ and for the upper limited case the domain is $(-\infty, h_{\max}]$. Examples of semi-restricted data fields are precipitation and river flow: both are lower bounded by the value of zero. In the semi-restricted case the standard deviation of the error is generally proportional to the magnitude of the data. For example, the uncertainty associated with determining a flow value for a river in flood from a stage measurement is higher than for a low flow event contained within the river banks, and the uncertainty associated with the flow value becomes zero as the river dries up.

In addition to increased error with magnitude, there is a chance that events occur that are not measured. This is especially true for precipitation. In this case an observation of zero cannot be assumed to have an uncertainty of zero. Although this case is not dealt with here, such events may be added when a rare value is chosen from a random sample.

As a first approximation, the standard deviation of the error term for semi-restricted value fields can be specified as

$$\sigma_1 = (h_k^o - h_{\min})\xi \tag{13}$$

for the lower limited case and

$$\sigma_1 = (h_{\max} - h_k^o)\xi \tag{14}$$

for the upper limited case. Here, σ_1 is linearly dependent upon the difference between the value h_k^o and the data limit $h_{\min}$ or $h_{\max}$ with the proportionality constant ξ. The offset is similarly formed as

$$\sigma_2 = (\hat{h}_k - h_{\min})\chi \quad \text{or} \quad \sigma_2 = (h_{\max} - \hat{h}_i)\chi \tag{15}$$

for the lower and upper limited cases respectively.

Applying a variational error to semi-restricted value data significantly reduces the bias associated with out-of-range values. Data such as precipitation have a lower bound of zero and a significant proportion of zero-valued data. If the unrestricted perturbation approach were applied, on average, half of the ensemble values that were originally zero would be perturbed outside the boundary, requiring truncation to zero to bring them back within the boundary and thus introducing bias. Using a variational error avoids this situation, because the applied error reduces as the boundary is approached, reducing (but not eliminating) the possibility for perturbed values to exceed the boundary. If out of range values are produced they are set to the boundary value.

Using a normally distributed error allows bias to be minimised through judicious choice of ξ and χ. The value of ξ and χ needed to reduce the chance of the perturbed data leaving the lower boundary is guided by the relationship

$$\xi, \chi \preceq \frac{-1}{z_k}, \tag{16}$$

where z_k is a $N(0, 1)$ random number and $\preceq$ is taken to indicate 'generally' less than or equal to. This equation is exact only if either ξ or χ is zero. As probabilities can be associated with the chance of a certain value of z_k being exceeded, the probability of domain exceeding values occurring can be estimated. For example, to reduce the chance of a domain violating error being introduced to less than one in one thousand, $\Pr(z_k \leq -3.125) = 0.0009$, which corresponds to ξ less than 0.32. A similar argument can be constructed for the choice of χ. Further discussion on this and the derivation of Eq. (16) is given in Appendix A. While Eq. (16) calculates an upper limit on possible values of ξ and χ, smaller values are be used and express that the forcing data is known with more certainty.

5.5. Restricted value fields

A restricted data type is physically constrained by an upper and lower bound ($h_{\min}$, $h_{\max}$). Applying a perturbation term to this type of data requires somewhat more consideration be given to the error distribution. A constant error such as Eq. (12) could be used, with any bias due to the truncation of domain exceeding values accepted. A better approach is to assume the maximum standard deviation occurs at the mid point of the domain and reduces linearly to zero at the domain boundaries, giving a triangular shaped distribution by

$$\sigma_1 = \begin{cases} \dfrac{\hat{h}_i - h_{\min}}{h_{\mathrm{mid}} - h_{\min}}\xi, h_{\min} \leq \hat{h}_i \leq h_{\mathrm{mid}}, \\ \dfrac{h_{\max} - \hat{h}_i}{h_{\max} - h_{\mathrm{mid}}}\xi, h_{\mathrm{mid}} < \hat{h}_i \leq h_{\max}, \end{cases} \tag{17}$$

where h_{mid} is $\frac{h_{\max}+h_{\min}}{2}$. An example of a restricted data type is cloud cover. Cloud cover data refers to the proportion of the sky covered by clouds with zero signifying clear skies and eight indicating completely cloudy skies. It is reasonable to associate an error distribution following Eq. (17) with cloud cover data, as it is easy to decide if the sky is completely covered or is completely free from clouds, but to determine whether cloud cover is four, five or six oktas is more difficult and subjective. Moreover, it has been found that cloud cover is more uncertain for midrange values. The applied error distribution takes this into account.

The offset is formed in a similar fashion with

$$\sigma_2 = \begin{cases} \dfrac{\hat{h}_i - h_{\min}}{h_{\mathrm{mid}} - h_{\min}}\chi, h_{\min} \leq \hat{h}_i \leq h_{\mathrm{mid}}, \\ \dfrac{h_{\max} - \hat{h}_i}{h_{\max} - h_{\mathrm{mid}}}\chi, h_{\mathrm{mid}} < \hat{h}_i \leq h_{\max}. \end{cases} \tag{18}$$

As with the semi-restricted case, the choice of ξ and χ values affects the probability that perturbed data leave the domain. For the restricted case the guiding relationship is

$$\xi, \chi \preceq \frac{h_{\min} - h_{\max}}{2z_i}. \tag{19}$$

The derivation of Eq. (19) is given in Appendix A. As with Eq. (16), probabilities can be assigned to different values of z_i. Thus Eq. (19) indicates that a ξ (or χ) value of less than one sixth ($=2 \times 3.125$) of the data range yields a probability of generating a domain leaving perturbed data value of less than one in one thousand. As with the semi-restricted case smaller values can be chosen to reflect the (un)certainty associated with a particular data.

6. Example application

The two methods described to initiate an ensemble and to add forcing error to a propagating ensemble are presented through an OSSE. The application is the assimilation of synthetic SST observations into a coastal hydrodynamic model. The proposed methods are tested by comparing results against those from standard methods taken from the literature.

6.1. Model and experimental framework

The data assimilation experiments presented relate to a hydrodynamic model of Port Phillip Bay (PPB), a shallow (approximately 20 m deep) enclosed bay situated in south eastern Australia (Fig. 1). The (Australian) Commonwealth Scientific and Industrial Research Organisation (CSIRO) Model for Estuaries and Coastal Oceans (MECO) is used to simulate the hydrodynamics of Port Phillip Bay. MECO is a finite difference model that solves the primitive equations using standard numerical techniques and is similar to freely available models such as the Princeton Ocean Model (Blumberg & Mellor, 1987). The atmospheric heat flux is applied based on the bulk parameterisation formulae of Gill (1982). For more details on MECO or the numerical techniques used, the interested reader is referred to Walker et al. (2002) and Herzfeld et al. (2002).

PPB has been modelled with 14 vertical depth layers and a 0.01 degree (~1 km) horizontal grid (~100,000 state variables). The model domain used in the experiments is presented in Fig. 1, showing an open boundary along the southern edge of the model. The main fresh water input is the Yarra River that enters Port Phillip Bay at Melbourne, although other minor riverine inputs are found in the north of PPB. The horizontal extent of the model domain is small and so a spatially constant atmospheric forcing was applied: for each ensemble all surface cells are forced by the same data. As none of the weather stations available collected all variables required by the model, the atmospheric forcing data were taken from different stations (Table 1). A split mode time step was used with a 6 minute time step for 3-dimensional modes and a 6 second time step for the 2-dimensional modes, although results were recorded 2-hourly.

Two sets of atmospheric data were used. One set was used to generate the synthetic truth, while the other was used in the assimilation simulations. A summary of the atmospheric stations used for the various data is given in Table 1 with locations shown in Fig. 1. Common data is cloud cover from Melbourne and incoming solar radiation which is derived theoretically using the algorithm of Zillman (1972).

Using the truth atmospheric data an initial run was made to simulate PPB conditions over the month of January 2003. A long spinup over a 2 year period was made to ensure stable conditions. Fig. 2 presents some diagnostic results from the long model run that give confidence in the model producing sensible results. The largest residual (mean) currents are located at the mouth of Port Phillip Bay. A small residual current runs along the eastern boundary of the bay. This in combination with the higher mean sea level in the north of the bay, suggests a net flux

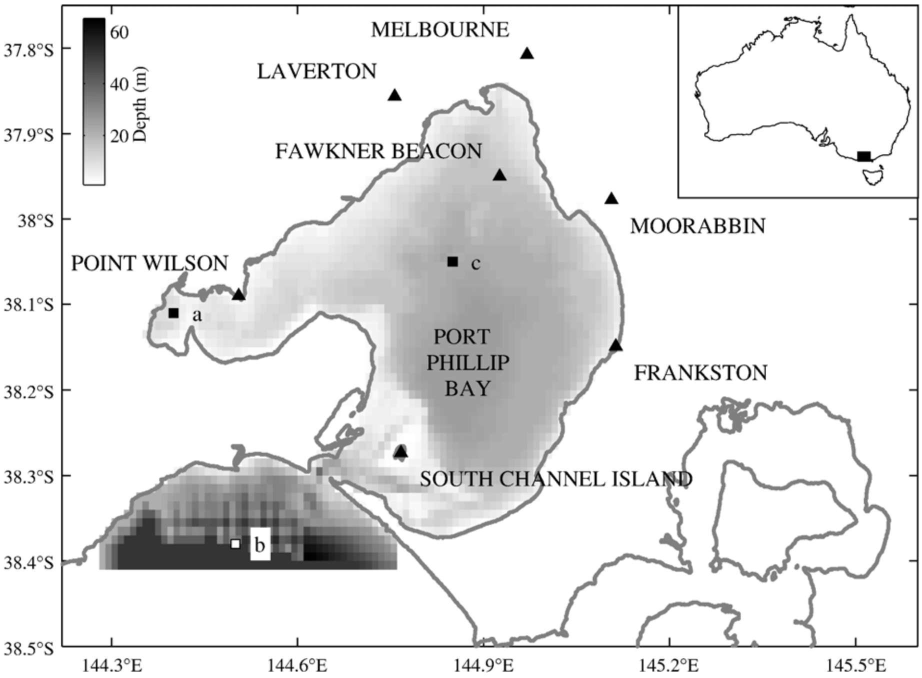

Fig. 1. Location of Port Phillip Bay with bathymetry showing model extent. The locations of weather stations used in the experiments are shown (triangles), together with 3 monitoring locations (squares) denoted a–c.

of water out of the bay. This is consistent with the river flow leaving the bay. The higher standard deviation of sea level outside the bay is also consistent with reality as the entrance to the bay strongly attenuates the tidal signal.

The results from the month-long simulation were denoted the synthetic truth. Surface water temperature snapshots were extracted from the truth every two days to represent satellite observed sea surface temperature. Independent random noise with a standard deviation of 0.5 °C was applied to the extracted field to generate realistic observations. This standard deviation was based on the advanced very high resolution radiometer (AVHRR) nonlinear SST algorithm (NLSST) error characteristics (Nalli & Smith, 1998).

For the assimilation simulation, the model used the second (assimilation) set of atmospheric forcing data and the states were initialised with spatially uniform values: temperature, 18 °C; salinity, 35 practical salinity units (PSU); u- and v-currents, 0 ms^{-1}; and sea level, 0 m. The assimilation analysis has been performed using an Ensemble Square Root Filter (EnSRF) as described in Evensen (2004). This filter was selected because its use of unperturbed observations means that all differences encountered in the experiments can be attributed to the forecast error. Based on the SVD of a long model run, an ensemble size of 20 was chosen for the assimilation. This ensemble size represents 84% of the system variance (Fig. 3). Testing with ensemble sizes of 10 and 50 members found that 10 ensemble members gave a less accurate result. Using 50 members gave a marginally improved result, however the large increase in computational expense for only a marginal improvement did not justify the use of 50 ensemble members.

While all model state variables can be updated during the analysis, only temperature was updated. The analysis was performed in all 3-dimensions (of temperature) simultaneously, rather than treating each spatial unit independently. The analysis of temperature only is based on an investigation which showed that geostrophic balance was not maintained in PPB and that salinity, currents etc. operate independently of temperature. In general, all variables should be included in an analysis unless it can be determined that they operate independently.

6.2. Ensemble initialisation

The ensemble initialisation method proposed in this paper using deviations generated from a SVD of (temperature) deviations extracted from a long model simulation was tested against the ensemble method with deviations generated from correlated random fields. Ensemble deviations were generated for both methods and added to the model state values of temperature (initialised at 18 °C) to generate the ensemble members. An initial ensemble spread of 1 °C was specified for both methods. At analysis an additional forecast error was added to the ensembles by means of another set of correlated random fields with a standard deviation of 0.5 °C. The same fields were used for both simulations.

Table 1
Atmospheric stations used for truth and assimilation simulations

Data type	Truth simulation	Assimilation simulation
Air temperature	Point Wilson	Frankston
Atm. pressure	Moorabbin	Laverton
Precipitation	Melbourne	Moorabbin
Wind	Sth. Channel Island	Fawkner Beacon
Relative humidity	Frankston	Point Wilson

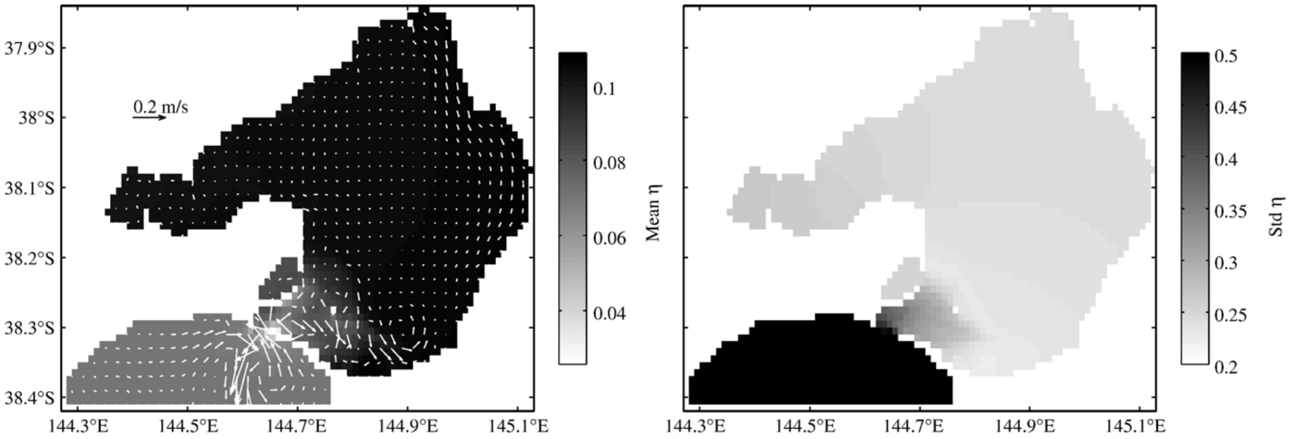

Fig. 2. Model results illustrating temporally averaged residual currents together with the mean and standard deviation of sea level η during the 2 year long model run.

A comparison of results from these two initialisation approaches are presented in Fig. 4 and show quite clearly that the proposed ensemble initialisation method gave superior results. The RMSE was calculated as the spatial mean squared difference between the truth and the ensemble mean. At the first analysis the RMSE of the proposed method rapidly reduced, while for the correlated random field method there was no corresponding reduction. Overtime, however the RMSE for both simulations tended to converge. The reason for this is that the hydrodynamic model is diffusive so that the impact of perturbations decay rather than grows overtime. Therefore, as the simulation progresses, the impact of the initialisation diminishes and the impact of model error becomes relatively more important.

6.3. Ensemble propagation

Model error is incorporated through the use of perturbed forcing data here. A justification for characterising the forcing data by data type is presented in Fig. 5. The plot contrasts two semi-restricted data types against two unrestricted data types. These are obtained by evaluating the spatial variability based on the data values at a number of atmospheric stations located around PPB. Individual points indicate individual time steps. As only two available stations record evaporation (panel b) the difference rather than the standard error is quoted. While the variability of the semi-restricted data types increases with value (panels a and c), the variability of the unrestricted data types appears constant over the main range of the data values (panels b and d). This illustrates the need to apply perturbations according to the error characteristics of the data, as well as allowing an estimate of appropriate magnitudes of perturbations and verifying which type of perturbation to use.

There are generally three distinct forcing data types used in hydrodynamic modelling: i) atmospheric forcing, ii) riverine forcing, and iii) open boundary forcing. Open boundary refers to the boundary between the model domain and the water body adjoining or surrounding it.

Atmospheric forcing has three functions within the model: i) momentum transfer via wind, ii) water level adjustment through changes in atmospheric pressure, and iii) heating and cooling of the water body through heat exchange with the

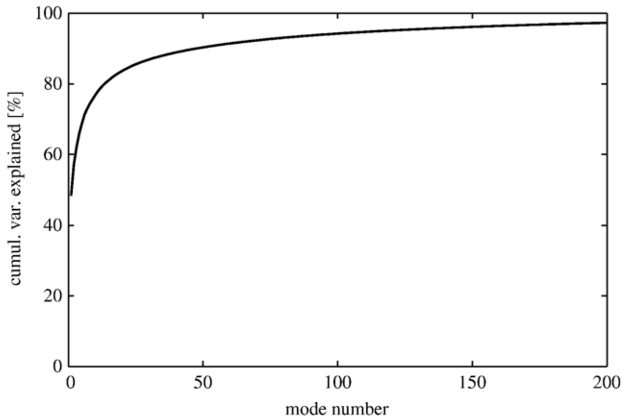

Fig. 3. Cumulative variance of the temperature in the PPB system explained by singular vectors. The number of singular vectors needed to describe a systems dynamics is a function of their relative singular values.

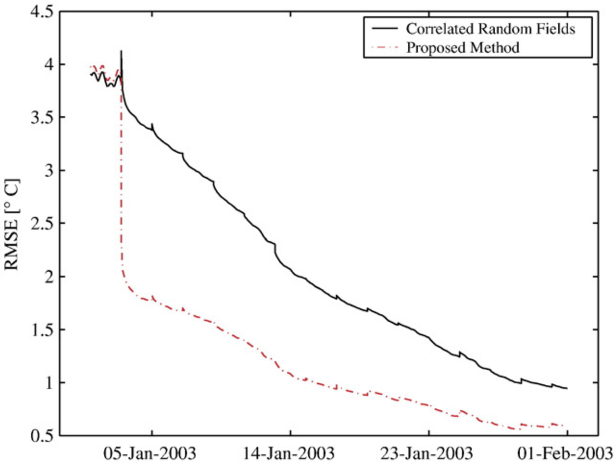

Fig. 4. Results of assimilation simulation testing the initialisation of a forecast ensemble. The method proposed in this paper is compared against initialisation using correlated random fields. RMSE of the two SST assimilation simulations is computed relative to a synthetic truth.

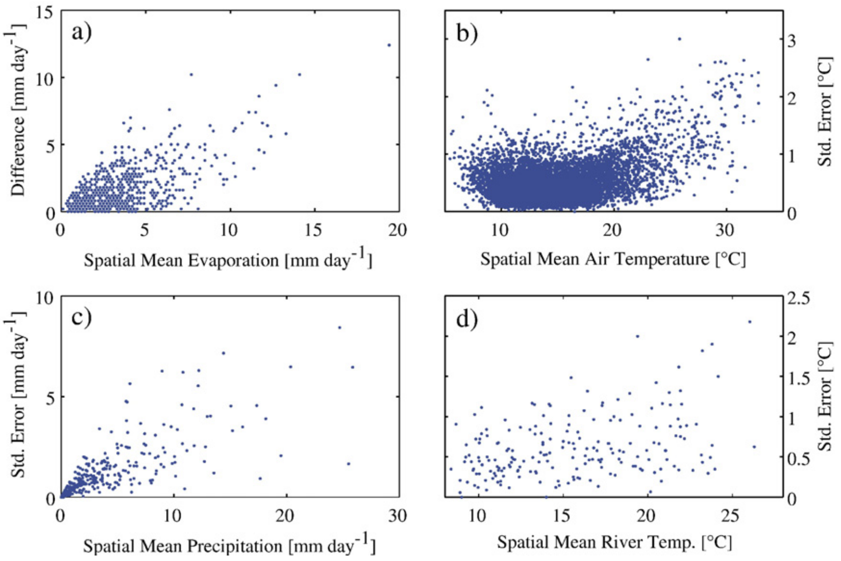

Fig. 5. Scatter plots showing spatial variation of a) evaporation, b) air temperature, c) precipitation, and d) river temperature.

atmosphere. Most of the data are related to the atmospheric heatflux calculations. Table 2 outlines the requisite data inputs for the hydrodynamic model together with a classification of its data type. Values of ξ and χ have been estimated from the spatial variations of data collected at various automatic weather stations located around PPB (Table 2). For example the gradient of the mid point line through the precipitation error value plot (panel c) has a slope of about 0.3 but this was reduced to 0.25 to limit the amount of variation at larger precipitation values, although possibly underestimating the error at lower precipitin values. For river temperature (panel d) the average error value is about 0.5 °C. Generally, the χ and ξ values were set equivalent, although for some data types different values were used to enhance the temporal correlation within the data. For air temperature the mean of the error over the standard range was about 0.6 °C, which was used for the χ value but the ξ value was set to 1.4 to gain more temporal variance within the ensemble and help account for the larger error at higher temperatures. River flow (in Table 3) is another example of varied ξ and χ values, where the larger χ value maintains the temporal pattern within the river flow data.

Table 2
Data types and values of ξ and χ for various atmospheric variables. The final column indicates the number of weather stations used to derive the ξ and χ values

Variable	Units	Data type	ξ	χ	No. of data sets
Air temperature	°C	Unrestricted	1.4	0.6	4
Wind vector	m s^{-1}	Unrestricted	2.5	0.7	8
Air pressure	Pa	Unrestricted	204	204	3
Precipitation	mm d^{-1}	Semi-restricted	0.25	0.25	4
Evaporation	mm d^{-1}	Semi-restricted	0.2	0.2	2
Short wave radiation	W m^{-2}	Semi-restricted	0.038	0.038	6
Relative humidity	%	Restricted	5.0	5.0	5
Cloud cover	oktas	Restricted	0.4	0.3	2

Riverine forcing data provide a source of fresh (or brackish) water to the model. As the river temperature is frequently different from sea temperature, rivers also act as a source or sink for temperature. The data type classification and best guess values for ξ and χ associated with riverine boundaries are summarised in Table 3. No ensemble was created for river salinity as the Yarra discharges fresh (zero PSU) water. For large saline estuarine systems however, a variable salinity boundary may be need.

Open boundaries predominantly control the momentum flux in hydrodynamic models by means of imposed sea level or velocity at the boundary. In this application a sea level boundary is used. If a model requires both prescribed elevation and normal boundary velocities these could be adjusted consistently using geostrophy. Open boundary conditions also control the temperature and salinity between the model and surrounding water bodies. These are also modelled through values imposed at the boundary. Referring to Table 4, the prescribed temporal error applied for salinity is zero. This is because a constant salinity boundary is being specified with the offset providing the variation between ensemble members.

Sea level data were treated separately to avoid introducing high frequency noise to the model. This is because the sea level data were recorded more frequently than other data sets: every six minutes rather than 3-hourly or daily. For the sea level data, normally distributed random numbers with a standard deviation of 0.05 m were generated at 12 hour intervals and a cubic spline

Table 3
Data types and best guess values of ξ and χ for riverine variables

Variable	Units	Data type	ξ	χ	No. of data sets
River flow	m^3 d^{-1}	Semi-restricted	0.05	0.2	N/A
River temperature	°C	Unrestricted	0.5	0.5	3
River Salinity	PSU	N/A	N/A	N/A	N/A

Table 4
Data types and best guess values of ξ and χ for various open boundary variables

Variable	Units	Data type	ξ	χ	No. of data sets
Elevation	m	Unrestricted	N/A	N/A	N/A
Water temperature	°C	Unrestricted	0.25	0.25	N/A
Water salinity	PSU	Unrestricted	0.0	0.5	N/A

was fitted through them. This produced a temporal correlation in the perturbation series which was then added to the original sea level time series.

6.3.1. Application

A sensitivity analysis has been performed on the values in Tables 2–4. The spread of water temperature predictions for ensemble members were most sensitive to open boundary water temperature, air temperature and short wave radiation values. However, the forecasts were fairly insensitive to changes of up to 50% in the χ and ξ values.

Fig. 6 shows some of the ensemble members generated from the original forcing set using the methods described here. These panels represent only a small time window of a longer time series and illustrate some of the details discussed above. As air temperature is unrestricted the range of the perturbed values (panel a) about the original value does not vary with time. For precipitation which has a lower boundary of zero, the generated perturbations are seen to increase with magnitude as the original precipitation values increase (panel b). Relative humidity (panel c) is a restricted data type (between 0 and 100) and the figure shows the magnitude of the applied perturbations reducing as the original value approaches the boundary.

To illustrate the effect of forcing data ensemble on model prediction ensembles, a 20 member ensemble run was made with each member using a different set of perturbed forcing values. In this case, members started from the same initial condition (16 August 2002). Three time series plots of water temperature are shown in Fig. 7 with the locations indicated in Fig. 1. These time series demonstrate the spread of ensemble members over time due to the perturbed forcing data alone.

The largest initial ensemble spread is observed at the open boundary (b) driven by the perturbed water temperature forcing at the open boundary. This spread is constant in time, constrained by the open boundary. Over time a gradual increase of variation in water temperature is observed throughout the bay (c). The effects of the perturbed atmospheric forcing data act more slowly than the open boundary. The spread grows according to the perturbations in the various atmospheric forcing data. The spread does not continue indefinitely but finds a limit based on the size of the perturbations. This spread due to atmospheric forcing was most pronounced at the edges of the bay and especially in the western arm (a). Thus the forecast error is not spatially uniform and varies depending upon the location. The variation will also depend on the state considered and the uncertainty of the forcing data type that dominates that particular state; i.e. the relative size of the forecast error may be significantly different for salinity or velocity.

The utility of perturbed forcing method described in this paper is demonstrated through a data assimilation experiment.

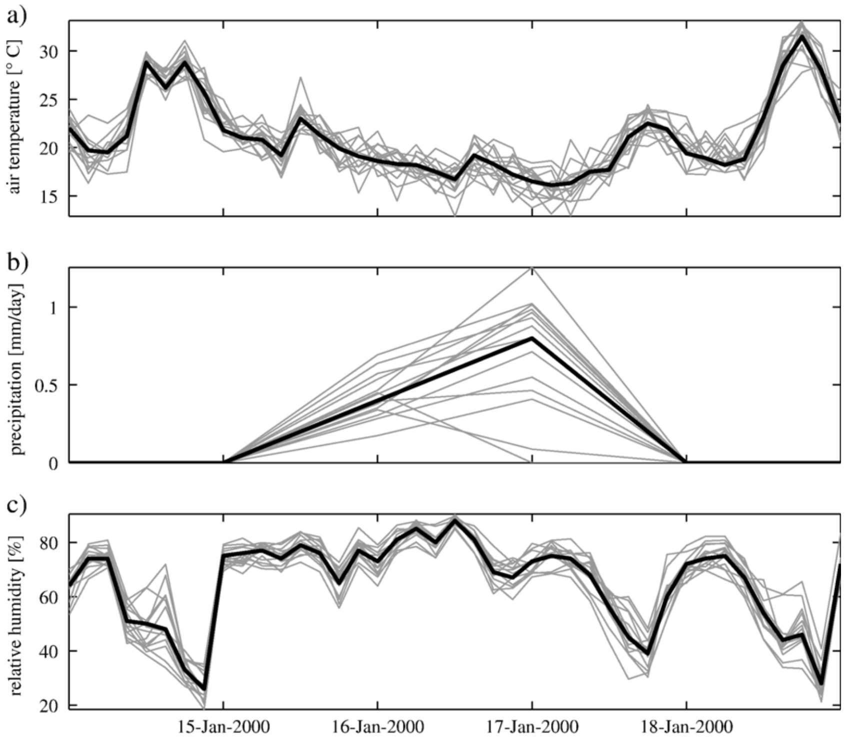

Fig. 6. Examples of forcing data field ensembles for a) air temperature, b) precipitation, and c) relative humidity using ξ and χ values from Table 2. Thin lines represent the ensemble and thick lines the original.

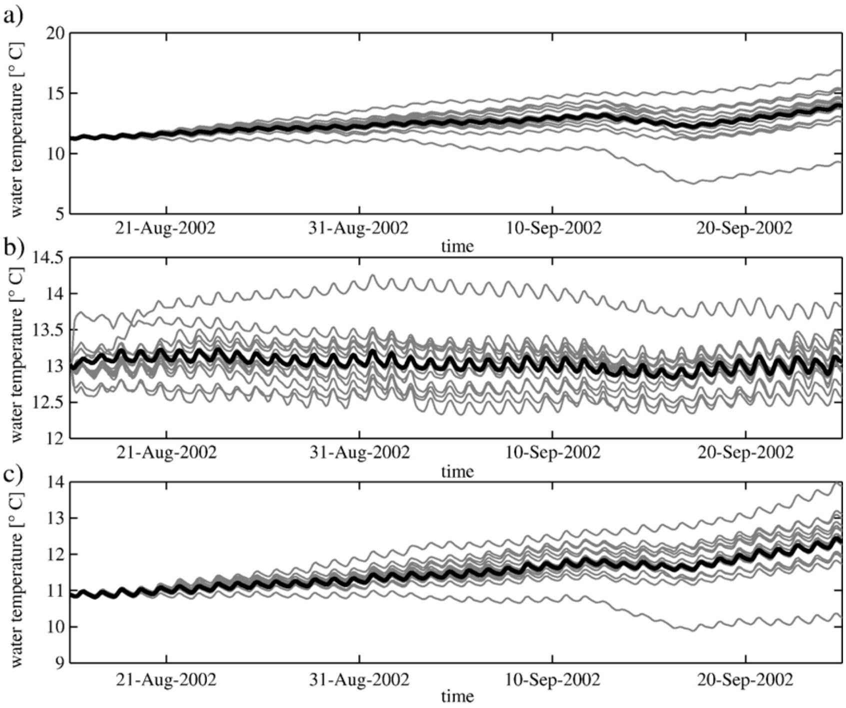

Fig. 7. Time series illustrating the effects of forcing data ensembles on model predictions of temperature at surface monitoring locations a) western arm, b) open boundary and c) centre of bay. Light lines represent the ensemble and dark lines the truth.

Here a simulation using perturbed forcing data method is compared against two other simulations: one where the same forcing data is used for each ensemble member but with random correlated fields added to the forecast at the analysis time, simulating the incorporation of model error, the other simulation used perturbed forcing data but with all of the data treated as an unrestricted type. This simulation allowed the possibility of biased forcing data to be introduced via the truncation of boundary exceeding perturbations. The correlated random fields were generated with a standard deviation of 0.5 °C. To adequately compare the perturbed forcing examples the ξ and χ values of the semi-restricted data needed to be adjusted. This was done by multiplying the existing values by the mean of the data value. The adopted values are listed in Table 5. The proposed ensemble initialisation methods described in this paper were used in all cases.

The results of the experiment are presented in Fig. 8. These results show that the assimilation with perturbed forcing has easily out performed the assimilation with correlated random fields and slightly out performed the assimilation using perturbed forcing with all data types treated as unrestricted. A reason for this is that introducing random correlated noise directly to forecast states at analysis time affects their dynamic relationships and possibly the numerical stability of the model. Introducing the error at the model boundaries, through perturbed forcing, alleviates this and allows the model to distribute the error dynamically through its domain. Also, (see Fig. 7) the error distribution for the perturbed method is spatially nonuniform whereas the random correlated fields impose a spatially uniform error. Thus forecast error would be

Table 5
Adopted values of ξ and χ for semi-restricted data types when treated as unrestricted to examine the impact of bias induced by boundary exceeding perturbations

Variable	ξ	χ
River flow	1.4	0.35
Precipitation	0.4	0.4
Evaporation	0.6	0.6
Short wave radiation	10	10

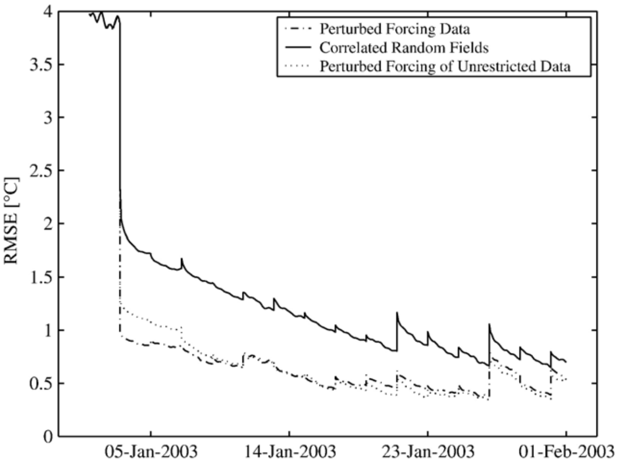

Fig. 8. Results of assimilation simulations testing the incorporation of forecast error. The use of perturbed forcing data generated by the procedure outlined is compared against adding forecast error using correlated random fields and against perturbed forcing with all data types treated as unrestricted. RMSE of the three assimilation simulations is computed relative to a synthetic truth.

underestimated at some locations and overestimated at others when using the correlated fields method. Furthermore, with the addition of a random correlated field, each forecast state element is perturbed by one random value introducing the possibility of sampling error. Using the perturbed forcing approach allows a state element to be influenced by a time series of random values, reducing the potential for sampling error to distort the results. This is particularly important where a small number of ensemble members are used. While the use of perturbed forcing data has clear benefits, the reduction of bias induced from boundary exceeding perturbed values by treating each variable according to its type (unrestricted, semi-restricted and restricted) also gave improvements over a simulation that treated all data as unrestricted allowing the introduction of bias from boundary exceeding perturbed values. This demonstrates the importance of attention to the unbiased generation of perturbed forcing data.

7. Discussion and conclusion

Use of an ensemble of model predictions to estimate model forecast error covariance is a common approach for nonlinear sequential data assimilation. While much attention in the literature has concentrated on the analysis, there is little direction given on the rigorous generation of the ensemble members and to the generation of ensembles of forcing data. This paper has described techniques to generate unbiased ensemble forcing data. A method of estimating appropriate ensemble size and generating domain spanning initial conditions was also introduced.

A framework for ensemble initiation and propagation that is applicable to practical data assimilation across a range of environmental fields has been outlined, and has been shown to achieve more accurate predictions and better estimates of error than standard approaches in the literature for an example application. The ensemble initiation uses the SVD of a long model integration to determine the optimal ensemble size and create an ensemble of independent model state conditions. Ensemble propagation uses an ensemble of model forcing data created using a perturbation and an offset term.

The importance of forcing ensemble members in data assimilation is that they can consistently and logically introduce model forecast error into an ensemble. However, the original forcing data used should provide a guide as to the appropriate levels of uncertainty for each particular forcing data type, and excessive addition of noise to the forcing data to generate larger model forecast uncertainty should be avoided. The values used in the example application for the ensemble generation framework presented here (Tables 2–4) are based on the observed variation of forcing data across the study area. The values adopted for a particular data assimilation exercise should likewise reflect the actual variation of the respective forcing data of that area. If large scale adjustment to the forecast uncertainty is required, both model structure error and forcing error should be explicitly included. Alternatively, the inflation factor method of Anderson and Anderson (1999) may be considered.

The methods presented in this paper apply to spatially uniform forcing data. While this is justified for the small area of application in the example given, a typical characteristic of many data assimilation applications, the method can be easily extended to account for spatially varying forcing. Extension to the two-dimensional case can be achieved by interpolating ensemble members at spatially distributed locations.

Acknowledgements

This work has been funded by a University of Melbourne-CSIRO Collaborative Research Support Scheme grant. The authors wish to gratefully acknowledge the following agencies for the use of their data; Melbourne Ports Authority, EPA Victoria, Bureau of Meteorology and Melbourne Water. Matthew Turner was supported by a Melbourne Research Scholarship. The suggestions of three anonymous reviewers is also acknowledged.

Appendix A. Choice of ξ values in semi and restricted data types

For the semi-restricted data type we require that the stochastic data point remain above a lower bound or below an upper bound. The following derivation of Eq. (16) is made by substituting Eqs. (13) and (15) into Eq. (10) for a realisation at the lower bound.

$$h_{\min} \leq h_k^o + (h_k^o - h_{\min})\xi z_k + (h_k^o - h_{\min})\chi z \tag{A.1}$$

$$\Rightarrow \frac{h_{\min} - h_k^o - (h_k^o - h_{\min})\chi z}{h_k^o - h_{\min}} \leq \xi z_k \tag{A.2}$$

$$\Rightarrow \frac{-1 - \chi z}{z_k} \geq \xi \tag{A.3}$$

and since $\xi_k > 0$, $z_k < 0$. For a given time series, χz is constant and has an expected value of zero. In which case Eq. (A.3) becomes

$$\xi \leq \frac{-1}{z_k}, \tag{A.4}$$

but the exact relationship still depends on the value of χz, and with this value being random, it is only known when the equation is applied. A similar derivation can be constructed for the upper bounded case.

Table A.1
The probability that a normally distributed random number z_k is less than a particular value as a function of ξ for semi-restricted and restricted variables

Semi-restricted ξ	Restricted ξ	z_k	Exceedence probability
1.0	$\frac{h_{\min} - h_{\max}}{2}$	−1	0.1587
0.5	$\frac{h_{\min} - h_{\max}}{4}$	−2	0.0228
0.33	$\frac{h_{\min} - h_{\max}}{6}$	−3	0.0014

As z_i is a normally distributed Gaussian random number, probabilities can be assigned to the possibility of z_k being less than a given value (Table A.1). Using the data in Table A.1 it can be seen that to reduce the probability of a domain exceeding value being generated to one in a thousand, ξ should be less than 0.33.

Similarly a derivation can be constructed for the restricted data type that limits values exceeding the restricted boundary. In this case the effect of the offset term has been ignored for simplicity, although as with the semi-restricted case it will have a small effect. Eq. (19) is derived by substituting Eq. (17) into Eq. (10)

$$h_{\min} \preceq h_k^o + \frac{h_k^o - h_{\min}}{\frac{h_{\max}+h_{\min}}{2} - h_{\min}} \xi z_k \tag{A.5}$$

$$\Rightarrow h_{\min} - h_k^o \preceq \frac{2(h_k^o - h_{\min})}{h_{\min} - h_{\max}} \xi z_k \tag{A.6}$$

$$\Rightarrow \frac{(h_{\min} - h_k^o)(h_{\min} - h_{\max})}{2_k(h_k^o - h_{\min})} \succeq \xi z_i \tag{A.7}$$

$$\Rightarrow \frac{(h_{\min} - h_{\max})}{2z_k} \succeq \xi. \tag{A.8}$$

Values of ξ can be associated with probability of exceedence values as indicated in Table A.1. These exceedence values refer to the probability of exceedence of one of these limits. The total probability to exceed either the upper or lower limit is twice as large.

References

Anderson, J. L., & Anderson, S. L. (1999). A Monte Carlo implementation of the nonlinear filtering problem to produce ensemble assimilations and forecasts. *Monthly Weather Review*, *127*, 2741−2758.

Blumberg, A. F., & Mellor, G. L. (1987). A description of a three-dimensional coastal ocean circulation model. In N. S. Heaps (Ed.), *Three-dimensional coastal ocean models* (pp. 1−16). American Geophysical Union.

Brusdal, K., Brankart, J. M., Halberstadt, G., Evensen, G., Brasseur, P., van Leeuwen, P. J., et al. (2003). A demonstration of ensemble-based assimilation methods with a layered OGCM from the perspective of operational ocean forecasting systems. *Journal of Marine Systems*, *40–41*, 253−289.

Evensen, G. (1994). Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *Journal of Geophysical Research-Oceans*, *99*(C5), 10143−10162.

Evensen, G. (2003). The Ensemble Kalman Filter: Theoretical formulation and practical implementation. *Ocean Dynamics*, *53*, 343−367.

Evensen, G. (2004). Sampling strategies and square root analysis schemes for the EnKF. *Ocean Dynamics*, *54*, 539−560.

Evensen, G., & van Leeuwen, P. J. (1996). Assimilation of geosat altimeter data for the Agulhas Current using the ensemble Kalman filter with a quasigeostrophic model. *Monthly Weather Review*, *124*(1), 85−96.

Gill, A. E. (1982). *Atmosphere–ocean dynamics* (pp. 297−322). New York: Academic Press Inc.

Hamill, T.H., Ensemble-based data assimilation: A review, *Unpublished manuscript*, University of Colorado and NOAA-CIRES Climate Diagnostics Centre, 2002.

Herzfeld, M., Waring, J., Parslow, J., Margvelashvili, N., Sakov, P., & Andrewartha, J. (2002, October). *MECO-model for estuaries and coastal oceans V4.0 scientific manual*, Ed. 1 Hobart: CSIRO Marine Research.

Houtekamer, P. L., & Mitchell, H. L. (1998). Data assimilation using an ensemble Kalman filter technique. *Monthly Weather Review*, *126*(3), 796−811.

Houtekamer, P. L., & Mitchell, H. L. (2001). A sequential ensemble Klaman filter for atmospheric data assimilation. *Monthly Weather Review*, *129*(1), 123−137.

Keppenne, C. L. (2000). Data assimilation into a primitive-equation model with a parallel ensemble Kalman filter. *Monthly Weather Review*, *128*, 1971−1981.

Miller, R. N., & Ehret, L. L. (2002). Ensemble generation for models of multimodal systems. *Monthly Weather Review*, *130*(9), 2313−2333.

Mitchell, H. L., & Houtekamer, P. L. (2000). An adaptive ensemble Kalman filter. *Monthly Weather Review*, *128*(2), 416−433.

Molteni, F., Buizza, R., Palmer, T. N., & Petroliagis, T. (1996). The ECMWF ensemble prediction system: Methodology and validation. *Quarterly Journal of the Royal Meteorological Society*, *122*(529), 73−119.

Nalli, N. R., & Smith, W. L. (1998). Improved remote sensing of sea surface skin temperature using a physical retrieval method. *Journal of Geophysical Research*, *103*(C5), 10527−10542.

Natvik, L. J., & Evensen, G. (2003). Assimilation of ocean colour data into a biochemical model of the North Atlantic — Part 1. Data assimilation experiments. *Journal of Marine Systems*, *40–41*, 127−153.

Pham, D. T. (2001). Stochastic methods for sequential data assimilation in strongly nonlinear systems. *Monthly Weather Review*, *129*, 1194−1207.

Reichle, R. H, McLaughlin, D. B., & Entekhabi, D. (2002). Hydrologic data assimilation with the ensemble Kalman filter. *Monthly Weather Review*, *130*(1), 103−114.

Reichle, R. H, Walker, J. P., Koster, R. D., & Hauser, P. R. (2002). Extended versus ensemble Kalman filtering for land data assimilation. *Journal of Hydrometeorology*, *3*, 728−740.

Robert, C., & Alves, O. Tropical pacific ocean model error covariance from Monte-Carlo simulations, to be submitted as a BMRC Research Report, Bureau of Meteorology Research Centre, Melbourne, Australia 25 August 2003.

Toth, Z., & Kalnay, E. (1993). Ensemble forecasting at NMC: The generation of perturbations. *Bulletin of the American Meteorological Society*, *74*(12), 2317−2330.

Toth, Z., & Kalnay, E. (1997). Ensemble forecasting at NCEP and the breeding method. *Monthly Weather Review*, *125*(12), 3297−3319.

Walker, S. J., Waring, J. R., Herzfeld, M., & Sakov, P. (2002). *MECO user manual.* Hobart: CSIRO Marine Research.

Zillman, J. W. (1972, August). Study of some aspects of the radiation and heat budgets of the southern hemisphere oceans, Bureau of Meteorology, Department of the Interior. *Meteorological Study*, *Vol. 26*.

Zupanski, M., Fletcher, S. J., Navon, I. M., Uzunoglu, B., Heikes, R. P., Randall, D. A., et al. (2006). Initiation of ensemble data assimilation. *Tellus*, *58A*, 159−170.