Quantitative methods for hydrological spatial field comparison

Stephen Russell Wealands

Submitted in total fulfilment of the requirements of the degree of Doctor of Philosophy

August 2006

Department of Civil and Environmental Engineering,
The University of Melbourne

Abstract

This thesis addresses the current lack of comprehensive, quantitative methods for comparing hydrological spatial fields. Comparison of spatial fields is needed for assessing hydrological models and for data assimilation. The methods that are currently used for quantitative comparison generally fail to consider the spatial arrangement of element values within spatial fields. Instead, there is a dependence on qualitative methods (e.g. visual comparison) to undertake comparison of many aspects (e.g. intermediate scale features), but such methods are non-repeatable, often biased and difficult to report on. This thesis advances the comparison methods available for use with hydrological spatial fields.

A detailed review of qualitative comparison and quantitative comparison methods used in other disciplines has been undertaken. The review identified the key aspects of comparison that were needed for hydrological fields. In addition to these aspects, existing comparison methods that could be adapted for hydrological comparison purposes were described. Three general notions emerged as being most important for comparison and have been pursued in the thesis: 1) importance; 2) tolerance; and 3) completeness.

Importance has been incorporated into new comparison methods by using data-driven (e.g. segmentation) and knowledge-driven (e.g. terrain analysis) pre-processing methods. These methods are used to modify a spatial field so that it only represents what is deemed important, which is then used for making subsequent specialised comparisons. Importance allows comparison measures to have a specific meaning and thereby reflect aspects of model performance that are application specific.

Tolerance has been implemented into new comparison methods by allowing spatially- and temporally-related elements to be considered as similar during comparison (unlike current methods that demand spatial coincidence). The notion of tolerance can also be thought of as accounting for uncertainty during comparison. The results of a tolerant comparison reveal only the significant differences (i.e. those larger than the tolerances) between two spatial fields, thus providing a measure of the true error (rather than a measure confounded by insignificant differences). During model assessment, this permits slightly different fields (e.g. with minor spatial shifts) to be evaluated as having equivalent performance, based on the requirements of the user.

By combining the current comparison methods with the new methods (i.e. importance and tolerance), an extensive suite of measures are made available. These measures are capable of quantifying a range of aspects and scales of spatial fields. Completeness is a general objective for comparison that depends on having access to such a range of measures. This thesis describes the process of devising a suitable comparison strategy that can achieve completeness by combining comparisons of complementary aspects and scales. This requires the user to plan the comparison and make clear decisions about what should be compared. Once made, the strategy can be applied to multiple fields and the results analysed with multi-criteria optimisation methods. This is a complete contrast to most current comparisons, which produce single results that are generally ambiguous (e.g. RMSE).

This thesis applied the new comparison methods within a customised comparison strategy to a real, hydrological modelling situation. The comparison results concurred with visual inspection and provided more detailed quantitative information than previously available. The unbiased and rigorous comparison of different aspects and scales produced clear, repeatable and reportable comparison findings that address the major weaknesses with current comparison methods. This bodes well for future application of the methods to larger, more complex fields and comparison tasks. At present, the hydrological community continues to develop its ability to observe and model spatial fields, but comprehensive model assessment is still in its infancy. The methods developed here provide a new suite of tools to the modeller and, if adopted, should significantly advance the art and science of spatial hydrological modelling.

Declaration

This is to certify that:

1. the thesis comprises only my original work towards the PhD;
2. due acknowledgement has been made in the text to all other material used; and
3. the thesis is less than 100,000 words in length, exclusive of tables, maps, bibliographies and appendices.

Stephen Russell Wealands

Acknowledgements

I take this opportunity to thank my primary supervisor, Rodger Grayson, for his assistance in focusing my research, keeping my eye on the bigger picture and his constructive feedback on this thesis. His well-timed bursts of encouragement helped to make things seem achievable. I greatly appreciate the structured and frank feedback from my other supervisor, Jeffrey Walker, which provided a balancing perspective on my research. I am also grateful to my overseas supervisor, Günter Blöschl, for the discussions and support afforded by him and his colleagues at TUWien, which helped to make my time in Austria very memorable.

I recognise the financial support from my Australian Postgraduate Award and Postgraduate Overseas Research Experience scholarships. I also appreciate the financial support given by the Cooperative Research Centre for Catchment Hydrology, the Department of Civil and Environmental Engineering and the Institute for Hydraulic and Water Resources Engineering (Vienna, Austria). I thank Andrew Western, Alan Seed and Andreas Güntner for the provision of spatial field data used in this research.

I've enjoyed the interactions with my enduring office mates – Matthew Turner, Christoph Rüdiger and Adam Smith – both the intellectual discussions and the frequent distractions. There are also many others from within the Department of Civil and Environmental Engineering that have made everyday life at The University of Melbourne enriching and many evenings at 'The P.A.' entertaining.

Finally, my wonderful wife, Karen, has been encouraging, accommodating and supportive whenever needed, despite dealing with her own thesis at the same time. I can't thank her enough. I've also had great companionship from Claude and Bernie during the many days and nights in front of the computer screen. And, as always, my family has been supportive of all my endeavours.

Table of contents

Chapter 1

Introduction

1.1 Importance of spatial field comparison

Hydrological spatial fields can be produced using spatially distributed models and/or a range of observation techniques (e.g. remote sensing, in-situ measurement). These fields are used to represent the spatial distribution and state of a specific hydrological attribute. Observed spatial fields, which represent reality (albeit with uncertainty), are increasingly available because of improved observation techniques and the widespread use of spatial data in natural resources management. Modelled fields, which aim to simulate reality, are also increasingly produced to help answer the 'spatial questions' that are posed by managers and researchers. This increased use of hydrological spatial fields has led to many new avenues of hydrological research. Much of the research is focussed on applications of different spatial fields, but there are important issues regarding the analysis of hydrological fields that still need to be addressed.

Observed spatial fields are used as simple data sources for constructing, parameterising and forcing hydrological models. These spatially distributed models are then used to simulate different hydrological or management-related attributes and their spatial arrangements (e.g. rainfall or land-use across a catchment) for a certain time period. The outputs from these models are often uncertain because of incomplete representation of the hydrological processes and poor definition of model parameters. Reducing this uncertainty is a major objective in hydrological research and requires comparisons to be made between the uncertain model outputs and observations of reality. Figure 1.1 illustrates the general process of spatially distributed hydrological modelling and shows that comparison is necessary for facilitating model assessment and/or data assimilation.

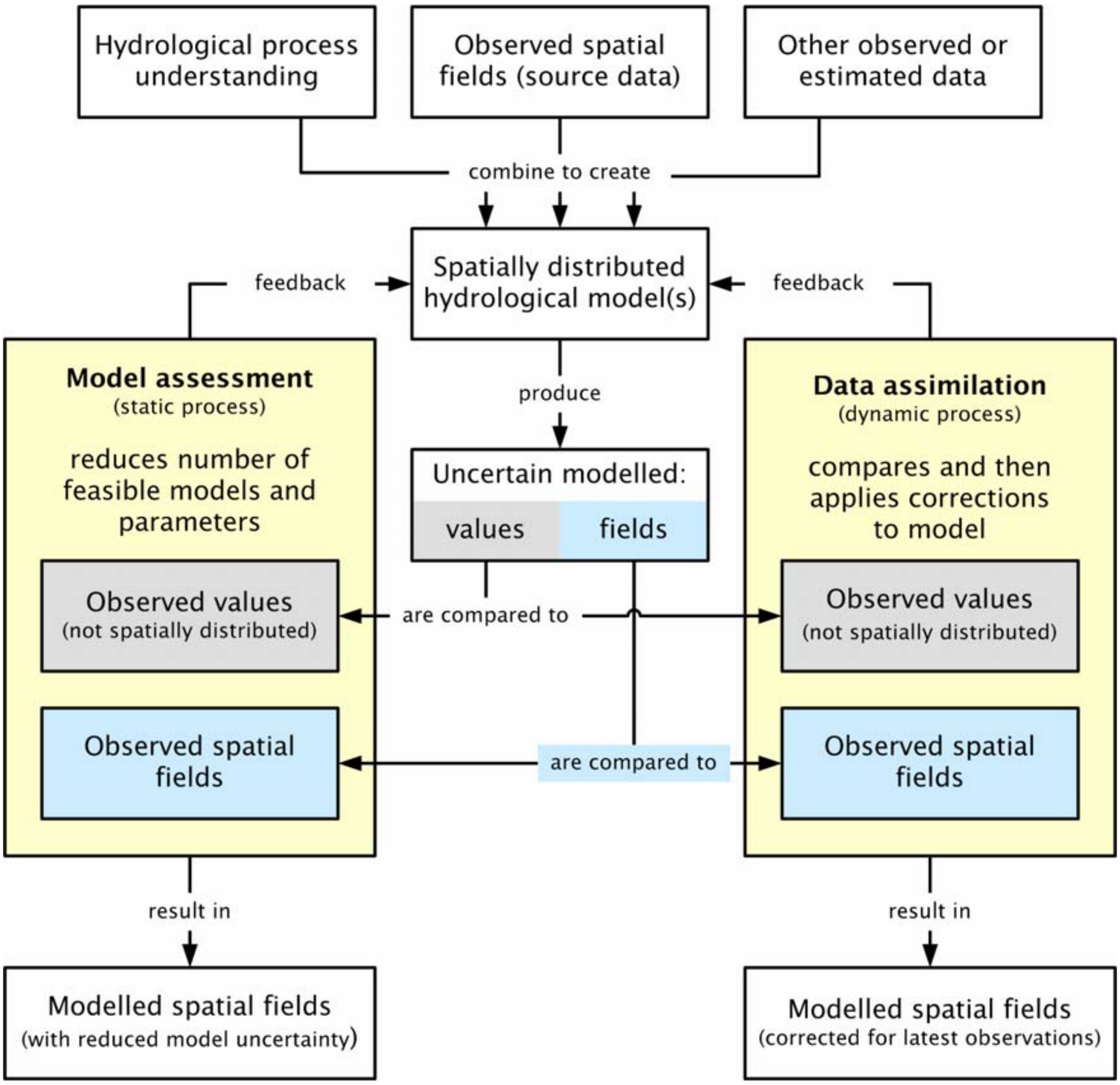

Hydrological process understanding
Observed spatial fields (source data)
Other observed or estimated data
combine to create
Spatially distributed hydrological model(s)
feedback
feedback
Model assessment
(static process)
reduces number of feasible models and parameters
produce
Uncertain modelled:
values
fields
Data assimilation
(dynamic process)
compares and then applies corrections to model
Observed values
(not spatially distributed)
are compared to
Observed values
(not spatially distributed)
Observed spatial fields
are compared to
Observed spatial fields
result in
result in
Modelled spatial fields
(with reduced model uncertainty)
Modelled spatial fields
(corrected for latest observations)

Figure 1.1 A simplified flowchart of the hydrological modelling process, specifically showing the roles for spatial fields. This thesis focuses on the key task of comparing observed fields with modelled fields (shaded blue) to provide the necessary information to facilitate model assessment and/or data assimilation (shaded yellow).

In most current hydrological studies, comparison is only undertaken between values representing an integrated hydrological response (e.g. a runoff hydrograph) and observations of a comparable attribute. Comparing integrated responses does not explicitly test the internal processes occurring in the model (Grayson et al. 1992), which are generally the most important outputs from a spatially distributed model. If these processes are not explicitly tested, a model may be producing the right answers for the wrong reasons. Observed spatial fields can explicitly represent these internal processes. Therefore, the addition of spatial field comparisons produces a more comprehensive test of model performance (Grayson and Blöschl 2000b; Jetten et al. 2003). Apart from situations where there are obvious constraints on spatial arrangement (e.g. a topographically-controlled flood inundation field (Horritt and Bates 2002)), spatial fields can help to avoid situations where a model produces the right answers for the wrong reasons (Klemeš 1986; Grayson et al. 2002).

Spatial field comparisons (i.e. comparison of modelled to observed field) are important for comprehensive model assessment and for data assimilation. Model assessment includes tasks such as model calibration, testing and understanding. For calibration and testing, quantitative comparison methods can be used to evaluate which models (or parameter sets) are acceptable, thereby constraining the range of feasible (or optimal) models. The observed spatial fields provide strong constraints on which models are considered feasible and are valuable for reducing model uncertainty, particularly when applied in an uncertainty framework (e.g. Beven and Binley 1992; Franks et al. 1998; Hunter et al. 2005) or multi-criteria optimisation (Gupta et al. 1998; Boyle et al. 2000). Both quantitative and qualitative spatial field comparisons are also used for understanding model performance by revealing where a model has errors and similarities, which can then be analysed by experts to understand the reasons why. Data assimilation can also benefit from the results of quantitative comparisons, which can be used to help determine the location and value of the corrections applied to different model states (e.g. Houser et al. 2000; Pauwels et al. 2001). These corrections adjust the model states so that it can produce modelled fields that are more similar to reality (based on the previously assimilated observations).

1.2 Statement of problem

Spatial field comparisons are already undertaken within hydrological modelling, although the current quantitative comparison methods used are not very comprehensive. These methods ignore the spatial arrangement within the spatial fields, instead treating each spatial field element independently. While the current methods have some benefits (e.g. simple, exact), there is potential for more advanced comparison methods to provide more useful information about error and similarity. Qualitative, visual comparison is commonly used to note similarities between spatial fields (i.e. using the relationships or patterns of arrangement). This is a powerful and important method, but the inability to quantify, repeat and report on such a comparison makes it unsuitable for operational use (e.g. in semi-automated uncertainty frameworks). These comparison methods are the current state-of-the-art in hydrology, yet they cannot comprehensively describe how spatial fields compare. The weakness of these methods has been recognised as a limiting factor in the uptake of spatial field comparisons in hydrological modelling (Grayson and Blöschl 2000b) and needs addressing.

There are other disciplines, such as image processing, pattern matching and landscape ecology that have been researching the problem of spatial field (or image) comparison for far longer than hydrology. Research in these disciplines has recognised that there are no "silver bullets that will solve a large range of general [comparison] problems" (Pavlidis 2003). However, the research does offer an existing collection of methods that may be applicable and/or adaptable to the comparison of hydrological spatial fields. The experiences from these other disciplines, as well as an understanding of the requirements for hydrological comparisons, can be used to unearth alternative methods that can address the weaknesses of the current comparison methods used.

1.3 Research objectives

The overall objective of this thesis is to advance the methods used for quantitatively comparing spatial fields in hydrological modelling. Within this objective there are a number of key research questions that need to be addressed:

- How are spatial fields currently used in hydrological modelling? Are there any problems or issues with their usage?
- When comparing spatial fields, what aspects of error or similarity can and cannot be quantified with the current methods?
- How can the quantitative comparison methods used in hydrology and other related disciplines be improved to quantify additional aspects of error or similarity?
- How do the new and improved quantitative comparison methods improve model assessment processes in hydrological modelling?
- How can the user decide which comparison methods are most suitable and how they should be applied for a given hydrological situation?

1.4 Scope

This thesis does not aim to define the comparison methods that are suited to every different hydrological situation. Nor does it aim to define what the 'best' comparison method is – this is a subjective decision that depends both on the needs of the user and the spatial fields being compared. Instead, this thesis aims to develop a range of quantitative, deterministic comparison methods that are widely applicable to hydrological fields and highlight the pros and cons of each. All of the methods should be considered for use in a semi-automated manner, where it is intractable for the user to visually inspect all of the fields.

The methods do not aim to remove the need for user input into the comparison process. Instead, they aim to encourage being explicit about the subjective decisions that help to define suitable and meaningful comparisons. All comparison methods require subjective decisions to be made, despite the fact that they are often referred to as objective measures within hydrology.

These comparison methods are developed and discussed for use in hydrological modelling, particularly for facilitating model assessment and data assimilation, although they also have application for model understanding. The term 'model' should be interpreted to be any modelled field, whether modelled using a distributed hydrological model or by an interpolation method. It may even refer to any field that is being compared against another field taken as the reality (e.g. assessing a remotely sensed field using point-based ground observations).

Finally, the spatial fields used for examples and analysis in this thesis are all regular-shaped element fields (i.e. rasters). These are used primarily because of their widespread availability and familiarity. However, the new methods presented in this thesis are described and can generally be implemented for other types of spatial fields (e.g. irregular-shaped elements) also.

1.5 Thesis outline

This thesis proceeds (0) by defining spatial fields and their general use in hydrology. Essential background information is given regarding spatial observations, spatial models and the various issues with their use. This background defines the context in which comparison methods are currently used and the tasks that they facilitate.

An extensive literature review is then undertaken (0), examining how spatial field comparison is done qualitatively and quantitatively within hydrology and many other disciplines. The research from other disciplines helps to guide the improvements made to current hydrological comparison. This review brings to light many analysis and processing methods that are valuable for hydrology and is an important contribution. From the review, three major themes emerge as being most important during comparison and they provide the basis for the methodology.

The major themes of tolerance, importance and completeness are incorporated into some new comparison methods for use in hydrology. The methods from other disciplines are adapted, applied and improved for use with the existing methods and hydrological spatial fields (Chapter 4). Details of the algorithm and usage for each method are given. The methods are tested using synthetic fields (with known deformations) to confirm their capabilities. A comparison flowchart and strategy are presented to show how these methods can combine together for use in specific applications, such as different types of hydrological model assessment.

Chapter 5 is a practical application of the comparison methods to a real hydrological situation. The hydrological example used is the Tarrawarra data (Western and Grayson 2000), in which modelled and observed spatial fields of soil moisture exist. These fields have been previously compared by experts using the current hydrological methods, thus providing a benchmark against which the performance of new methods can be evaluated. This chapter devises a comparison strategy for model assessment at Tarrawarra, then applies it and evaluates the results.

The discussion (Chapter 6) draws on evidence presented in the practical application (Chapter 5) to describe how subjective user decisions combine with the general methods to create specialised comparison measures. An example comparison strategy is devised for a different hydrological situation, thus demonstrating how the methods can be adapted to the needs of the user. Attention is also given to issues regarding analysis (e.g. significance testing) and implementation (e.g. computational demands). A software comparison tool is also described (see Appendix A) and the reader is encouraged to experiment with this tool to better understand the methods.

The conclusions are presented in Chapter 7, summarising the contributions made in this thesis and their expected impact on the field of hydrological modelling. The thesis then concludes with a discussion about the directions for further research on this topic.

1.6 Definitions

The following table of definitions is given to clarify the way certain terms are used throughout this thesis. There is no common language used in the literature when referring to spatial fields and their processing. This table defines a standard set of terms relating to spatial fields, as well as defining other ambiguous terms from hydrology (e.g. calibration, testing).

Table 1.1 Definitions for all major terms that are ambiguously used in hydrological literature or have not previously been clearly defined.

Term	Definition
Spatial field	a collection of associated elements with defined topology (i.e. the physical or logical relationship between elements) that are representative of a certain attribute at a certain time (i.e. an instance or duration)
Spatial pattern	a synonymous term for a spatial field that is avoided in this thesis because it implies some amount of trend or structure within the field (which may not exist)
Element	the building block of a spatial field; has a defined location (i.e. an area) and value (i.e. one attribute)
Element value	a value representing the magnitude (i.e. interval or ratio number) or type (i.e. ordinal or nominal category label) of the attribute being represented
Continuous field	a general grouping for any field representing an attribute with interval or ratio values (e.g. real numbers)
Categorical field	a general grouping for any field representing an attribute with ordinal or nominal category labels (including binary)
Spatiotemporal series	a series of spatial fields that are representative of a certain attribute over a range of times, usually with a regular interval between each time step or field
Observed spatial field	a field produced from observations that represents an attribute during the observation time (e.g. rainfall on a certain day); some processing may be used to convert raw observations into the observed field, but this is considered as part of the observation process (even though it is formally a type of model output)
Modelled spatial field	a field produced from model output (i.e. via a defined algorithm), such as that produced by a distributed hydrological model
Hydrological model	an algorithm representing real hydrological processes that is applied to a particular domain (using parameters and forcing data); the model algorithm, parameters and forcing data combine to produce modelled spatial fields

Term	Definition
Model calibration or optimisation	the process of refining the model algorithm or parameters so that the model output(s) best represent the calibration data set (e.g. observed time series or fields)
Model testing	the process of testing the capabilities of the model using independent observed data; often referred to as model validation
Model assessment	a holistic term to describe the process of understanding model outputs via comparison with observed data; may encompass calibration, testing, analysis, etc.
Comparing spatial fields	the process of using quantitative comparison methods to obtain comparison measures of the error and/or similarity between two spatial fields
Local scale comparison	a comparison method that uses the numerical relationship between each pair of elements having a certain spatial relationship (e.g. spatially coincident) to determine an field of local error or similarity, before summarising that field into a single comparison measure
Global scale comparison	a comparison method that uses a characteristic of multiple elements from each field, rather than the numerical relationship for each individual element in computing the comparison measure
Comparison measure	the numerical result of a comparison method being applied to a field or a pair of elements; interpreting this measure correctly requires understanding of the method used and fields being compared; often referred to as an objective measure or objective function in hydrology
Intermediate measure	usually refers to a field of local comparison measures, such as is created during local comparison; can also refer to a table or graph that is an intermediate stage of processing during a comparison method; intermediate measures are useful for visually analysing the contributions to a final comparison measure
Error measure	a comparison measure that uses the numerical difference (or residual) between field elements in its calculation; the measure is expressed in same units as the source field; interpreted relative to known characteristics of the source field; accumulates all errors and gives a measure of 'how badly' the fields match (i.e. a penalty-based comparison); often referred to as an absolute measure
Similarity measure	a comparison measure that uses the numerical relationships between field elements and also a certain reference (e.g. observed mean, linear fit) in its calculation; produces a measure with a limited range, commonly bounded by zero and one; value of one denotes perfect similarity, while zero can denote no similarity or similarity equal to the reference (depending on method); only elements that are similar contribute to the resulting measure; gives a measure of 'how well' the fields match (i.e. a reward-based comparison); often referred to as a relative measure
Intra-comparison	the process of analysing measures obtained from using one comparison method with one set of related spatial fields (e.g. many modelled fields compared to one observed field); the measures can be interpreted relative to one another or to a certain reference
Inter-comparison	the process of analysing measures obtained from using one comparison method with many sets of spatial fields (e.g. with different locations, times and/or attributes); the measures must be interpreted relative to a certain reference or otherwise analysed as ranks to be inter-comparable

Chapter 2

Background for spatial fields

2.1 Chapter overview

Spatial fields are observed and modelled using a range of techniques within hydrology. Observations can be made in-situ (i.e. on the ground) using techniques that measure hydrological responses at a point (e.g. a rain gauge) or over an integrated area (e.g. a stream gauge). In-situ measurements are made at multiple locations to represent the spatial variability of the response. Observations can also be made using remote sensing techniques that measure different aspects of hydrological response (e.g. RADAR, surface temperature). When measurements are available for multiple locations and the locations are spatially related, an observed spatial field is produced.

Spatial fields are also produced by hydrological models, which use equations to calculate and distribute hydrological quantities across an area. The modelled values (e.g. surface runoff, soil moisture percentage) are produced for all locations within the area and usually for multiple times. Comparison of the modelled fields against an observed field is the focus of this thesis, although understanding the meaning of spatial fields is necessary first.

This chapter provides the background on spatial fields in hydrology, beginning with a definition of scale and spatial fields. A number of observed and modelled fields from hydrology are then shown to illustrate their typical application. Issues with observations and models that the reader should be aware of are then discussed, as these underlie the research questions tackled in this thesis.

2.2 Understanding spatial fields

2.2.1 Defining scale

The term 'scale' is used in many contexts and disciplines, however, in the Earth sciences it refers to the spatial and/or temporal dimensions of an object or process. Along with the term scale comes other terms, such as grain, extent and resolution that are all sometimes used to refer to scale or some aspect of it. In the hydrological literature, scale is usually defined using the 'scale triplet' (Blöschl and Sivapalan 1995). This triplet defines the spacing, extent and support of an element or process in space and time (Figure 2.1). The spatial and temporal scales of the elements that make up a spatial field define what the field truly represents, and subsequently what it can be used for.

The spatial 'spacing' is the distance between neighbouring elements, spatial 'extent' is the bounds of all elements in the field and spatial 'support' is the area each individual element represents. For a single spatial field, the temporal spacing is undefined and the temporal extent and support are the duration (or instant) that the field represents. If a spatiotemporal series exists (e.g. model outputs), the temporal spacing is the time in between output of the fields, the temporal extent is the total time period between the first and last fields and temporal support is the duration (or instant) each field represents.

For example, a spatial field of near-surface soil moisture obtained from time domain reflectometry (TDR) samples, the scale triplet could have values of 20m spatial spacing

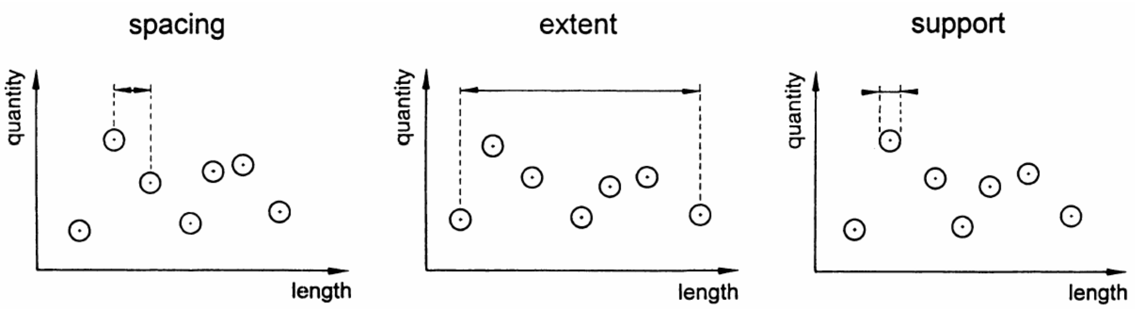


Figure 2.1 Schematic of the scale triplet, which can be defined for spatial and temporal scales (from Blöschl and Grayson 2000). The scale of a spatial field defines what it truly represents and therefore the applications it can be used in.

(the sampling interval), 200m spatial extent (the sampled area) and 10cm spatial support (the average diameter that influences a TDR measurement) (Blöschl and Grayson 2000). The temporal extent and support of a single TDR-derived field could be 12 hours (the time taken to sample all locations in the field), although that would only be true if stationarity were assumed. If this same field was created from by an airborne passive microwave sensor, the values could be 50m spatial spacing, 2000m spatial extent and 70m spatial support. The temporal extent and support would be almost instantaneous (or the time taken for the sensor to scan the area). In both examples, the fields represent near-surface soil moisture (albeit over different depths), but are measured in different ways. By defining the scale triplet whenever data are being managed, a clearer understanding of the true meaning of the data is obtained that is essential when working with models.

Comparison between spatial fields can only truly be done when the fields being compared represent the same thing. Otherwise, scale differences in space and time can lead to detection of false differences or similarities between fields. This is true from a conceptual point of view, but in practice this constraint would preclude almost any comparisons being used in modelling. Some models produce fields and series with scale characteristics that are not reproducible by any observation methods. Instead, scale differences should be understood and managed to limit their influences. For example, spatial fields with matching spatial scales but different temporal scales (e.g. instantaneous versus daily) may be comparable for some purposes. Spatial fields with different spatial supports are also often treated as being comparable (e.g. a rain gauge to a RADAR rainfall element). The scale triplet provides a simple framework in which the user can understand scale differences. Once understood, the differences can sometimes be mitigated by pre-processing the data (i.e. upscaling or downscaling). In practice, scale differences are often overlooked and ignored, but this can lead to false or unclear findings when comparing data sets.

2.2.2 Changing spatial scale

A change in one or more components of the scale triplet results in a significantly different representation of the elements in a spatial field. The impact of each component of the scale triplet can be seen in Figure 2.2. If the support is too large, each

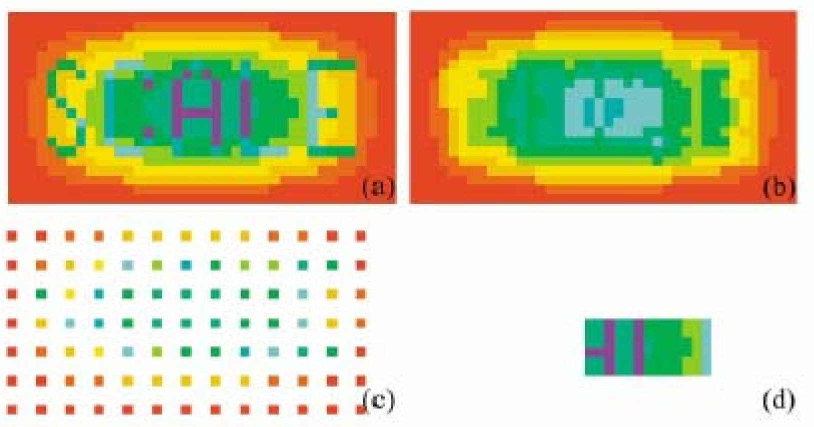


Figure 2.2 The effect of changing each component of the scale triplet for a spatial field: (a) true field; (b) the effect of increasing support; (c) the effect of increasing spacing; and (d) the effect of decreasing extent (from Western et al. 2002).

element value will be an integrated response that smoothes out smaller scale variability (Figure 2.2b). If the spacing between elements is too large, the local variability may not be represented at all (Figure 2.2c). If the smaller spacing (as in Figure 2.2a) is used, then the extent that can be represented may be reduced due to time or money constraints (Figure 2.2d). This will manage to represent local variability, but fail to represent the variability existing for the rest of the study area. This is true for any kind of spatial field, whether produced using in-situ or remote techniques.

Remotely sensed fields often suffer from the situation shown in Figure 2.2b, where they provide a continuous coverage of a large area but have too large a support, thus averaging the small scale variability. Hydrological models can also suffer from this problem, although they can be run using computational elements with smaller support (provided the small scale processes can be represented). Measurements taken in-situ are more often like Figure 2.2c, although they are commonly irregularly spaced. They usually have small support and large spacing, leading to an intermittent coverage of the area that can miss small scale or local variability. Figure 2.2d shows the situation that exists when high-resolution imagery is obtained but is too expensive for complete coverage of the area. A comprehensive field is obtained that represents that local area well, but fails to represent the remaining area.

For comparison methods that deal with spatial fields from different sources (e.g. from models and independent observations), it is essential that scale inconsistencies be recognised and managed as best possible. Some techniques exist that can be used to effectively change scales, but all assume certain relationships between scales to achieve this. Western et al. (2002) presents a review about scaling soil moisture that identifies various approaches to changing between scales with common hydrological data. “The essence of successful scaling is to distil the key patterns from information at one scale and to use these to make good predictions at another scale” (Western et al. 2002, p.151). Changing scale involves manipulating one or more components of the scale triplet to produce a new version of the data with a different characteristic scale. This can require interpolating the values in between or around observed elements using the relationships within the original data (e.g. Figure 2.2c, d to Figure 2.2a). Another common situation is where the data has large support (e.g. an areal average) and this is disaggregated to show the variability at a smaller scale (e.g. Figure 2.2b to Figure 2.2a). This task uses specific rules or statistical relationships to estimate the small scale element values. More commonly, data are aggregated (or averaged) from a finer-scale to produce a coarser scale representation.

There is no universal theory of scaling that permits data to be easily produced at many different scales, which could be used to make observed and modelled fields more comparable. In hydrology, interpolation and extrapolation methods that use the statistical relationships within and amongst spatial fields are used for changing the spatial scale of spatial fields (where the arrangement is important) (Western et al. 2002). These methods use the spatial correlation within a single field or correlation between fields (e.g. relationship between soil moisture and terrain) to describe the variance in space. The variance is a function of distance, with nearby elements usually having similar values and those further away being more variable. Where the extent of a field is to be made greater, the relationships with other fields that cover the larger extent (e.g. terrain models) are used to facilitate extrapolation. The relationships between or within fields are used to define the parameters for interpolation methods such as ordinary kriging or co-kriging (Journel and Huijbregts 1978), which are used to distribute these relationships in space. Numerous interpolation methods exist that use different approaches to distributing the random variables (with known statistical structure) in space. A more technical description of the way spatial correlation is characterised is given in Chapter 3 and more details on estimation methods can be obtained in Isaaks and Srivastava (1989).

Understanding scale relationships in hydrology is a continuing area of research, which can potentially provide better methods for changing the scale of particular spatial fields. Scaling relationships are also used in hydrological models for representing sub-element variability. Further developments are expected with increased data availability, particularly where the same attribute is being measured by sensors at multiple scales (e.g. different resolution remote sensing platforms).

2.2.3 Spatial field structure

The definition of a spatial field refers to a set of associated elements, where each element contains a value and a location. Elements are related to one another via a specified topology, which may or may not be inherent to the field structure. Elements are the building blocks of spatial fields and determine the overall appearance of the field. Most studies in hydrology deal with one of three different types of spatial fields, which have quite different visual appearance (Figure 2.3). The type of field structure is usually determined by the methods used to create it, but the structure can be changed

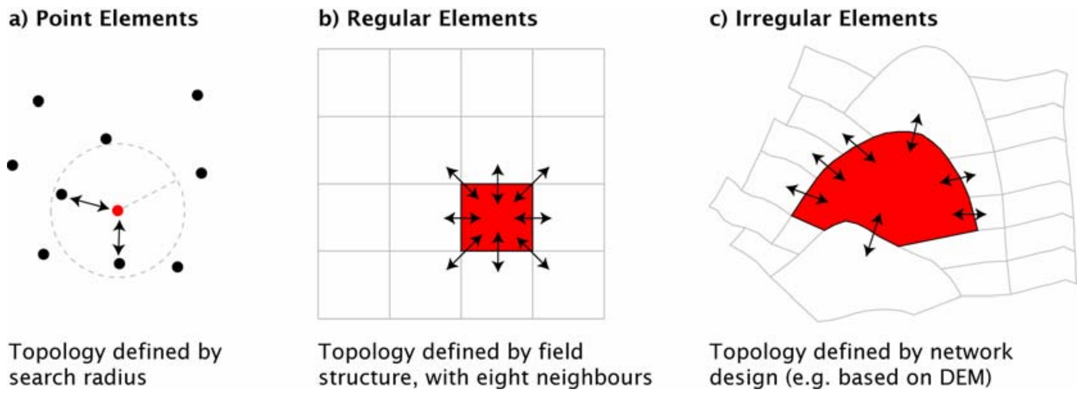

a) Point Elements
Topology defined by search radius
b) Regular Elements
Topology defined by field structure, with eight neighbours
c) Irregular Elements
Topology defined by network design (e.g. based on DEM)

Figure 2.3 Spatial fields are made up of elements (shown in red), each with a value and location for which the value applies. The relationship with other elements (shown by arrows) is defined in various ways, depending on the element type. The examples shown are similar to situations in hydrology: a) shows rain gauge observations, with nearby gauges considered to be related; b) shows a remote sensing observation, where neighbouring elements are related; and c) shows a modelled field based on contours and flow lines, with pre-defined relationships (e.g. using an elevation model).

with different types of pre-processing. The field structure determines how an extent is discretised, or more specifically, how the location that a value represents is defined in space.

Observed fields collected by in-situ measurements are usually point elements (Figure 2.3a), which have spatial relationships defined in some manner. When known relationships exist between the point elements, these fields are usually processed into a regular element form (Figure 2.3b), using pre-processing or scaling methods. Point elements sometimes have a fairly sparse coverage (i.e. with large spacing), so the relationship between nearby points is used to reduce the spacing and more clearly represent the spatial pattern between the point elements using regular elements (Grayson et al. 2002). If the relationships between sparse points cannot be defined, they should be treated independently (i.e. no topology) and therefore not compared as a field (but can be compared as independent points). Remotely sensed measurements are provided as regular elements on a regular grid. Due to the continuous representation of a certain extent with remotely sensed measurements, each element is considered connected to all of its immediate neighbours. This field structure is also used by many hydrological models, which produce modelled fields with the same structure. The model equations may allow connectivity with either the four-way or eight-way neighbours, depending on the model. This is the most common type of field used in hydrology (Figure 2.3b). It can be thought of as synonymous with an image or a raster, made up of pixels or cells respectively.

Irregular elements are also used frequently as model structures, where the element boundaries are defined with contours, flow lines or other topographic features (Figure 2.3c). Topology is defined within the hydrological model, which is commonly based on a terrain model (with upslope and downslope connectivity). Some observed spatial fields also use an irregular element structure (e.g. soils maps), but they are often processed back onto a regular structure to be consistent with other data sets. Therefore, by using some assumptions and/or scaling relationships the elements can be changed between different structures. This involves some form of interpolation, with methods such as splines (Hutchinson and Gessler 1994) and ordinary kriging (Journel and Huijbregts 1978) being most popular. Other methods such as Thiessen polygons

(Dingman 2002) are used for certain situations, such as when a point element is assumed to represent all locations it is the nearest to. Interpolation methods facilitate making fields with different structures comparable, although they rely on making assumptions about the relationships between elements.

Due to the widespread use of regular element fields, they are the type used throughout this thesis. Most of the comparison methods presented can also be applied with the other field structures, albeit with minor modifications in some cases. When working with multiple spatial fields and comparing them, the fields should be of the same type. This follows from the earlier discussion about scale consistency between fields during comparison. If the scale and structure cannot be matched via pre-processing or assumed to be equivalent, then comparison between the spatial fields should not occur.

2.2.4 Spatial field values

Spatial fields can have either continuous or categorical values stored for each element. The type of value stored is usually determined by the measurement method. For example, observed fields of soil moisture produced from TDR measurements will have continuous values of percentage moisture content (Figure 2.4a). Other fields can have categorical values that have a hierarchy of values (i.e. ordinal values), such as erosion mapping with low/medium/high categories (Figure 2.4b). Categorical fields of land use assign a nominal (i.e. named) category to each element. The simplest type of value is binary categorical, in which elements can only have one of two values, such as saturated and unsaturated areas (Figure 2.4c).

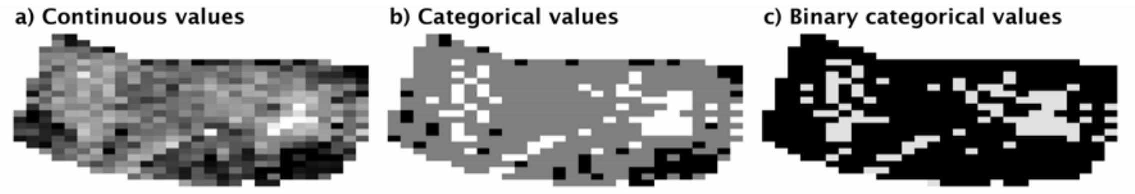


Figure 2.4 A spatial field represented with three different element value types: a) shows a continuous field of soil moisture with moisture content from 22% to 46% (dark to light); b) shows a categorical field, with soil moisture recorded as low/medium/high (dark to light); and c) shows a binary categorical field with saturated (light) and unsaturated (dark) areas.

Continuous spatial fields contain the values with the most information. Continuous values can be readily simplified into categorical values if needed, using various assumptions and pre-processing. However, due to the quantitative nature of hydrology, continuous fields are usually used without much simplification. In land use modelling, continuous fields (containing reflectance values of multiple wavelengths) are almost always simplified into a categorical field (i.e. a classified land use map). Categorical fields allow boundaries to be defined between values that are close together, which produces contiguous regions with the same value. These regions (or features) open up alternative analysis options that are fundamental in disciplines such as landscape ecology. There are few situations in hydrology where there is a need to simplify continuous fields as it leads to information loss, although for visualisation this is often undertaken (e.g. using categorised colour ramps).

Hydrological spatial fields are represented by continuous values wherever possible. Some examples are soil moisture percentage, precipitation amount, flood level or snow depth. However, there are frequently situations where the measurement methods cannot capture the hydrological attribute sufficiently to provide continuous values. Jetten et al. (2003) mapped low/medium/high erosion due to difficulties with accurately measuring the volume of erosion from an area. In flood hydrology, there is no feasible way to obtain water levels at every location during a flood. Also, snow cover is often derived from reflectance values in remotely sensed imagery. One solution is to sacrifice the detail of the value to cover a greater extent. Remote sensing can be used to record the limits of the flood water inundation, thus providing a binary categorical field that can be used during modelling (Bates et al. 2004). The other option would be to measure the water level as a continuous value at fewer locations. Blöschl and Grayson (2000) refer to making a trade-off between obtaining detailed fields and obtaining detailed values. The value of the different options must be considered, in particular the ability of the measurements to constrain or assess model performance. In many instances, covering a larger spatial extent with less information in each value can provide a more useful test of model performance than few points with more informative values (Jetten et al. 2003; Hunter et al. 2005).

The type of value determines the calculations that can be undertaken with the spatial field. Continuous values can permit differences, ratios and magnitudes to be calculated between element values, whereas categorical fields generally only permit agreement or disagreement between categories. If the categories have some kind of order, ranking can also be done. For comparison of spatial fields, the type of value is a major control on the available comparison methods. It must be the same between the fields being compared (or else made to agree). Possibilities for comparison with continuous or categorical spatial fields are examined further in Chapter 3, although the focus for improving methods in this thesis is on continuous fields, due to their prevalence in hydrology.

2.3 Observed fields in hydrology

Observed spatial fields are used for many purposes in catchment hydrology. They can serve as parameters or forcing data for spatially distributed hydrological models (e.g. Western et al. 1999b). They can be analysed to obtain a detailed understanding about the spatial variability of hydrological attributes (e.g. Merz and Plate 1997). They are also used for a myriad of other spatial analyses, ranging from estimating quantities to allocating resources. A review of common observed spatial fields within hydrology is given here. Two associated issues that arise when dealing with observed spatial fields – the use of surrogates and understanding measurement noise – are also reviewed. These considerations are necessary in making sure that observed fields are fit for use in all aspects of modelling. When used for calibration and testing, observed fields provide the representation of reality against which modelled fields are compared. Ensuring that this is truly representative of reality is vital for making meaningful comparisons.

2.3.1 Examples of observed hydrological fields

Foufoula-Georgiou and Vuruputur (2000) used continuous fields of rainfall, collected via RADAR measurements, to assess the ability of stochastic models to simulate space and time variability of rainfall (Figure 2.5a). They found this to be possible due to the dynamic scaling behaviour of rainfall fields. Houser et al. (2000) also used continuous fields of rainfall, this time interpolated from a network of rain gauges (i.e. point elements), to assess the suitability of remotely sensed brightness temperature as a surrogate for rainfall in semi-arid environments. They found that spatial variability of rainfall must be represented in their model to correctly simulate hydrological response. Houser et al. (2000) also used continuous fields of soil moisture to assess the performance of different data assimilation schemes applied to their model. Data assimilation uses observations to correct the model states and inputs, which should improve the model predictability by ensuring that the model conditions are up-to-date.

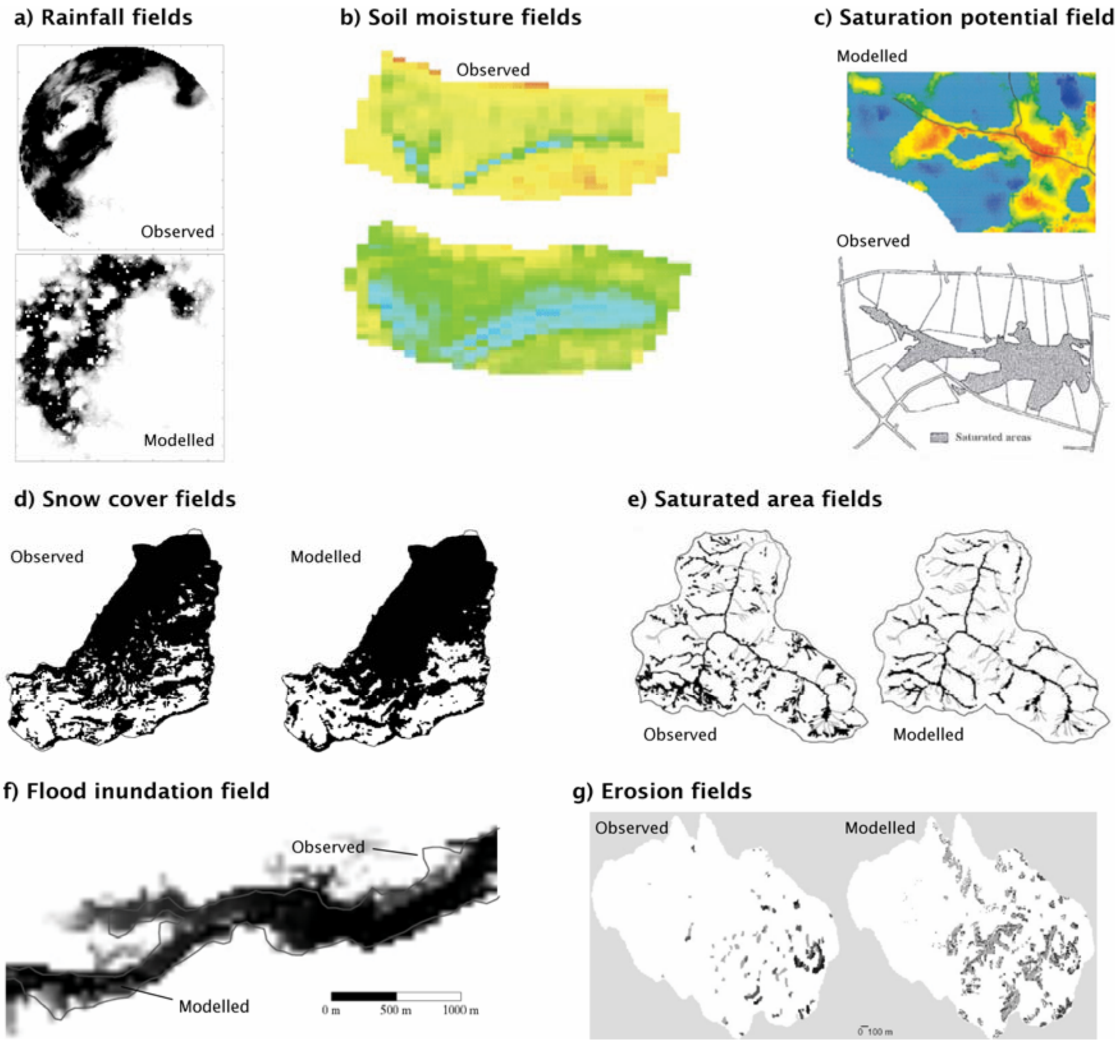


Figure 2.5 Examples of spatial fields that have been used in different areas of hydrology: a) rainfall amount (black is high rainfall) (from Foufoula-Georgiou and Vuruputur 2000); b) soil moisture (blue is wet, yellow is dry) (from Western and Grayson 2000); c) saturation potential index (red is high saturation potential) (from Troch et al. 2000); d) snow cover (white is snow) (from Tarboton et al. 2000); e) saturated areas (black is saturated) (from Güntner et al. 2004); f) flood inundation (line is observed flood extent, dark is high likelihood of inundation, light is low) (from Aronica et al. 2002); and g) erosion amount (white is low, grey is medium, black is high) (from Jetten et al. 2003).

Pauwels et al. (2001) assimilated two different types of information obtained from observed fields of soil moisture – the statistical summary and the actual element values. They assessed performance based on catchment discharge values and found that assimilating the statistics of the field led to most improvement. When assimilating the field, a minor improvement was recorded, although this is difficult to evaluate because the assessment did not include spatial field comparison.

Western and Grayson (2000) observed continuous fields of soil moisture at multiple times (Figure 2.5b). These were analysed as point element fields and also upscaled (via geostatistical interpolation) to match the modelled fields. By visually comparing the observed and modelled fields, important processes for the model were determined. Amongst other things, they found that variable soil properties were not needed, while preferential flow through cracked soils would have improved model performance. Chirico et al. (2003) also used continuous fields of soil moisture, amongst other data sets, to determine the processes that should be represented to achieve sufficient model performance. The final spatial simulations of the model were quantitatively compared against the observed fields to assess model error. Wilson et al. (2005) has used the same type of fields for testing the performance of regression-based models of soil moisture variability.

Blöschl et al. (1991) used binary categorical fields of snow cover to assess simulations from a distributed model (Figure 2.5d). Adding complexity to the model (solar radiation and wind drift parameters) improved the model performance when compared to the observed fields. Continuous spatial fields of snow water equivalent were used by Tarboton et al. (2000) to assess their snow model. They found improved performance when wind drift processes were included, while other processes led to insignificant performance gains. Another example of binary categorical fields exists with saturated area modelling. Güntner et al. (2004) mapped saturated areas in the field, then used these to evaluate the best performing model from a collection of simple and complex terrain-based models (Figure 2.5e). They considered scale differences between the modelled and observed fields during comparison. Without observed fields, no quantitative assessment of these models would have been possible. Continuous fields representing saturated areas have also been derived from time series of synthetic aperture radar (SAR) images (Troch et al. 2000). By recognising the areas with low variability over time, an index of saturation potential can be produced and used during modelling (Figure 2.5c). This models was evaluated using a surrogate field based on in-situ mapping of saturated areas.

Franks et al. (1998) used mapped saturated areas to constrain model predictions of soil moisture. The addition of an observed spatial field into the generalized likelihood uncertainty estimation (GLUE) methodology allowed significant reduction in the predictive uncertainty of the model. A synonymous approach has also been used for flood inundation modelling, where an observed binary field of flood extents (Horritt et al. 2001) is used to provide additional assessment of model performance. Bates et al. (2004) and Aronica et al. (2002) both used this type of additional assessment with the GLUE methodology to constrain the possible model parameters and reduce predictive uncertainty (Figure 2.5f). Both of these examples show how the simplest binary categorical fields can provide useful information to improve modelling. Categorical fields of erosion amount (with more than two categories) were mapped in the field by Jetten et al. (2003). The LISEM (Takken et al. 1999) erosion model produced continuous modelled fields, which were changed into categorical fields to make them comparable (Figure 2.5g). They found the observed field particularly useful for calibration and testing of the model and recommend using this type of data for assessment rather than integrated responses at the catchment outlet.

2.3.2 Surrogate observations

When the available measurement techniques cannot directly measure the attribute of interest, surrogate measurements can be used to represent the attribute, albeit with some uncertainty. In the strictest sense, virtually all measurements are surrogates, although the surrogates referred to here generally lack high correlation with the attribute of interest (Grayson et al. 2002). Surrogates are variables that have some correlation with the attribute of interest, but they are usually easier to obtain over a large spatial extent. Due to the lack of high correlation, surrogate fields that are used for comparison as observed fields should consider measurement errors (as should any observed field not assumed to be a perfect reality). When using surrogates, the scales over which the correlation holds must be considered, otherwise the spatial arrangement of the surrogate will not be an accurate representation of the hydrological attribute, thus making it unsuitable for comparison.

Hydrologists most commonly use terrain or some characteristic of the terrain as a surrogate. As such, digital elevation models (DEMs) are considered the base data in most modelling studies. They are widely available as a continuous field and for many hydrological attributes elevation is a major control on spatial variability (Moore et al. 1991). Recent technologies, such as airborne scanning laser altimetry (Cobby et al. 2001; Flood 2001) and SAR interferometry (Rabus et al. 2003), facilitate the rapid collection of high quality DEMs at varying scales and across large spatial extents. A complete description of the uses of terrain models in hydrology is given by Moore et al. (1991). To better understand these uses, as well as other surrogates used in hydrology for representing the spatial distribution of hydrologic attributes, some examples are given here.

DEMs are often used as a surrogate for spatially distributing rain gauge measurements, due to the orographic influence of elevation on precipitation. Mean annual rainfall and elevation are usually highly correlated, but when hourly rainfall is needed the correlation is far weaker (Blöschl and Grayson 2000). Many studies use terrain as a surrogate for runoff generation though the use of the topographic wetness index (Beven and Kirkby 1979). This index makes the assumption that saturation excess is the only runoff generating process and should be used with this in mind. It provides a simple model that transforms slope and contributing area values into a useful index (Quinn et al. 1995). Terrain aspect is used as a surrogate for solar radiation, which is useful in snow modelling (e.g. Blöschl et al. 1991) and soil moisture modelling (e.g. Western et al. 1999b) for considering snow melt and evapotranspiration effects respectively.

Remote sensing measurements are often considered surrogates, as they measure reflectance values in different wavelengths. These values are processed into values that represent attributes of interest. For example, fields of land use categories are produced using combinations of reflectance values from different wavelengths. Passive microwave sensors measure brightness temperature values from the Earth and use models to convert these to near-surface soil moisture (Schmugge et al. 2002). Active microwave sensors (e.g. SAR) record backscatter values, which respond to the varying dielectric constant of the soil and can be translated into moisture content values (Ulaby et al. 1996).

Surrogate fields are also used to produce fields that are needed for forcing, parameterising and assessing models. Where a particular model parameter cannot be measured directly, surrogates provide the means of obtaining spatially distributed parameter values. For example, a categorical field of soil types can be transformed into a field of soil hydraulic properties by using ‘pedo-transfer functions’ (e.g. Rawls et al. 1983). For forcing a model, Wilson et al. (2005) showed how regressed terrain characteristics could be used to prepare spatially distributed soil moisture initial conditions. In model assessment, surrogates are often produced because the measurements are not readily usable in their raw form. Verhoest et al. (1998) used multiple SAR fields to produce a surrogate map of potential saturated areas. Horritt et al. (2001) used the smooth areas in a SAR field to represent the area inundated during a flood event. In these situations, the measured fields are made comparable to the modelled fields. Grayson et al. (2002) also explain that models can be designed to produce fields that are comparable to the measurement methods available. Land surface models are designed with a shallow ‘skin layer’ of soil moisture to facilitate assessment using measures from remote sensors (that only represent the response from the shallow layer near the soil surface).

2.3.3 Measurement error

There is measurement error in any measurement made, whether made in-situ or remotely. The measurements used to produce observed spatial fields are no exception. Measurement errors are either systematic or random (i.e. noise). Systematic errors are usually due to some characteristics of the measurement method, such as equipment set up or performance over different ranges (e.g. instrument calibration). Systematic errors can often be corrected, provided that some more accurate measurements are available (e.g. calibration data for a particular measurement type). Surrogates often have systematic errors from the hydrological attribute they are representing. If these errors are well understood, then the surrogate can provide a useful representation of the spatial field.

Random errors can be significantly reduced via averaging. This is done by having multiple measurements at each location and averaging them, or otherwise by averaging spatially-similar element values. Reducing the amount of noise in a field helps to reveal the true spatial variance of the attribute (i.e. the signal), assuming that systematic errors are well managed. If the noise is much higher than the variance in the field, the signal is hidden (e.g. Figure 2.6c). If the noise is less than the variance, then the signal will be more apparent (e.g. Figure 2.6a). Western et al. (1999b) experienced this effect when analysing soil moisture fields throughout the year. During summer, the measurement error was as high as the variance, thus hiding any signal or pattern. In winter, the variance of the field was much higher than the measurement error, thus revealing a clearer arrangement.

The amount of improvement in measurement error by averaging is limited by the independence of the errors (Blöschl and Grayson 2000). If the errors are independent, then measurement errors decrease with the inverse of the number of samples being averaged. Averaging also changes the characteristic scale of the measurement, as it produces an average over time and/or space, thus changing the support of the element. Scale changes must be considered so that the averaging does not make the meaning of the field unsuitable for the purpose. However, averaging multiple fields over time can be used to reduce the measurement error in remotely sensed fields and thus produce

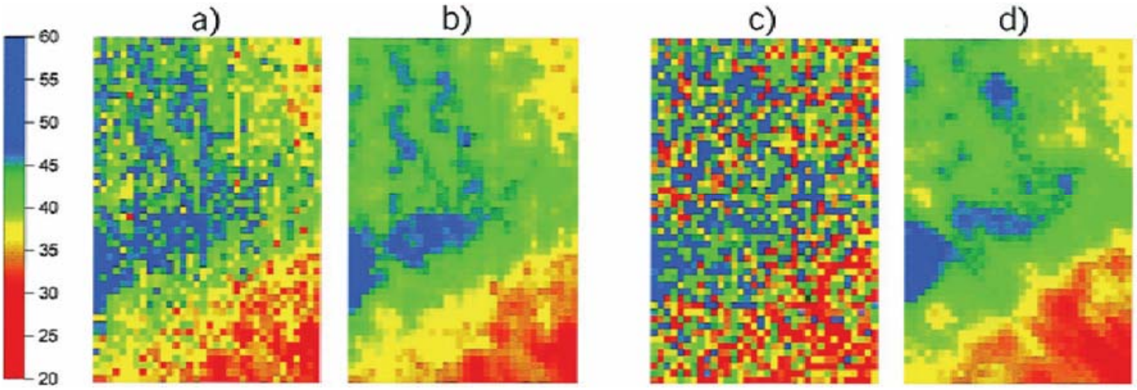


Figure 2.6 Examples of high resolution soil moisture fields (2m x 2m elements) with different degrees of measurement error: a) measured soil moisture field (Western et al. 1998); b) measured field, smoothed using ordinary kriging with nugget of 3% v/v; c) noisy field, created by adding random noise ($\sigma^2 = 40\%$ v/v) to measured field; and d) noisy field, smoothed using ordinary kriging with nugget of 43% v/v (from Blöschl and Grayson 2000).

fields with higher signal-to-noise ratio. Verhoest et al. (1998) used this approach to determine saturated areas from multiple SAR fields, which individually had weak correlations with soil moisture response.

Averaging in space rather than time is achieved by using nearby measurements (i.e. aggregating). This is often termed filtering or smoothing. 'Moving window' filters can be used to achieve this (e.g. Mastin 1985), although where spatial correlation exists in the errors, geostatistical methods can produce better realisations of the smoothed field. Blöschl and Grayson (2000) presented an example where a field of soil moisture with measurement error of 3% v/v (Figure 2.6a) was filtered using ordinary kriging (Figure 2.6b). The resulting field shows a clearer signal or pattern, but not vastly different from the original field. When additional noise is added (Figure 2.6c), the field produced by smoothing the measurement error (Figure 2.6d) has a much clearer pattern than the noisy field. In this situation, the aggregation method (i.e. ordinary kriging) uses the nugget effect to determine how the smoothing is handled. This analysis shows that when the errors are understood, systematic errors can be minimised and the effects of measurement errors can be removed. In observed fields containing measurement error, pre-processing to reduce noise is beneficial because it helps make the observed field more representative of reality.

2.4 Spatial models in hydrology

Observed spatial fields provide a picture of the state of a hydrological attribute and are essential for understanding catchment response. However, to understand the dynamics of catchments, a framework to facilitate hypothesis testing is needed. Computer-based modelling is used throughout hydrology for this purpose. There are many different types of models, ranging from those that estimate bulk quantities to those that produce spatially explicit estimates across an extent. In this thesis, the focus is on spatial models, which are used for testing hypotheses about the behaviour of hydrological systems. Models provide the platform on which conceptualisations of hydrological processes are combined to simulate hydrological response. If models prove to adequately simulate a certain response, they can also be used for predicting the effects of changed conditions on hydrological response (e.g. land use change). Typical examples of the types of spatial models used in hydrology are given here. This is followed by a discussion of the main issues to consider when working with spatial models, which centre on the choice of model complexity. The way these issues are managed during the processes of model calibration and testing is then discussed. These tasks are where spatial field comparisons can influence and advance modelling practice.

2.4.1 Types of models

There are many comprehensive reviews of hydrological modelling available, which provide examples and classifications of models (e.g. Singh 1995; Abbott and Refsgaard 1996; Grayson and Blöschl 2000a). These reviews highlight that models can be classified in many ways, although a common approach is to look at three main aspects of any model to help describe it: 1) the spatial representation (i.e. lumped or distributed); 2) the algorithms (i.e. empirical, conceptual or physically-based); and 3) the type of inputs and parameters (i.e. deterministic or stochastic).

Lumped hydrological models have a single element to represent the entire study extent, thus failing to represent any spatial variability. Typical examples are the STANFORD watershed model (Crawford and Linsley 1966) and the Sacramento model (Burnash 1995). Lumped models only provide an average value for the extent, thus being useful for addressing questions such as “how much?”, but not questions such as “where?” In contrast, distributed models discretise the study extent in some manner and produce values for multiple model elements, providing a representation of variability in space. Typical examples include SHE (Abbott et al. 1986), Thales (Grayson et al. 1995) and TOPMODEL (Beven 1995). These models can use different types of structures to represent the study extent, although regular grid-based structures are common. The model element size determines the degree of spatial variability that can be represented and also the scale at which processes must be represented. Any variability that occurs at a smaller scale can only be included in the model by using additional parameterisation.

Mathematical algorithms are used to calculate values for each model element, whether it is a single lumped element or many distributed elements. In the simplest case, empirical relationships are used to describe how the output varies relative to the input. This type of algorithm does not try to represent each individual process occurring. A more common approach is to determine relationships that represent individual processes. A conceptual understanding of how the processes interact is then used to combine the individual algorithms to produce the desired model outputs. This is a classical approach that is still in use. TOPMODEL is an example of a conceptual distributed model (Refsgaard 1997) that uses simple topographic relationships to provide the spatial variation in model outputs. Physically-based algorithms use the governing equations of water flow over and through soil and vegetation to model hydrological response. These models aim to represent the physical processes and their interactions, using parameters that are, in principle, measurable physical quantities (rather than abstract parameters). The processes that are represented in a given model are determined by the purposes of the model, as well as the knowledge that exists about the phenomena being modelled. The SHE and Thales models are both physically-based spatially-distributed models that are used for hydrological research (e.g. Western et al. 1999b; Chirico et al. 2003; Boegh et al. 2004; Abu El-Nasr et al. 2005). For any additional processes to be included in models, additional parameters are needed. It is often difficult to obtain measured values for all the parameters, leading to a situation where model calibration is needed to refine parameter estimates. As such, physically-based models are sometimes referred to as complex conceptual models. The amount of complexity in a model (i.e. the number of processes represented) must be balanced with the purposes of use and the availability of data, as more complex models have greater data requirements. As models become more complex, observed spatial fields become more useful for model assessment via comparison. Issues relating to model complexity are discussed further in the next section.

All models require inputs to produce modelled outputs. The inputs are the forcing data (e.g. the meteorological conditions that drive the system) and the parameters (i.e. the controls on how the processes represented in the model act). These can be provided as deterministic or stochastic inputs. Deterministic inputs are the most common type, where a single set of inputs are used to produce a single set of outputs. Stochastic (or statistical) inputs are being increasingly used in hydrology to more completely define the range of possible inputs that could exist for the model. Statistical distributions are provided for some or all of the parameters and forcing data. In most situations, a number of parameter sets are then sampled from the distributions and run through the model, producing a range of outputs that can be used to analyse uncertainty (Figure

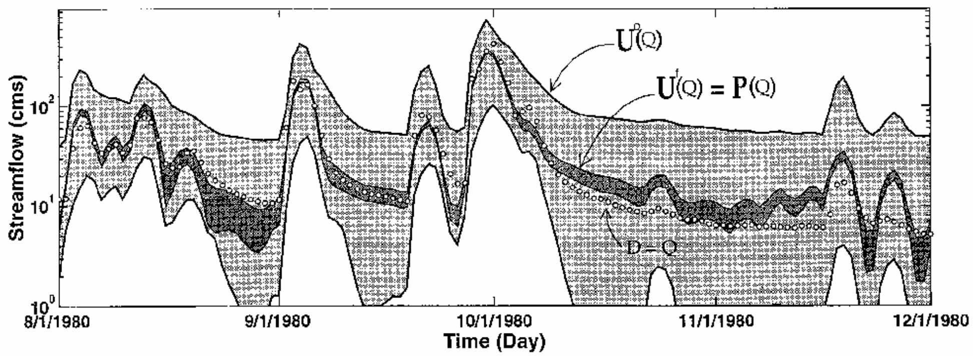


Figure 2.7 Stochastic inputs to models are used to produce stochastic outputs, which can define the bounds of model simulations. This example time series (from Gupta et al. 1998) shows the observed data (dots), the bounds for all parameter sets considered (light grey area) and the bounds of the optimal parameter sets (dark grey area).

2.7). This is conceptually similar to the ideas behind methods such as GLUE (Beven and Binley 1992). In some models, the statistical distributions are applied within the process equations (e.g. Entekhabi and Eagleson 1989), although these approaches usually need the distributions to be simplified to make them computationally manageable. The increased use of stochastic inputs is in response to the increased need for estimates of uncertainty on model outputs.

2.4.2 Issues with spatial models

Model complexity is the major control on the capabilities of spatial models and also the issues related to their use. Determining the optimum model complexity for a study is a task that primarily relies on the experience of the modeller. However, there are two general approaches that are used when building a model – the downward and the upward approach (Klemeš 1983; Grayson and Blöschl 2000a). The downward approach begins with the simplest model possible, only adding in additional processes if they improve the ability of the model to simulate the observed data (Sivapalan et al. 2003). Even if a certain hydrological process is known to occur in the study extent, it is only included in the model if it improves performance (as evaluated via comparison). This results in a model that reproduces the observed data as best possible, but can lack the ability to simulate correctly when other processes occur. In contrast, the upward approach includes all the processes that could be operating in the study extent. It assumes that if the processes are represented correctly (i.e. the conceptualisations and parameters are correct), then the model simulations will be correct. This approach can produce models that are too complex and cannot be adequately parameterised or tested. If the observed data are insufficient, the model processes may need calibration (i.e. to assign parameters) and/or the performance of the model may not be adequately tested. When this occurs, models are said to have identifiability problems, because the data does not allow identification of whether the model is indeed producing the right answers for the right reasons (Klemeš 1986). These problems are also termed non-uniqueness or equifinality (Beven 2001).

The upward and downward approaches to modelling can lead to quite different representations of reality, none of which will be entirely correct. Instead, hydrologists are forced to accept some level of "pragmatic realism" (Beven 2001), in which the real processes are represented using 'less than ideal' algorithms or relationships. Overall, modelling approaches should endeavour to use representations of processes or relationships that hold true over the key spatial and temporal scales being investigated (e.g. Beven 2001; Littlewood et al. 2003; Sivapalan et al. 2003). To achieve this, it is necessary to find a balance between representing sufficiently complex processes while working within the limitations of the available data. This is where the upward and downward approaches converge and suitable models are developed.

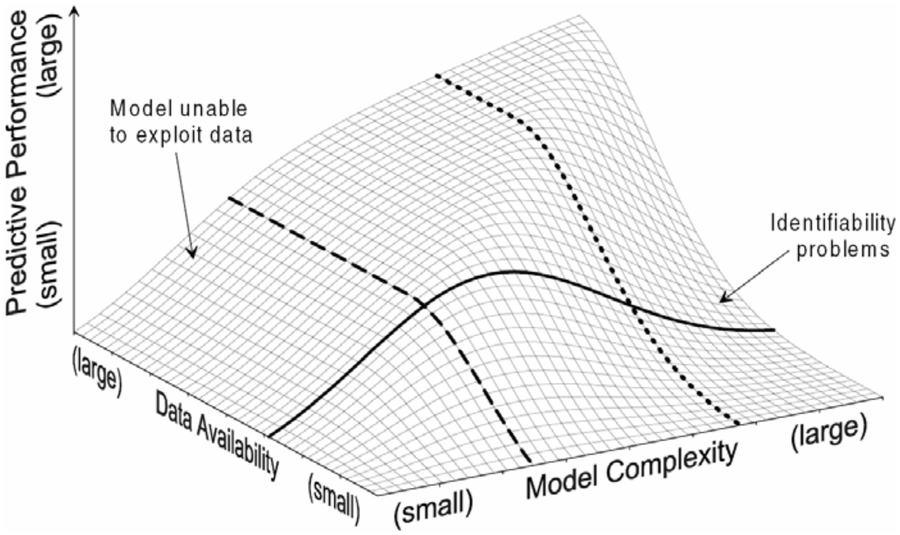


Figure 2.8 A conceptual relationship between model complexity, data availability and predictive performance (from Grayson and Blöschl 2000a). The solid line represents certain data availability, while the dashed lines represent two models with different complexity.

At some point, a model will have sufficient complexity to represent the dominant processes, while still having sufficient data to parameterise and test that the model is simulating the study extent correctly. This is where the optimal predictive performance will be found for a given model/data combination (Figure 2.8). By finding this balance, modellers can ensure their model performs as optimally as possible, while still being able to explain why it performs as it does. With spatial models lacking available data, poor predictive performance is usually encountered due to identifiability issues and the large uncertainty that goes with these simulations. To rectify this situation, efforts in collecting data for models have increased over the past decade (Grayson et al. 2002). In most cases, data availability is the limiting condition on model complexity and assessment. As observed spatial fields are used more in modelling, the predictive performance of models is expected to improve. There are many complex models that are not adequately tested. They should be tested more thoroughly before more complexity is added. This thesis focuses on the way that these complex models can be better tested via spatial field comparison.

2.4.3 Model calibration and testing

Model calibration and testing are traditionally separate processes. The observed data are often split, with half being used for calibration and half for testing. Model calibration is the process used to determine the set of parameters that produces the best model performance. The optimal parameters are discovered through calibration by comparing model simulations with observed data. By adjusting the parameters, the calibration converges on the optimum set. Then, using this optimal parameter set, the model is used to simulate the period for which observed data are available. Testing is the process of evaluating particular aspects of model performance to ensure the model is fit for a certain purpose. This is done by comparing the model simulations against a reference defined by the modeller, which is usually some independent observed data. Spatial field comparisons are required when the observed data are spatially distributed. Rykiel (1996) describes the general types of testing that are done in modelling studies, ranging from expert assessment to statistical tests. Testing is used to assess the 'fitness of use' of a model, so the purpose and context of the model must be well understood prior to undertaking model testing.

As models have become more complex and the number of parameters has increased, calibration procedures have been found to less frequently converge to a single set of parameters that is optimum. Instead, a number of parameter sets are found to all produce optimal performance. This is an identifiability problem (Figure 2.8), which is also termed non-uniqueness or equifinality (Beven 2001). There are two main approaches for managing this problem: 1) to limit the number of free parameters to help simplify the calibration process; or 2) to run simulations for all of the feasible parameter sets and reject any non-behavioural simulations. Non-behavioural simulations are those in which the model is not a plausible explanation of the observed field. In both cases, observed data are needed to evaluate the model performance, whether it is done visually or quantitatively. The first approach is discussed by Refsgaard (1997), who explains that model calibration with distributed models is possible if the ratio between independent observed data and the number of free parameters is kept low. This is done by carefully parameterising the model using observed data, experience from similar studies or surrogates (e.g. using a soil map to spatially assign relative soil hydraulic properties). If the processes represented in the model are physically-based, then observed data (collected at the correct scale) can be used to assign parameters and avoid most calibration (e.g. Western et al. 1999b). If the many parameter values can be reduced down to a manageable number, calibration can still find an optimal solution. After calibration, the optimal model must be validated to ensure it is fit for use, which could be achieved using comparison with independent observed fields.

The second approach is to reject the idea of calibration entirely (Beven 2002). Rather than searching for the optimal parameter set for a certain model, all possible models and possible parameter sets are analysed and then constrained. Models and parameters sets are rejected if they are not physically feasible. Once the set of possible models and parameters are defined, they are used to simulate the hydrological attributes. The model simulations are compared with observed data to further reject any unfeasible models. At this stage, all models and parameter sets that are judged as non-behavioural are removed. The remaining, behavioural models are used to define the cumulative distributions of the simulated attributes. The GLUE methodology follows this general approach and is increasingly being used in distributed modelling studies (e.g. Franks et al. 1998; Bates et al. 2004). Under this approach, the observed data are used to determine all possible simulations and assign their probabilities, thus providing a representation of the uncertainty in the model. This conditioning process replaces the usual calibration and testing steps, although predictive testing (i.e. testing of model predictions using future observations) would still be desirable (Rykiel 1996). This approach to modelling is still being researched and further developed, but is expected to have a major impact on hydrological modelling once it is commonly used and accessible (Grayson and Blöschl 2000a).

To complete any calibration, testing or conditioning process, the model simulations must be compared to a reference to assess performance. Rykiel (1996) describes the three things a modeller must specify for testing (which also apply to calibration): the purpose of the model; the context of its use; and the tests the model must pass (i.e. performance criteria). The first two criteria are often not considered when validating models. Instead, the 'common' suite of tests are run between the data sets and evaluated. However, if a model is going to be used to determine where to spatially allocate resources, then its ability to get the spatial arrangement right must be validated. Jetten et al. (2003) make the point that it is more economical to over allocate in one location than allocate to the wrong location. Testing different aspects of model performance is critical to the whole model assessment process. Qualitative tests, such as visual inspection and expert opinion, can be used, but quantitative tests are more repeatable and useful for automated calibration and testing.

2.5 Chapter summary

A concise definition has been given to clarify what constitutes a spatial field (i.e. a set of elements with a defined relationship between them). The specific definition of a field is also controlled by its characteristic spatial and temporal scales. Understanding the scale triplet for a field is important to ensure that it is comparable with another field. Many scale differences are often overlooked so that spatial field comparisons can actually be used, as it is uncommon for the fields to have matching scale definitions. Pre-processing methods, such as interpolation, can enable scale differences to be resolved and have been discussed.

Typical examples of observed and modelled hydrological spatial fields show the diversity of hydrological attributes that are represented in a spatially distributed manner. These are the application areas in which spatial field comparisons are most likely to be used. Issues that must be considered when working with these are discussed. These include the removal of measurement errors to make the field more representative of reality and also the use of surrogates when the hydrological attribute cannot be directly observed.

Finally, the types of hydrological models used to produce spatial fields have been reviewed and their issues discussed. As models become more complex, they demand different approaches to calibration and testing. These model assessment tasks are where spatial field comparison is most able to advance hydrological modelling practices. The development made in the remainder of this thesis address this potential.

Chapter 3

Review of spatial field comparison

3.1 Chapter overview

The comparison of spatial fields (or images) is a fundamental task in many disciplines, including image processing, pattern matching/recognition, medical imaging and landscape ecology. In hydrology, spatial field comparisons are used primarily for model assessment tasks, but the methods used for making these comparisons are currently in their infancy.

Hydrologists rely heavily upon qualitative visual comparisons, which are very useful but lack the rigour and repeatability needed for model assessment. To address this, simple quantitative methods are applied to produce comparison measures that can be directly used in model assessment tasks. However, the simple quantitative methods treat the elements within a field independently, which makes them incapable of comparing any aspects of spatial relationships (apart from strict agreement). These quantitative comparison methods can potentially be improved by using: 1) an understanding of how visual comparisons work; and 2) the knowledge about comparison methods from other, more mature disciplines.

This chapter begins by reviewing the key tasks undertaken during human visual inspection and comparison. This highlights some aspects of visual comparison that would be desirable to emulate in a quantitative manner. The current state-of-the-art for quantitative comparison in hydrology is then presented. This reveals where the methods are currently limited and where advances would be most beneficial. An investigation of other related disciplines is then given, which identifies any existing comparison methods that could be applied or adapted for use in hydrological comparison. Each of these three major sections are summarised individually. The chapter then closes with a proposal for advancing the methods used for quantitatively comparing spatial fields in hydrological modelling, based on the findings of this review. Major components of this chapter have been published in Wealands et al. (2003; 2005a).

3.2 Visual comparison

Visual comparison is considered the most powerful and comprehensive method for comparing spatial fields (Grayson et al. 2002; Hagen 2003). All modelling studies in which observed and modelled fields are available tend to use this method before any other, mainly due to its immediacy and versatility. Visual comparison is a method that is innate to humans because it is simply a task of visualising multiple fields and then 'playing spot the differences' between them.

When visualising spatial fields, element values are commonly represented using a colour from a specific colour ramp (i.e. a scale of colours used to represent the numerical values). The research in cartography and other disciplines suggests some colour ramps that are suitable for certain types of data, although there are no firm guidelines about the best colours to use (Brewer 2003; Light and Bartlein 2004). In all situations, the number of colours chosen (i.e. the discretisation of the element values) has a major influence on the visual appearance of the field. When few colours are used to represent the full range of element values in a spatial field, the true variability in the field can become generalised (Figure 3.1b). For categorical fields, a distinct colour is usually applied to each distinct category value. For continuous fields, continuous colour ramps should be used so that a different colour is assigned to every value, resulting in very similar values having very similar colours (Figure 3.1c). Light and Bartlein (2004) describe continuous colour schemes that use one hue for each category, with varying intensity to reveal variability in that category (i.e. from light to dark). Continuous colour ramps allow all the variability between elements to be shown and leave the responsibility with the human observer to make judgements about features that may or may not exist.

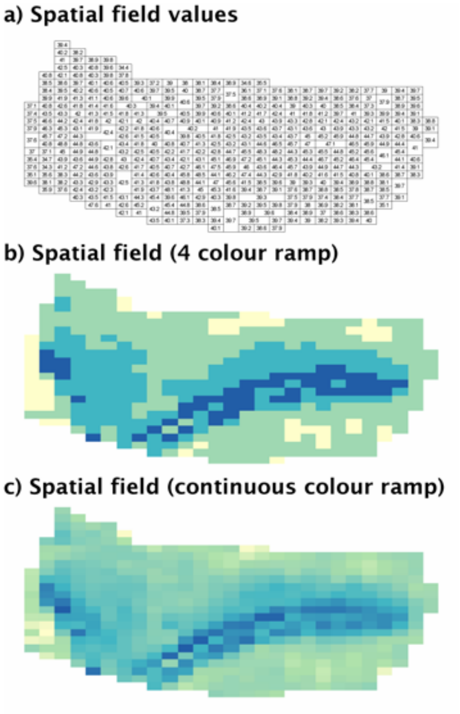

a) Spatial field values
b) Spatial field (4 colour ramp)
c) Spatial field (continuous colour ramp)

Figure 3.1 The method used to visualise a spatial field plays a major role in how it is interpreted during visual comparison. a) shows the numerical element values; b) shows the element values represented with four colours (using equally-spaced category values); and c) shows the element values on a continuous colour ramp.

A spatial field is presented to a human observer for visualisation by using a colour ramp to assign colours to each field element. When the field is visualised, the innate abilities of the human visual system are then used to make a number of mental notes about characteristics of the field. The human visual system uses a highly active process to convert the complex visual stimuli (i.e. the colours in the field) into organised information (Definiens Imaging 2003). For example, a human observer will note the general colour of the entire field, the colour (and variability) in different parts of the field and the approximate amount of the field covered by each colour. Mental notes about whether the colours are smoothly varying or whether there is high variability will also be made. During this process, cues to perceptual organisation, such as proximity, connectedness, continuity and closure (Sarkar and Boyer 1993) are also used to assist in recognising homogeneous regions within the field. These regions can then be described using the background knowledge of the observer. These many powerful capabilities of human visualisation are explained in later sections. The characteristics of the fields that are noted during visualisation depend on the observer. An experienced observer who understands the data and the purpose of its use will make more rigorous notes about the field compared to an independent observer with little background knowledge. This inconsistency is one of the primary weaknesses with visual comparison, which results from all human observers having different background knowledge and different biases.

To conduct a visual comparison between two fields, the most common approach is to visualise them together, either next to each other or toggling between them on a computer screen (Aerts et al. 2003). Each characteristic from each field can then be inspected and compared individually. Human observers tend to generalise the characteristics of the fields (in colour or location), which leads to comparisons that are not exact. The results are often just mental notes or they may be translated into a type of scoring system, in which similarities are rewarded and differences are penalised. An alternate approach is to use some type of quantitative comparison method first (e.g. compute the residual field) and to then visually interpret the resulting field (e.g. Western and Grayson 2000). In this case, the difference between each element value is captured in the resulting field and used to determine the comparison measure. Regardless of the approach taken, when visual methods are used for comparison, the results of the comparison are difficult, if not impossible, to quantify. The true nature of the comparison is also ambiguous due to the different ways that human observers judge similarity.

During both visualisation and comparison, there are large demands placed on the ability of the human visual system. This system has evolved through time and is a powerful tool for explaining the visual stimuli presented (Definiens Imaging 2003). However, the visual system uses various mechanisms to simplify the scene being observed, so that it can be feasibly processed in real-time (Itti et al. 1998). Intermediate-level visual processes appear to select a subset of the available stimuli before further processing (Tsotsos et al. 1995). This approach, combined with the organisation of a field into regions, makes it possible for humans to process complex scenes. When complex spatial fields (i.e. multiple and/or highly detailed fields) need to be compared, these processes are employed by visual comparison methods. These methods generalise the fields and limit the extents over which comparison is applied. Visual comparison methods cannot rigorously compare all elements in complex fields and therefore produce results that are biased by the structure imposed and the region investigated.

This review of visual comparison methods has revealed the three biggest problems with visual comparison: 1) repeatability, as no specific set of characteristics are compared; 2) quantification, as there is no method to translate the mental notes into measures; and 3) rigour, as the human visual system imposes subjective organisation and importance onto the data to make processing complex data feasible. These problems can be addressed to some extent by using statistics and image processing algorithms. Repeatability, quantification and rigour are all achieved by specifying a set of algorithms that use the element values to produce numerical comparison measures between multiple, complex fields. The difficulty lies in finding algorithms which are capable of comparing the characteristics of the fields that are suitable for the purpose. The following sections provide a discussion about the general aspects of human vision and highlight the actual tasks that may be captured into algorithms.

3.2.1 Comparing multiple characteristics

One of the obvious strengths of visual comparison is the ability to compare many different characteristics of spatial fields and then formulate an assessment of similarity. The characteristics that are compared are determined by the observer and are influenced by their understanding of the data and the purpose. There are some characteristics of spatial fields that are always useful to compare as they give general information about the fields. These include the general colour, amount of variability, smoothness of the variations and the distribution of the colours (both in quantity and spatial arrangement). Power et al. (2001) acknowledge that humans intuitively identify a number of similarities between two images, by firstly noticing overall similarity and then focusing on the finer details. Other characteristics will also be of interest depending on the situation, although the experience of the user is needed to determine which ones. The user can then formulate an overall assessment of similarity by combining the knowledge about how similar or different certain characteristics are.

This aspect of human vision depends largely on user experience, which cannot be replaced by a single algorithm. Instead, different algorithms must be used to obtain comparison measures that inform about specific similarities or differences. These can then be combined into a single, overall measure (i.e. a combined index) or interpreted separately. In hydrology, reporting multiple comparison measures (e.g. Legates and McCabe 1999) or undertaking multi-objective calibration (e.g. Boyle et al. 2000) are the most common way of achieving this.

3.2.2 Focusing on important parts

Research into eye movements and visual attention suggests that humans interpret complex scenes by selecting a subset of the available sensory information before further processing (Itti et al. 1998; Bruce 2003; Maeder 2005). This subset is considered to be the perceptually important information. By working with a subset, the processing is simplified and the effect of unimportant information is removed. An approach such as this could be considered biased and incomplete, or it could alternatively be considered efficient and avoiding the interference of irrelevant information. The real power of the human visual system is in its ability to rapidly identify the information that is important. This capability varies with the individual, the data and the task (i.e. the question being asked).

Studies that aim to model visual attention begin by trying to understand what parts of a scene are perceptually important. A number of studies have looked at physical evidence such as saccadic eye movements (those initial movements where the eye jumps between different points of interest in a scene) (Findlay et al. 1995; Itti et al. 1998). Other studies have used conceptual models based on cognitive and information theory (Osberger and Maeder 1998; Tompa et al. 2000). In both cases, a model is used to represent the characteristics that draw visual attention. One common finding is that high local contrast is perceptually important (Itti et al. 1998; Tompa et al. 2000). The model of visual attention is then used to produce an importance (or salience) map (Figure 3.2b). From this point, the map of importance can be used to limit the extents of analysis or to define the starting points for subsequent processing. During comparisons, this approach can be used to prioritise or limit the parts of the fields that are compared. Tompa et al. (2000) has achieved this by using information maps to weight the comparisons, thus reducing the impact of differences between element values that are considered unimportant.

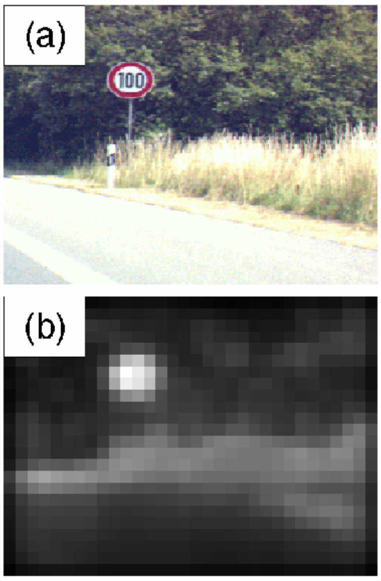


Figure 3.2 A real-world scene (a) and the computed importance (or saliency) map (b) (from Itti et al. 1998). Lighter colours in the importance map represent greater importance. In this example, visual attention has been modelled by combining three factors – intensity contrast, colour contrast and orientation contrast.

This aspect of visual comparison can be captured into an algorithm, although the algorithm generally requires user input to define the characteristics that are deemed important for the application. These issues of visual attention and using an existing knowledge base are rarely considered in most computer vision examples (Sarkar and Boyer 1993). In hydrology, a synonymous approach has been used with time series data. Gupta et al. (1998) defined different parts of a stream hydrograph (e.g. the rising and falling limbs) that are functionally important, thus making use of expert knowledge to help focus the comparison method. This is not widely practiced in hydrology, although recent research by Shamir et al. (2005) shows that it continues to be useful for assessment of temporal hydrological models.

3.2.3 Recognising organisation and regions

The human visual system depends on focus and simplification to process the vast amount of stimulus it receives. When looking at spatial fields or images, humans rapidly recognise regions, which are effectively groups of elements. This type of organisation is sought by using an innate ability that has evolved in a world that is very structured (Sarkar and Boyer 1993). To illustrate this, an experiment by Smith (1986) is referred to, in which a group of subjects were given a random visual stimulus and asked to reproduce it. The reproduced set was then used as the stimulus for the next group of subjects. After 12 iterations, there was an obvious structure that had been imposed onto the random stimulus. With fields containing sharp boundaries, such as a land use map or an image of the built environment, regions tend to be very apparent. In less structured fields, such as an elevation model or a satellite image of vegetation health, regions are much more difficult to recognise. However, in all cases the human visual system will impose some kind of organisation onto the field. The detection of organisation by human observers is explained by the laws of grouping, as determined by the Gestalt school of psychologists and summarised by Rock and Palmer (1990). They found that the visual system organises 'parts into wholes' using cues such as proximity, similarity, closure, continuation and connectedness. This organisation process can be influenced by attention, intention, attitude and other background information (Sarkar and Boyer 1993).

Once regions have been recognised, the next process is to describe what the region actually represents. This is quite simple when the object is familiar and distinct, but in less structured fields representing abstract attributes (such as those produced by hydrological models), the description will be less clearly defined.

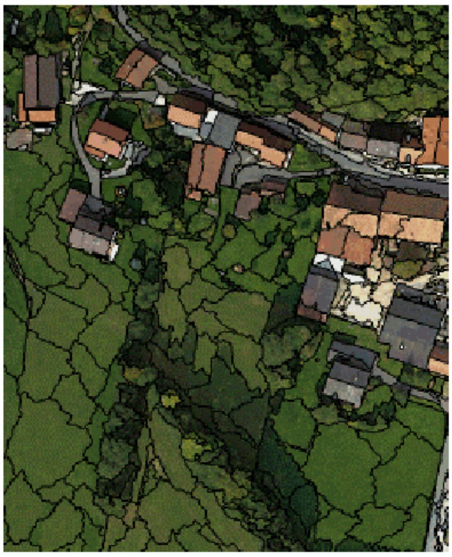


Figure 3.3 An example of regions that have been 'recognised' using an image segmentation approach (from Baatz and Schäpe 2000). Notice the distinct regions found in the structured, built areas and the somewhat arbitrary boundaries specified in the less-structured, natural areas.

The recognition of regions within images is an active area of research in computer vision and image processing. Many methods exist for defining regions within images, although they can be difficult to assess due to the variability that exists within human-derived recognition (e.g. Martin 2002). Image segmentation methods that work at multiple scales (e.g. Baatz and Schäpe 2000) may also be useful for providing a strong recognition of regions within a field with unknown structure (Figure 3.3). These methods are reviewed later in this chapter (section 3.4.1). Once regions are defined, there remains the problem of describing or labelling the regions, as occurs during visual analysis. Application-specific solutions have been found to achieve object recognition in certain situations by using known spectral or textural characteristics to assign labels to regions. When these characteristics are not suitable, information from surrogate data or expert knowledge may be useful for providing the context for labelling regions.

3.2.4 Multiple scales

The use of multiple scales is an important aspect of visual comparison. Unlike most analysis methods, visual comparisons operate across multiple scales. Scale can be considered as the 'window of perception' (Hay et al. 2003) over which fields are observed. As the scale changes, the visual appearance of a field will vary accordingly. Many of the characteristics observed in a field actually appear in different forms at different scales. For example, a forest can be observed as a homogeneous region in a remote sensing image, but when looking at a smaller scale the forest is actually made up of patches of different trees, which are each made up of different individual trees. Clearly, the scale of observation or analysis plays a role in determining what characteristics of a field can actually be analysed or compared. A major strength of human vision is the ability to recognise the 'correct' scale for analysis, a task that utilises extensive background information and an understanding of context.

Performing comparisons at multiple scales is not an original idea, but it is not currently practised in hydrology. In hydrology, multiscale methods have only been used for some spatial analyses, including the characterisation of rainfall fields (e.g. Zepeda-Arce et al. 2000) and the derivation of terrain attributes (Gallant and Dowling 2003). The current quantitative methods used for comparison perform all analysis at one defined scale. Costanza (1989) describes a comparison method that performs analysis across multiple scales and then combines the findings into a single summary. The field is upscaled by increasing the support of each element, thereby smoothing the appearance of the field. Hay et al. (2003) shows a range of methods for recognising regions by analysing multiscale representations of a field. The recognised regions define the elements at the intermediate scale and can be used for analysis. As the scale increases, the regions get increasingly larger and eventually they represent the entire field as a single element (i.e. at a global scale). Once a field is represented at multiple scales, comparison methods can be applied to each scale to produce information about the similarities and how they vary (or persist) across scales (e.g. Pontius et al. 2004b).

3.2.5 Tolerance for differences

During visual comparison, human observers recognise regions and generalise the appearance of the fields. Through simplifying and generalising, the human observer allows characteristics that may be slightly different to still appear similar. The main characteristics that get generalised are value (or colour) and location (which influences both position and shape). During comparison of these generalised fields, the observer ignores any differences that cannot be visually discriminated. This human approach is flexible, while computerised comparison methods are strict.

Hagen (2003) illustrates these different approaches using checkerboards (Figure 3.4), the first with a white box in the upper left corner, the other with a black box. A strict comparison finds agreement at a global level, but total disagreement at local level of observation. A human observer instantly recognises the similarity between the fields, but would have great difficulty in describing it with a value. Some observers would

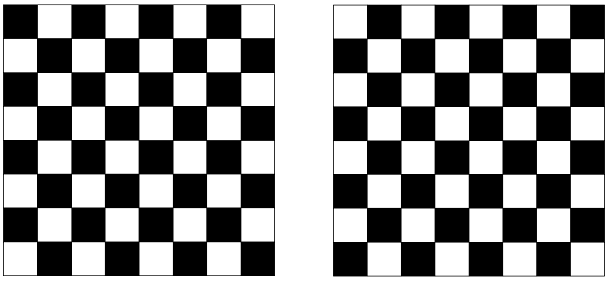


Figure 3.4 Contrasting checkerboards provide a good example of the contrasting approaches taken by visual comparison and computerised comparisons. Visual comparison will recognise the similarity between the fields, while a strict computerised comparison finds total disagreement

state that there is a shift, but otherwise the fields are identical, while others would state that the fields are opposite and therefore completely different.

The strict nature of computerised comparisons can make them unsuitable for certain tasks. For finding the absolute differences between fields, no generalisation is desired and strict methods are suitable. However, in many tasks, some tolerance for differences is necessary. In accuracy assessment for remote sensing, Foody (2002) recognises that there should be some level of positional tolerance incorporated into comparison methods. At present, if any alignment issues occur between fields, they will appear as model errors, even though they are often understood and tolerated by the users. Visual comparison has approximate ways of incorporating this, but it could be more specifically considered in the comparison algorithm. Hagen (2003) uses fuzzy set theory to incorporate mismatches in location and category value between fields. In hydrology, the fuzzy approach has been applied to a study of saturated area mapping (Güntner et al. 2004) so that locational differences that arise due to scale differences are tolerated. A multiscale approach can also be used to indirectly tolerate differences, because the larger scale versions of the original field will have generalised locations and values. Pontius et al. (2004b) described an approach that attributes the cause of the errors to either location or attribute differences, but does not produce a field showing where the problems occur. In contrast, the method of Hagen (2003) explicitly allows for the mismatches at each individual element and produces a field showing any errors. These types of tolerant methods are potentially applicable for hydrological purposes.

3.2.6 Emulating visual comparison methods

Qualitative visual comparison is widely used in hydrology because of: 1) the convenience of the method; and 2) the many powerful capabilities of the human visual system. Some of the powerful capabilities can be emulated by computerised methods, as described in the previous sections. By emulating these familiar approaches to comparison, it may be possible to calculate useful quantitative comparison methods that are very familiar to hydrologists (i.e. the methods can answer the same questions that were being answered visually). Table 3.1 summarises the key capabilities of visual comparison and the general methods identified to emulate them.

Table 3.1 Summary of visual analysis and comparison approaches and how they are used. The potential for emulating these methods using computerised methods is also summarised.

Aspect of visual comparison	Description of use and potential approaches for emulation
Evaluating multiple characteristics	Used to comprehensively assess where and how fields are different or similar; emulated by combining a variety of quantitative measures into an index or by using multi-criteria model assessment
Focusing on important parts of the field	Used to simplify field, either by changing the extent being analysed or by generalising the details within the field; emulated by subsetting field into important regions, as defined by the user or by characteristics of the field
Recognising organisation	Used for simplifying field into recognisable features or objects, thereby reducing the information being analysed and allowing other characteristics of regions to be used for comparison; emulated using segmentation or other image processing tools
Evaluating multiple scales	Used to identify 'correct' scale of analysis based on knowledge, or otherwise to analyse the field at a range of scales; emulated by combining quantitative measures from different scales, which depend on having a method for changing representation of field at different scales
Tolerance for differences	Used to recognise aspects that are 'slightly similar' between fields, or otherwise to ignore differences that are too minor to visually discern; emulate by making comparisons with other nearby elements or by comparing at multiple scales

In emulating any aspect of visual comparison, the computerised methods will require some parameters to be specified. For example, region detection methods need some criteria to determine what defines a region. Multiscale comparisons must have a method for scaling the data. Limiting the extent of analysis requires some decisions about what is important. To tolerate differences the magnitude of the differences must even be defined. During visual comparison, these values are innately applied based on the experience of the user. Over time, it is expected that similar experience could also develop for using computerised methods.

3.3 Current methods used in hydrology

The current, quantitative, computerised comparison methods used in hydrology make use of two general approaches: 1) characterisation followed by comparison (i.e. global comparison); and 2) comparison followed by characterisation (i.e. local comparison).

In the first approach, each spatial field is either characterised into a number (e.g. summary statistics, landscape indices) or otherwise characterised into a graph from which the characteristics are derived (e.g. variogram). These global characteristics are then numerically differenced to produce a measure of global similarity or error.

The second approach operates on each pair of elements that have a specific spatial relationship. The numerical relationship between a characteristic of one modelled and one observed element is used to derive a local measure between each pair. These local measures can then either be visually analysed as an intermediate field, or otherwise summarised into an error or similarity measure between the fields (e.g. mean absolute error).

The following sections describe the current methods used in hydrology, starting with those that 'characterise then compare' (i.e. global comparison methods). This is a thorough review of the comparison methods used for spatial field comparison in hydrology, including methods used for both continuous and categorical fields. Some of the methods are for interactive analysis of fields, but they are still reviewed here because they illustrate how hydrologists currently deal with comparison problems. In some of the following sections, the sets of example fields in Figure 3.5 have been used to illustrate interesting aspects of the comparison methods.

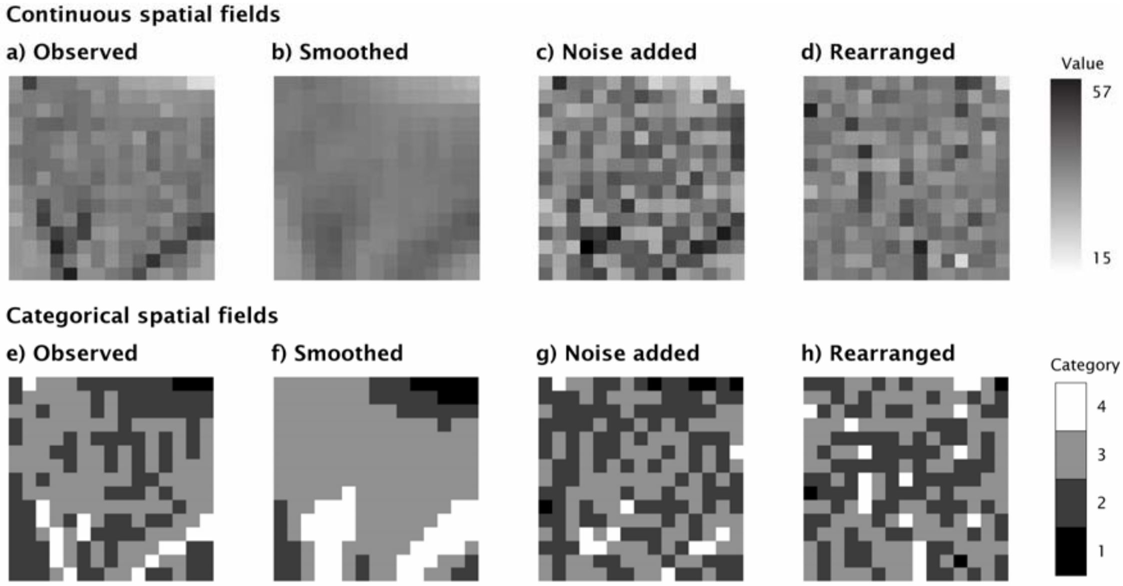


Figure 3.5 Example fields used to illustrate current comparison methods in hydrology. All fields are generated from the observed field (a). The other fields are: b) smoothed with mean filter; c) one realisation of random noise added (σ = 5); and d) one realisation of randomly rearranged elements. The categorical fields (e-h) are created from the continuous fields (a-d) by using four equal categories.

3.3.1 Statistical characterisation

The elements within a spatial field can be characterised (or summarised) into a numerical value or plot. When multiple fields are characterised in this way, the characteristics can be numerically differenced to describe how the fields compare. Throughout this section, the equations use x_i to denote the location of element I within the field, $f(x_i)$ to denote the value of element i and N to denote the number of elements in the field.

3.3.1.1 Basic statistical characterisation

Statistical characterisation treats every element in each spatial field equally, with no regard to their location or size/shape. Statistical characterisation focuses on describing the distribution of the element values, rather than capturing anything about their absolute arrangement (i.e. their interdependence). These methods are considered to work in measurement space. Hydrologists commonly use these methods to understand large datasets and subsequently draw conclusions from the data. When used to make comparisons between fields, they are most often used to evaluate bias and differences in variability or distribution.

The following equations (3.1-3.4) are commonly used to describe the distribution of continuous values within a spatial field. The mean describes the central tendency, standard deviation describes variability, skewness describes the symmetry of the distribution and kurtosis describes the peakedness of the distribution .

$$\text{MEAN} = \mu = \frac{1}{N}\sum_{i=1}^{N} f(x_i), \tag{3.1}$$

$$\text{Standard Deviation (SDEV)} = \sigma = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\left(f(x_i)-\mu\right)^2}, \tag{3.2}$$

$$\text{Skewness (SKEW)} = \frac{1}{N\sigma^3}\sum_{i=1}^{N}\left(f(x_i)-\mu\right)^3, \tag{3.3}$$

$$\text{Kurtosis (KURT)} = \frac{1}{N\sigma^4}\sum_{i=1}^{N}\left(f(x_i)-\mu\right)^4. \tag{3.4}$$

Table 3.2 The basic statistical characteristics used to describe the distributions of element values within the continuous spatial fields shown in Figure 3.5. BIAS is shown for comparison with the observed field.

Field	MEAN	SDEV	SKEW	KURT	BIAS
a) Observed	36.3	4.1	0.4	2.9	-
b) Smoothed	36.4	2.5	-0.7	1.5	0.1
c) Noise added	36.3	6.5	0.1	0.0	0.0
d) Rearranged	36.3	4.1	0.4	2.9	0.0

All of these measures are used to characterise the distribution of values in a spatial field, yet the higher-order characteristics (i.e. skewness, kurtosis) are often unreliable when outliers are present. Alternate summary characteristics, such as L-moments (Hosking 1990), can be used to better describe distributions when outliers exist, although such methods are less well-known. Where possible, many users inspect the empirical probability distribution functions (pdf) rather than summarising them. While this is useful, it is not a quantitative method.

The most common global comparison measure is bias (BIAS), which is the difference between the mean values for two fields. Other characteristics can also be numerically differenced but are not commonly used. There are some measures for comparing the probability density functions, such as the Kolmogorov-Smirnov test, which are detailed later (section 4.3.4.2).

All of these measures allow the distributions to be compared, but they only provide a summary of a spatial field. When categorical fields are used, summary measures more commonly contain counts or percentages of each category value. Table 3.2 shows the statistics for the continuous example fields from Figure 3.5. From this information, it can be observed that the smoothed field has less variation (but slight BIAS), while the noisy field is much more variable. Smoothing changed the skewness of the distribution substantially, while adding noise removed the obvious peakedness of the distribution. The random field has identical characteristics to the observed field, yet they are visually quite different. This has been done intentionally to illustrate the inability of basic statistics to incorporate any notion of spatial arrangement, yet they do identify the global changes between the fields.

3.3.1.2 Geostatistical characterisation

Geostatistical methods utilise the distance between elements to describe the relative arrangement of element values within a field. This is a similar approach to the lag correlation measures that are often used for analysing time series in hydrology. Journel and Huijbregts (1978) introduced the variogram as a measure of the relationship of a single variable over a certain distance or lag. The variogram is composed of measures of variance (the semivariance) that are dependent on the lag. It is defined by

$$\text{SEMIVARIANCE} = \gamma(h) = \frac{1}{2N_h}\sum_{i=1}^{N_h}\left[f(x_i) - f(x_{i+h})\right]^2 , \tag{3.5}$$

where N_h is the number of pairs of elements separated by the distance h. At any distance, $\gamma(h)$ should be evaluated against the global variance (σ^2). Where $\gamma(h) < \sigma^2$ there is strong spatial correlation, which gets weaker as $\gamma(h)$ approaches σ^2. When the semivariances are plotted against the lag, a sample variogram is created which can be analysed to reveal characteristics of spatial variability. The lags can be limited to certain directions during analysis, although omnidirectional variograms are used here.

The characteristics of a variogram can be used for comparing spatial fields, but variograms are more commonly used for interpolating and generating stochastic data. A thorough treatment of these applications is provided by Blöschl and Grayson (2000). An experimental variogram (i.e. that summarises a spatial field) can be prone to sampling issues (Skoien and Bloschl 2006) and therefore may not characterise the field adequately. Variograms are also poor at describing multiscale characteristics.

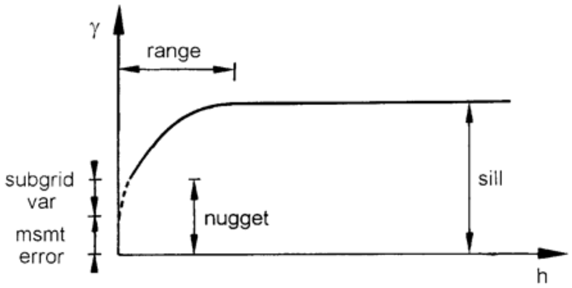


Figure 3.6 An idealised variogram, showing the definitions of the sill, range and nugget (adapted from Blöschl and Grayson 2000). In hydrological fields, the nugget is considered to be made up of measurement error and sub-grid (or small scale) variability.

Western et al. (1998) have used variograms extensively on observed spatial fields of soil moisture to understand spatial processes. They obtain the characteristics of a sample variogram by fitting it with a theoretical variogram model. From the theoretical variogram, the characteristics to describe spatial variability – the sill, the range and the nugget – are obtained (if they can be defined) (Figure 3.6). The sill is the semivariance at which the variogram flattens, but this only exists if the process is stationary (i.e. there is no spatial trend). The range is a measure of the distance over which spatial correlation remains, with large values denoting values that are smoothly varying over greater distances. The nugget describes the variance between pairs of elements at small lags and is often used to represent measurement errors and/or sub-element variability.

If the geostatistical characteristics of spatial fields can be defined, numerical comparison of the characteristics can describe how the spatial relationships within the fields differ. In Western et al. (1998), exponential variogram models (with a nugget) were successfully used to describe observed soil moisture fields. The characteristics (sill, range and nugget) were numerically compared between different fields and the change in the characteristics over time was visually analysed to describe trends. In general, variograms are not widely used for quantitative comparison in hydrology, but the work by Western et al. (1998) illustrates how theoretical variograms can quantify specialised characteristics of spatial fields. Extensions to the geostatistical approach that make use of categorical or binary data exist, but are not widely used in hydrology for comparison.

3.3.1.3 Characterisation using landscape metrics

When working with categorical fields, the options for undertaking comparisons change entirely. However, there is still a clear difference between characterisation and comparison. Categorical analysis is suitable when the element values are nominal or ordinal. In most instances, this will occur because of the observation method used. For example, binary fields are often observed in hydrology to map presence/absence of snow cover (e.g. Blöschl et al. 1991) or flood inundation (e.g. Horritt et al. 2001). Spatial fields can be converted from continuous to categorical, but this reduces the information content in the field and should only be done when sharp boundaries are required for use in analysis.

Characteristics of categorical fields have been used recently in hydrology by Güntner et al. (2004) to compare modelled and observed saturated area fields. They made use of landscape metrics to characterise the general size, shape and arrangement of saturated area patches. They then compared these characteristics (derived from observed and modelled binary fields) to provide quantitative support to their visual observations. The landscape metrics used were the mean patch size and mean shape index. Mean patch size is the average area of each contiguous region within each category, while mean shape index is the average of the shape index. The shape index commonly used is

$$\mathrm{SHAPE} = \frac{\mathrm{P}(\mathrm{r_i})}{4\sqrt{\mathrm{A}(\mathrm{r_i})}}, \qquad (3.6)$$

where P(r$_i$) is the perimeter of patch/region r$_i$ and A(r$_i$) is the area of that region (McGarigal and Marks 1995). A value of 1 indicates a perfectly compact shape, which is considered to be a square. These two metrics have been applied to the example fields from Figure 3.5, with the results listed in Table 3.3.

Table 3.3 The results of applying two landscape metrics to characterise the size and shape of the categorical example fields in Figure 3.5.

	Mean Patch Size (cells)				**Mean Shape Index**			
Field	**Cat 1**	**Cat 2**	**Cat 3**	**Cat 4**	**Cat 1**	**Cat 2**	**Cat 3**	**Cat 4**
e) Observed	3.0	6.0	1.6	1.8	1.0	1.2	1.5	1.0
f) Smoothed	8.0	7.0	3.8	1.8	1.2	1.2	1.2	1.3
g) Noisy	1.0	5.0	4.0	1.0	1.0	1.2	1.2	1.0
h) Random	1.0	4.0	6.0	1.0	1.0	1.1	1.2	1.0

The characteristics listed in Table 3.3 reveal that the patches (or regions) in the smoothed field are the largest, followed by those in the observed field. In all fields, the patches for categories 2 and 3 (i.e. the mid-range values) are largest. In terms of the patch shape, most of the category patches are close to square (i.e. close to 1), although those in the observed category 3 are more complex. The shape measures are influenced by having many single element patches in the noisy and random fields, thus resulting in an average size and shape close to a single square element. When derived for multiple fields, the numerical differences between these characteristics gives information about how the spatial field structures compare.

There are many more landscape metrics available than those currently used in hydrological applications. These include measures that aim to characterise overall fragmentation, patchiness, spatial cohesiveness, overall diversity, clumping and the amount of aggregation (Li et al. 2005). Research in the field of landscape ecology has frequently used these to characterise patterns in landscapes, such as the appearance of land use and forest plots (e.g. Lundquist et al. 2001). They are less frequently used for quantitative comparison, although Güntner et al. (2004) showed an application of the basic metrics to spatial fields of saturated areas in hydrology. A number of researchers debate the ability of landscape metrics to quantify patterns rigorously, although they agree that when clear boundaries exist, the simple patch based metrics (e.g. those used in Table 3.3) can provide a useful summary of overall structure (Gustafson 1998; Li and Wu 2004; Li et al. 2005).

In hydrology, as with most other disciplines working in the natural environment, the creation of clear boundaries in a continuous landscape is a subjective process. For example, Güntner et al. (2004) had to define what was considered a saturated area for their mapping, as well as defining the threshold at which a modelled element is considered as saturated. This issue makes the meaning of landscape metrics dependent on the methods used to define categories. If descriptions of landscape structure are desired in the natural environment, the patches (or regions) should either be observed as regions (e.g. levels of erosion (Jetten et al. 2003)) or otherwise defined using a rigorous and repeatable pre-processing method. Such methods are explored later in this review.

3.3.2 Local comparison

Local comparison is the most widely used method for comparing hydrological data sets, including time series and spatial fields. These methods all work at the finest scale, with a comparison being made for every pair of spatially coincident elements from the fields (i.e. from an observed and a modelled field).

3.3.2.1 Current methods used for continuous data

Hydrologists are frequently faced with comparing observed and modelled data, which can be achieved using a range of summary measures that are familiar amongst the hydrological community (e.g. Legates and McCabe 1999). For each pair of spatially coincident elements, a local measure of error or similarity is calculated and stored in an intermediate field (e.g. residual field). The intermediate field is usually summarised into a final measure, which is used for model assessment tasks. These are 'compare then characterise' methods, which are referred to through this thesis as local comparison methods.

The following four equations (3.7-3.10) are the most widely used comparisons in hydrology. Root mean squared error (RMSE) summarises the errors at each element, penalising larger errors more heavily (through squaring). Mean absolute error (MAE) also summarises error, but all errors are treated according to their magnitude (i.e. no additional penalty). The coefficient of determination, also known as R^2 (RSQ), describes the proportion of variance in the observed data that is explained by the modelled data. It evaluates how good the linear fit is between the two sets of element values, with a value of 1 denoting perfect agreement. The coefficient of efficiency (COE) evaluates the error in the modelled data against the variance present in the observed data. If the modelled data has less error than the variance of the observed data, COE will be a positive value (up to 1). If the model error is greater than the variance, the COE will be negative. These measures are calculated using

$$\mathrm{RMSE} = \left\{ \frac{1}{N} \sum_{i=1}^{N} \left[f(o_i) - f(m_i) \right]^2 \right\}^{0.5}, \tag{3.7}$$

$$\mathrm{MAE} = \frac{1}{N} \sum_{i=1}^{N} \left| f(o_i) - f(m_i) \right|, \tag{3.8}$$

$$\mathrm{RSQ} = \left\{ \frac{\sum_{i=1}^{N} (f(o_i) - \mu_{f(o)})(f(m_i) - \mu_{f(m)})}{\left[\sum_{i=1}^{N} (f(o_i) - \mu_{f(o)})^2 \right]^{0.5} \left[\sum_{i=1}^{N} (f(m_i) - \mu_{f(m)})^2 \right]^{0.5}} \right\}^2, \tag{3.9}$$

$$\mathrm{COE} = 1.0 - \frac{\sum_{i=1}^{N} (f(o_i) - f(m_i))^2}{\sum_{i=1}^{N} (f(o_i) - \mu_{f(o)})^2}, \tag{3.10}$$

where $f(o_i)$ is an observed element value, $f(m_i)$ is a modelled element value and $\mu_{f(x)}$ is the mean of all element values in a specific field (Legates and McCabe 1999). The continuous example fields (Figure 3.5) are compared using these methods, with the results listed in Table 3.4.

The observed field shows the values that denote perfect agreement with all of the measures. Lower values for RMSE and MAE indicate closer agreement, while for RSQ and COE the values close to 1 denote better fit. With the example fields, it is apparent that the noisy field, while having similar sized errors to the rearranged field, exhibits much better fit (or similarity) with the observed data. The RMSE for the rearranged field shows the impact that penalising large errors can have, as without the penalty the MAE shows no worse performance than the noisy field. The COE values indicate that only the smoothed field is a better representation of the observed field than the observed mean would be. This suggests that the errors in spatial arrangement in the noisy and rearranged fields cause greater errors than exist by having no spatial variation.

Legates and McCabe (1999) discussed these measures in regard to their ability to assess goodness-of-fit between modelled and observed data. They found that the measures each have individual strengths, but that no single measure can fully describe how fields compare. For assessing absolute error, both RMSE and MAE provide a measure of the typical error. For judging similarity between the datasets, the use of a normalised measure such as COE is favoured, as it can be evaluated relative to a known benchmark. In contrast, RSQ values must be evaluated based on experience, making interpretation across a range of different data sets more difficult. These measures form the basis of any current quantitative comparison between spatial fields in hydrology, but are also

Table 3.4 The widely used local comparison methods (RMSE, MAE, RSQ and COE) are used to compare the continuous example fields from Figure 3.5 to the observed example field.

Field	**RMSE**	**MAE**	**RSQ**	**COE**
a) Observed	0.0	0.0	1.0	1.0
b) Smoothed	2.7	2.0	0.6	0.6
c) Noise added	5.0	4.3	0.4	-0.5
d) Rearranged	5.9	4.3	0.0	-1.0

routinely used in many other disciplines.

3.3.2.2 Current methods used for categorical data

Local comparison methods work in the same general way for categorical and continuous fields – that is, each pair of spatially coincident element values are compared and their comparison result is stored in an intermediate measure which is then summarised. For categorical fields, the inability to numerically compare nominal or ordinal category values means that an intermediate field of local error measures cannot be created. Instead, the only information that can be stored on a per element basis is whether the elements match or not. This intermediate field can be summarised to give a measure of ‘percentage correct’ (PCOR), which is the most basic comparison measure for categorical fields. PCOR can also be determined by adding the diagonal elements from the contingency table (Table 3.5).

Table 3.5 The structure of the contingency table for four categories (Pontius et al. 2004b). Components of the table are denoted by $P_{i,j}$, which is the proportion of elements with observed category i and modelled category j. The $_{+}$ denotes all category values.

	Modelled				**Total**	**Loss**
Observed	**Cat 1**	**Cat 2**	**Cat 3**	**Cat 4**		
Cat 1	$P_{1,1}$	$P_{1,2}$	$P_{1,3}$	$P_{1,4}$	$P_{1,+}$	$P_{1,+}$ - $P_{1,1}$
Cat 2	$P_{2,1}$	$P_{2,2}$	$P_{2,3}$	$P_{2,4}$	$P_{2,+}$	$P_{2,+}$ - $P_{2,2}$
Cat 3	$P_{3,1}$	$P_{3,2}$	$P_{3,3}$	$P_{3,4}$	$P_{3,+}$	$P_{3,+}$ - $P_{3,3}$
Cat 4	$P_{4,1}$	$P_{4,2}$	$P_{4,3}$	$P_{4,4}$	$P_{4,+}$	$P_{4,+}$ - $P_{4,4}$
Total	$P_{+,1}$	$P_{+,2}$	$P_{+,3}$	$P_{+,4}$	1	
Gain	$P_{+,1}$ - $P_{1,1}$	$P_{+,2}$ - $P_{2,2}$	$P_{+,3}$ $P_{3,3}$	$P_{+,4}$ - $P_{4,4}$		

More advanced local comparisons record the differences between spatially coincident elements in an intermediate measure called a contingency table (or confusion matrix) (Monserud and Leeemans 1992; Foody 2002; Pontius et al. 2004b). This table (Table 3.5) contains values describing the proportion of elements in the fields that have each possible combination of category values. The values within the contingency table are the basis from which summarised comparison measures are determined for categorical fields. In this way, it performs a similar role to the residual field when comparing continuous fields.

The most widely used comparison measure for categorical fields is Cohen’s ‘kappa index’ (Cohen 1960):

$$\text{Kappa index (KAP)} = \frac{\sum_{i=1}^{k} P_{i,i} - \sum_{i=1}^{k}\left(P_{+,i} * P_{i,+}\right)}{1 - \sum_{i=1}^{k}\left(P_{+,i} * P_{i,+}\right)}, \tag{3.11}$$

where k is the number of categories and the proportion values are obtained from the contingency table (Table 3.5). This index uses the PCOR value (i.e. the sum of $P_{i,i}$) and adjusts it for any chance agreement that could be expected between categories (based on chi-square). A value of 1 denotes perfect agreement. It is widely recognised that PCOR can be heavily influenced by the number of elements within each category. When one category is dominant, PCOR is often an artificially high value (Foody 2002). This is where the kappa index is preferred and it is considered a standard for accuracy assessment between categorical fields. Both PCOR and KAP are calculated in Table 3.6 for the categorical example fields from Figure 3.5.

Table 3.6 The percentage correct and kappa measures are calculated for the categorical example fields from Figure 3.5. The KAP values have the range [-1, 1].

Field	PCOR	KAP
e) Observed	100.0%	1.00
f) Smoothed	58.7%	0.31
g) Noise added	56.9%	0.25
h) Rearranged	39.6%	-0.05

The percentage correct measure shows that both the smoothed and noisy fields are evaluated as almost equally similar to the observed field. However, in both cases, the kappa values show less agreement (as they are normalised by chance agreement). The randomly rearranged field obtained a near zero score for KAP, indicating that it offers no improvement over chance agreement (as would be expected), while the PCOR value still finds almost 40% agreement.

In hydrology, categorical measures based on the contingency table are used mainly for comparing binary or categorical spatial fields. One major application is for flood inundation modelling, where methods are being used to compare observed inundated areas with categorised flood surfaces (e.g. Aronica et al. 2002; Bates et al. 2004). These studies generally create their own, specialised measures from the contingency table, so that false positives and/or true negatives (e.g. elements that are correctly modelled as dry) do not confound the comparison measures (e.g. Hunter et al. 2005). The contingency table can be used as the basis for many other comparison measures, such as the measures of skill used in meteorology (Jolliffe and Stephenson 2003). Some further discussion about how to interpret the contingency table is given in the following section.

3.3.3 Intermediate measures

Local comparisons compare every pair of elements and then summarise the results into a single measure. This is the opposite approach to global comparison. During local comparison, every element location is compared explicitly, which allows the variability of error or similarity to be analysed. A number of simple analysis methods are used to do this, which are collectively termed intermediate measures. These include residual fields, scatterplots and the contingency table. Intermediate measures are usually analysed visually (i.e. as a field or a graph) and are particularly useful for explaining the causes behind a specific comparison measure.

3.3.3.1 Intermediate and residual fields

An intermediate field allows the spatial distribution of the local error or similarly values to be analysed. The field is produced during local comparison, with each element containing the comparison measure computed between the observed and modelled fields. When the numerical difference (or residual) is used as the comparison measure, the field is commonly known as a residual field. Intermediate fields show the location and magnitude of all error or similarity values.

Intermediate fields are usually analysed visually. Residual fields are often visualised using a dichromatic colour ramp, with red for positive values and blue for negative values. Visual analysis is often used to explain the reasons why a comparison measure (i.e. a summary of the residuals) obtained a specific value. For example, if the RMSE is large, the residual field can reveal whether this was caused by some elements with very large errors, or whether it was an equal contribution of error from all elements. Intermediate fields are widely used in research papers for presentation of results and for discussion. However, as with visual comparison of fields, visual analysis of intermediate fields is neither repeatable nor rigorous.

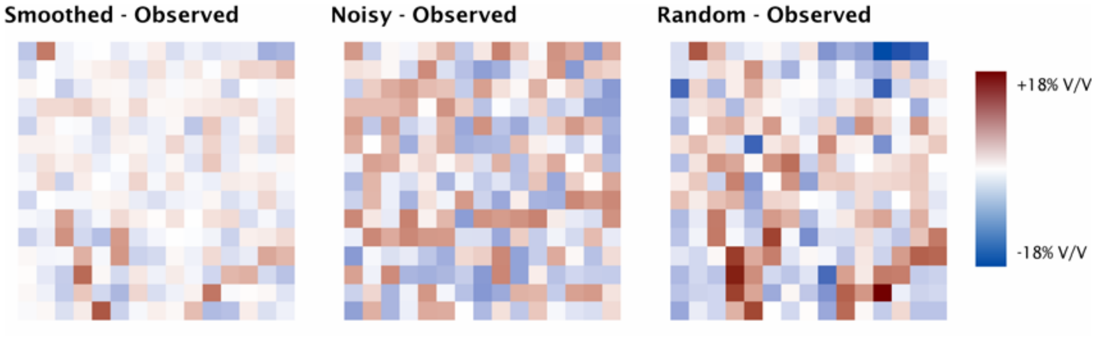


Figure 3.7 Residual fields produced by comparing the continuous example fields from Figure 3.5 against the observed field.

Residual fields are shown in Figure 3.7 for local comparisons of the continuous example fields from Figure 3.5. The colour ramp used allows over- and under-prediction to be visually assessed. It is apparent that the smoothed residuals are lower in value than all of the other fields. The residuals from the smoothed and noisy fields lack any clear spatial structure, whereas the residuals from the randomly rearranged field show structure similar to the original observed field. This illustrates a situation where the model (i.e. the rearranged field) fails to represent the spatial arrangement of the observation well, which causes the arrangement to remain in the residuals.

Basic statistical characterisation can also be used to summarise the distribution of values in the intermediate fields. These can give useful summary measures of individual fields, such as the error variance. Geostatistical characterisation of the residual fields is also a useful way of evaluating the residuals. Western et al. (1999a) used this approach to assess a spatially-distributed soil moisture model. When clear spatial correlation remained in the residual fields (produced via local comparisons of modelled and observed soil moisture fields), the model performance was deemed to have poor spatial predictive performance.

3.3.3.2 Scatterplots

A common method for comparing observed and modelled data sets is to plot the values on a scatterplot (or dot plot). Each element is plotted using the observed value as the x-coordinate, and the modelled value as the y-coordinate. If the values are identical, this produces a linear pattern in the scatterplot. The scatterplot is an intermediate measure that is related to the RSQ measure. It visualises the errors between spatially coincident elements relative to a linear fit.

Figure 3.8 shows the scatterplots for the comparisons between the continuous example fields (from Figure 3.5). It is apparent that the noisy field has a linear trend, but is widely spread. The smoothed field appears like a compressed and slightly modified version of the linear trend. In contrast to these, the randomly rearranged field is scattered widely and shows poor agreement with the observed field.

Scatterplots can also be used to analyse relationships between individual element errors and other surrogate variables. For example, the residual field can be calculated for a pair of observed and modelled fields. The residuals can then be plotted against another surrogate field using a scatterplot. This can help to analyse if any relationship exists between the error and another variable. It is particularly useful in hydrology, where model processes are often related to physical attributes that are available as spatial fields, such as terrain or climatic attributes.

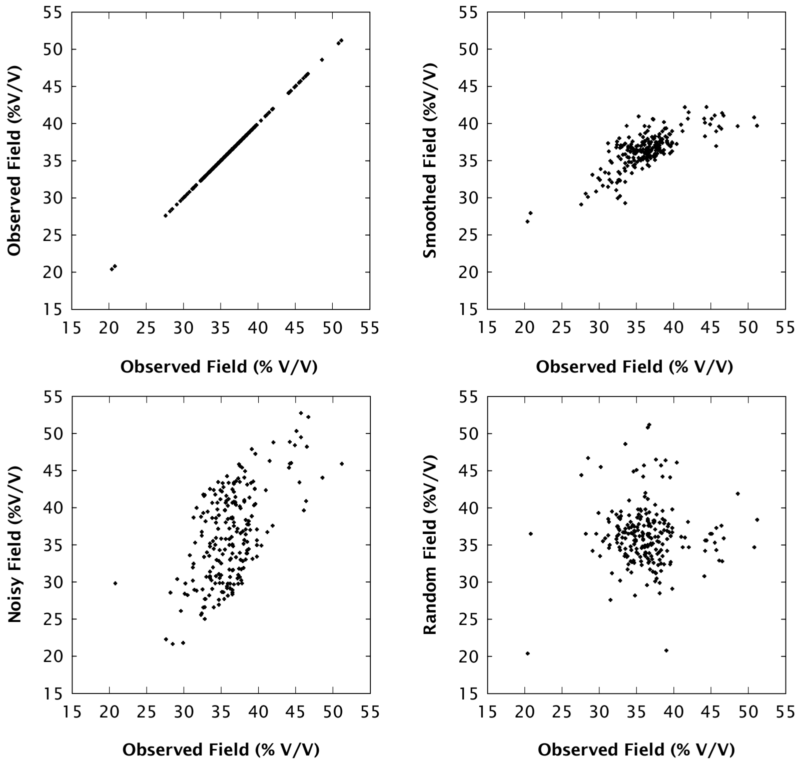


Figure 3.8 Scatterplots for comparing each continuous example field in Figure 3.5 to the observed field.

3.3.3.3 Contingency table

The contingency table explained earlier is an intermediate measure for categorical comparisons. It can be analysed in many different ways, to reveal information about how individual categories have changed between two fields. This table is still an active research topic in land use change and accuracy assessment (e.g. Pontius et al. 2004b; Foody 2006). When analysing fields with more categories (e.g. land use maps), the confusion matrix can be used to determine useful information about change for each category (e.g. how much gain, loss or swapping has occurred). Pontius et al. (2004b) have also shown that producing confusion matrices at multiple scales can enable the distance over which the changes occur to be determined.

In hydrology, the contingency table is predominantly used as the basis for deriving comparison measures between binary fields (see previous section), rather than as an intermediate measure that is analysed more thoroughly. The simple nature of the categorical fields used in hydrology (i.e. binary) limits the applicability of these recent advances in analysing categorical fields.

3.3.4 Cross-section analysis

One of the difficulties when dealing with spatial fields is the increased number of degrees of freedom that exist with any processing. During time series analysis, time can be simply plotted against value. However, during spatial field analysis or comparison, location requires two axes for representation, therefore making it more difficult to plot. Due to the familiarity with analysing and comparing time series in hydrology, it is often desirable to simplify a field into a cross-section or transect, thus enabling simpler visual analysis.

Grayson et al. (2002) identifies that transects, when placed so that they follow important terrain features, can reveal shifts and highlight problems with model structure. For example, they analysed a transect placed across a gully to inspect whether model error was related to terrain position. Blöschl et al. (1991) used transects to compare fields of binary snow cover observed from aerial photography with modelled snow water equivalent, and to see if they were related to elevation. The transects used in both these examples are shown in Figure 3.9. The soil moisture example reveals better model

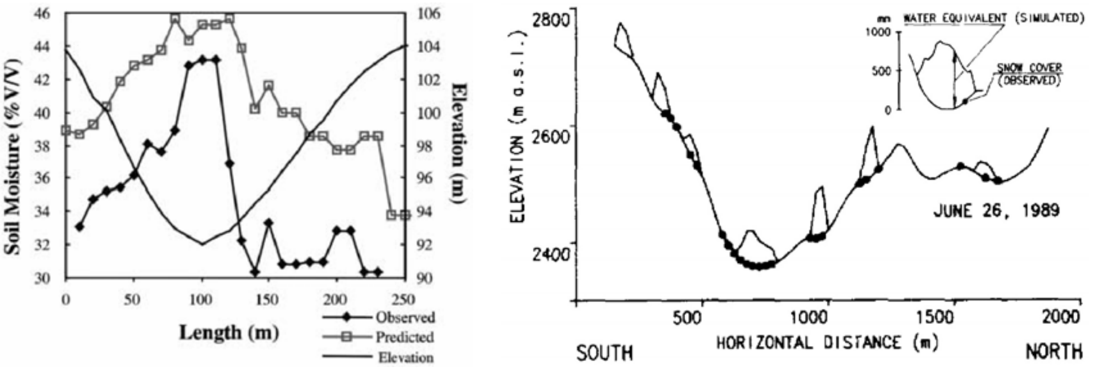


Figure 3.9 Two examples of transects being used to analyse spatial fields are shown. Observed and modelled soil moisture (left) is compared with elevation. Observed snow cover (right) is compared against modelled snow water equivalent and elevation (adapted from Blöschl et al. 1991; Grayson et al. 2002).

agreement at the bottom of the gully than on the hillslopes. The snow example shows that the observed snow cover did not match the model in high elevation areas. These interactive forms of comparison are useful, but the results are heavily dependent on the location of transects. For example, detecting shifts (in any direction) requires carefully placed transects.

This type of intermediate measure is considered as a useful analytical tool for explaining why error or similarity is occurring. It has limited value for quantitative comparison, but does represent a more specialised comparison method used within hydrology.

3.3.5 Comparison of 'analysis windows'

The majority of methods used for comparing spatial fields treat every element independently of the neighbouring elements. However, to compare the spatial relationships within spatial fields, neighbouring elements should be considered in the comparison method. Defining an 'analysis window' is a common way of doing this. The analysis window specifies which neighbouring elements around each 'focus' element will also be used during analysis. As each focus element in a field is processed, the window of neighbouring elements changes accordingly. This allows information about spatial relationships (e.g. other nearby values) to be included during analysis (or comparison).

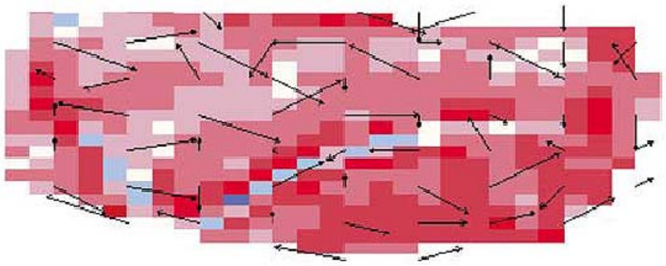


Figure 3.10 An example of the output from an optimal local alignment comparison method. The arrows start from an observed element and represent the direction and magnitude of the most likely matching element in the modelled field (based on the window of nearby elements). The details of the underlying spatial field are irrelevant for the purposes of this example (adapted from Grayson et al. 2002).

A comparison method termed 'optimal local alignment' was introduced by Grayson et al. (2002) to enable the investigation of locational shifts between spatial fields. This method uses techniques from the image processing discipline of optical flow analysis (Barron et al. 1994). The optimal local alignment method determines the direction and magnitude of the most likely shift (i.e. the optimal fit) for each 'window' of elements in the observed field. A correlation measure is calculated between each 'observed window' and every possible 'modelled window' (i.e. a window that is moved across the entire modelled field, element by element). The pair of windows that have the highest correlation are used to define the optimal local alignment. The output of this method is a spatial field of vectors showing the direction and magnitude of the most likely shift (Figure 3.10).

The user can visually analyse this output to identify areas with consistent shifts, which may be explained by issues with the model and/or observations. As with transects, this approach is an intermediate measure that is useful for understanding how fields compare, but not for summarising the overall comparison for use in model calibration or testing. Optimal local alignment requires variation within the spatial field for it to be able to recognise shifts. Therefore, with binary categorical fields, shifts can only be recognised at edges (i.e. where there is variability). This is due to the use of a cross-correlation measure to evaluate the fit between the groups of elements.

Another method that utilises an analysis window to include spatial relationships into comparison has been recently applied in hydrology (Güntner et al. 2004; Wealands et al. 2005a). This comparison method subjectively defines the degree of 'locational similarity' between an element and each of its neighbours. This is a form of fuzzy comparison, as originally detailed by Hagen (2003). Using a fuzzy definition for location, an element might be defined as having similarity of 0.8 with its immediate neighbours. The fuzzy definition effectively specifies the size of the analysis window, which contains all elements considered to be similar in location. In the example shown in Figure 3.11, the similarity of location is defined by a linear function that decreases with distance (up to a radius of 5 elements). Güntner et al. (2004) used this method to compare modelled and observed binary fields of saturated areas.

As Figure 3.11 shows, each modelled element is compared against all observed elements that are within the analysis window. If a matching value is found within the window, it is assigned a local similarity based on the fuzzy definition of locational similarity. This allows some level of spatial relationship to be included into comparison, which is considered vital for advancing spatial field comparison methods (Wealands et al. 2005a).

This approach has only been applied to binary fields in hydrology and categorical fields in other disciplines (as discussed in the following sections). While it has not been used further in hydrology, it is a very promising comparison method. A comparison method using the idea of analysis windows and fuzzy similarity is pursued further in this thesis for application with continuous hydrological fields.

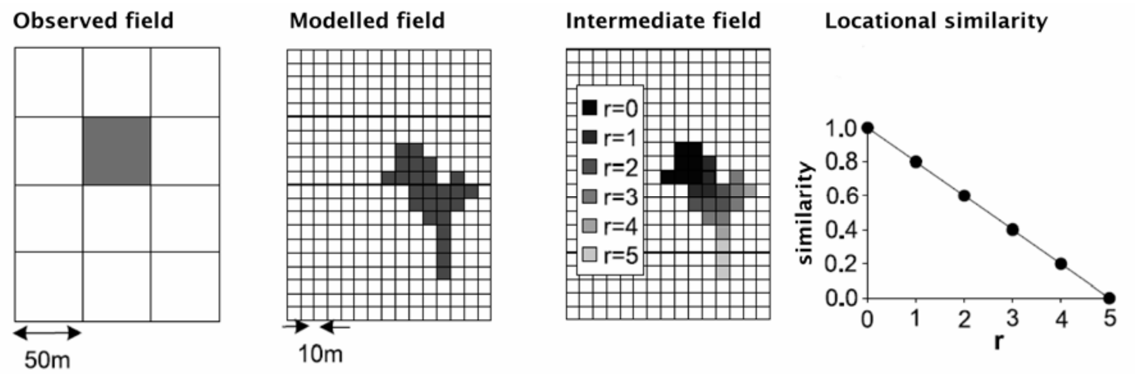


Figure 3.11 An example of fuzziness of location, in which the modelled field is compared with an observed field (collected at a coarser resolution). In the intermediate field of similarity values, a low value (dark shading) indicates better agreement (higher similarity) than a high value (light shading). Note that even though some modelled elements are not spatially coincident with the observed element, they are still assigned similarity (adapted from Güntner et al. 2004).

3.3.6 Limitations of current comparison methods

This section has described the variety of quantitative methods that are currently used for comparing spatial fields in hydrology. Some of these methods, such as basic statistical characterisation and standard local comparison measures (e.g. RMSE), are widely used and will continue to be used because of their simplicity and versatility to any data set. Other methods, such as geostatistical characterisation, are potentially useful but often overlooked because they are seen as too difficult to compute and require experience to interpret. It certain situations where numerical measures are not required, intermediate measures (e.g. residual fields, scatterplots, transects) are combined with visual analysis to help understand how the error or similarity varies across the field. The most recently developed methods, which use analysis windows to consider spatial relationships, are promising examples of where spatial field comparisons can progress. Table 3.7 summarises the current comparison methods, their general uses and major limitations.

Table 3.7 Summary of the current methods used for quantitative comparison in hydrology. The shaded methods are used further throughout this thesis

Current comparison method	Description of use and limitations
Basic statistical characterisation	Used to summarise element values within a field into characteristics, which can then be numerically compared with other fields; operate at global scale in measurement space only, thereby ignoring the spatial arrangement of elements
Geostatistical characterisation	Used to characterise the spatial relationship between elements within a field; characteristics of variograms can be numerically compared; does not explicitly retain information about spatial arrangement
Landscape metrics	Used to characterise various aspects of contiguous regions within a categorical field; requires region boundaries to be defined if applied to continuous fields; handles information at an intermediate scale by deriving characteristics of regions rather than elements
Local error and similarity methods	Widely used through hydrology for comparing any type of temporal or spatial data set; enforces strict spatial coincidence during comparison, while ignoring spatial relationships with neighbours; determines error or similarity measures for every element, which are summarised to give comparison measure; works efficiently at the finest scale
Contingency table	Widely used for local comparison of categorical fields; provides information for understanding matches and mismatches between equivalent categories; used to derive summary measures for binary and categorical field comparisons; can also be used as intermediate measure for analysing category changes

Current comparison method	Description of use and limitations
Intermediate fields	Used for analysing the arrangement of error or similarity values within a field, which combine to produce local comparison measures; can be characterised using basic statistics or geostatistics to give useful summaries of the comparison; useful for visually understanding which elements are most influential in resulting comparison measures
Scatterplots	Used to examine the relationships between values from both fields, or otherwise between residuals and surrogate fields; works at local scale and is strictly for spatially coincident elements
Transects	Used to simplify representation of spatial field, by examining only one cross-section and the values along it; analysis results are heavily reliant on placement of transect
Analysis windows	Used to perform analysis and comparison on groups of spatially related elements, rather than individual elements; allows spatial relationships to be explicitly considered during comparison; methods have had limited development for hydrological spatial fields (i.e. continuous fields)

At present, local and global comparison methods address the extremes of comparison, in terms of both scale (i.e. the finest and coarsest level of comparison) and rigour (i.e. how strictly the comparison is performed). However, the current methods generally fail to address the gap in between these extremes, which includes analysis at intermediate scales (e.g. regions) and the consideration of explicit spatial relationships and arrangement (e.g. recognising similarity of location). These issues are currently managed by using visual methods. To identify quantitative methods for addressing them, the remainder of this review looks at other disciplines for potential solutions. The knowledge from other disciplines can then be combined with the current state-of-the-art in hydrology to define a plan for advancing spatial field comparisons in hydrology.

3.4 Comparisons within other disciplines

The task of producing a quantitative measure of similarity between two spatial fields is encountered in other disciplines, ranging from image processing and pattern recognition to landscape ecology and multimedia. Other disciplines generally work with fields that are single- or multi-band images. These disciplines usually rely on a number of fundamental methods for processing images and conducting comparisons, some of which could be applied to hydrological applications.

The suitability of a new method depends on whether it can work with hydrological spatial fields, which have particular characteristics such as the lack of obvious and definable features and the presence of moderate to high noise. These demands are different to other disciplines dealing with comparison, which generally work with pre-determined images or features (e.g. a face or fingerprint), operate under controlled imaging conditions (i.e. with minimal noise and uncontrolled changes) and must be performed very quickly (i.e. the details are often simplified).

The reason for making a comparison varies slightly between disciplines, although in all cases it is used for reducing the complexity (or amount) of results to be analysed. In content-based image retrieval, comparisons are done to find the most likely match from within a database of images (e.g. an image search engine) (Pauwels and Frederix 1999). Industrial applications use comparison to operate machinery on a production line, such as for rejecting products that do not match a prototype closely enough (Veltkamp and Hagedoorn 1999). Surveillance applications also make use of comparison for recognising change between subsequently captured images (Radke et al. 2005). Land-use and land-cover change is an integral component of geographic, economic and ecological research that uses comparison to assess the accuracy of a model (i.e. define the error) and to identify where change has occurred (Pontius and Malizia 2004).

All of these disciplines are particular application areas for the comparison of images (or spatial fields), as is hydrology. Their ongoing need for comparison has led to the refinement of many common processing methods, most of which stem from the broad discipline of image processing. When referring to methods from image processing, the terms image and pixel are often used instead of spatial field and regular element, but they are actually synonymous. The following sections describe the comparison methods used in other application areas and how their specific approaches may be suited to hydrology. This review is organised into the following sections: 1) the fundamental image processing methods used during comparison; 2) the structural features used in other disciplines; and 3) change detection methods used in other disciplines.

3.4.1 Image processing for comparisons

Image processing encompasses many fundamental methods that are used throughout other disciplines as part of the comparison process. These methods are used to undertake the key tasks of conditioning (i.e. modifying pixel values) and grouping (i.e.

defining the organisation or structures present within an image) (Haralick and Shapiro 1992). Both of these are general pre-processing tasks that prepare images for use in different types of comparisons. Hydrologists routinely use conditioning methods for making spatial fields ready for comparison (e.g. georeferencing, upscaling spatial fields). In some situations, conditioning may be the only pre-processing needed before comparison. Other comparison methods work with structural features from the fields, so grouping methods need to be applied. These methods are not widely used in hydrology. The following sections detail the methods used for conditioning and grouping in other disciplines.

3.4.1.1 Conditioning an image

Conditioning methods are generally used for removing artefacts of prior processing and correcting known problems. Conditioning methods that work purely in measurement space (i.e. with the distribution of values only) include standardisation and histogram matching. These methods are used to change the pixel values so that their numerical distribution meets certain criteria. Standardisation makes the distribution have mean = 0 and variance = 1. Histogram matching is a more involved process that makes the cumulative distribution function (cdf) of one image match the cdf of another. Both of these methods are useful when trying to compare images, as they make sure the measurement spaces coincide (to some degree) (Hadjidemetriou et al. 2004). For hydrology comparisons, changing the distribution of elements may be undesirable because it removes the real-world interpretation of the values (which are generally important). These methods are more suitable when the magnitude of the element values does not have a particular interpretation or when the element values being compared represent different attributes (e.g. comparing soil moisture to a vegetation index). In these cases, assuming a particular relationship between the distributions (e.g. both have equal mean and variance) is needed to enable comparison. The RSQ correlation measure implements a form of standardisation implicitly in its derivation (Legates and McCabe 1999).

Conditioning methods that work with image space information (i.e. consider spatial location and relationships) include noise filtering and smoothing. Both of these methods commonly use ‘analysis windows’ to incorporate information from neighbouring pixels. Noise filtering modifies pixel values that are dramatically different from their neighbours by replacing them with a value calculated from the neighbourhood (such as the average) (Mastin 1985). In hydrology, the removal of noise is commonly achieved using these methods or geostatistical interpolation. Both methods can also be used for upscaling (or smoothing), which uses neighbours to determine the upscaled pixel values. These methods are quite widely understood and are not detailed further in this thesis.

3.4.1.2 Defining structural features within an image

Grouping methods are generally used to process an image into a set of structural features that are required for comparison or recognition. Structural features are often defined by using edges that are apparent in the image. Edge detectors use the change in pixel values in certain directions within a neighbourhood to assign an edge strength value to each pixel. For example, edges are useful for enhancing the dominant features in facial images (e.g. eyes, mouth), as these features appear as sudden changes in contrast and produce strong edge response. The edge detection methods use a number of different approaches (see Haralick and Shapiro 1992; Bow 2002), although they tend to be sensitive to image noise. They are infrequently applied to natural scenes due to a lack of clearly defined edges in the natural environment, although Horritt et al. (2001) have used edge features (combined with additional image processing) for identifying flood inundation extents from SAR imagery. Hydrological spatial fields are often smoothly varying, thus limiting the use of edge-based methods for region detection. A more common approach for detecting structural features or regions (which ideally represent objects) is image segmentation. This is the process of partitioning an image into a set of non-overlapping regions. After segmentation, every pixel within the image will be assigned to a particular region.

Segmentation can be achieved using knowledge-driven or data-driven approaches. In knowledge-driven approaches, the image is analysed to extract a predetermined type of target object, such as finding individual tree crowns in a satellite image of a forest. Hydrological fields rarely have such pre-defined target objects, although pre-defined target regions could potentially be extracted from surrogate fields, such as a terrain model (e.g. to define a low slope feature). In data-driven approaches, the image is segmented into 'image objects' (Hay et al. 2003) based on some additional information. Some of the common methods used for data-driven segmentation are thresholding, clustering, region growing and region merging. A range of methods are covered in the reviews about segmentation and its various applications (e.g. Pal and Pal 1993; Chang and Li 1994; Jain et al. 1999; Udupa and Saha 2003). One of the important points from the broader literature on comparisons and segmentation is that there is no single best method (Smeulders et al. 2000), although some are more applicable to the types of data encountered in hydrological applications than others.

3.4.1.3 Segmentation using pixel values

One of the most common and simple segmentation methods is thresholding (Haralick and Shapiro 1992). Thresholding involves defining a fixed value that then determines whether each individual pixel is labelled as high or low. Figure 3.12a shows a spatial pattern of soil moisture that has been grouped using a threshold of 38% volumetric soil moisture. This data are from the Tarrawarra data set presented in Western et al. (1999b). The threshold used was manually selected to identify the wet gully areas of the image, whereas other approaches exist that adaptively compute the threshold from

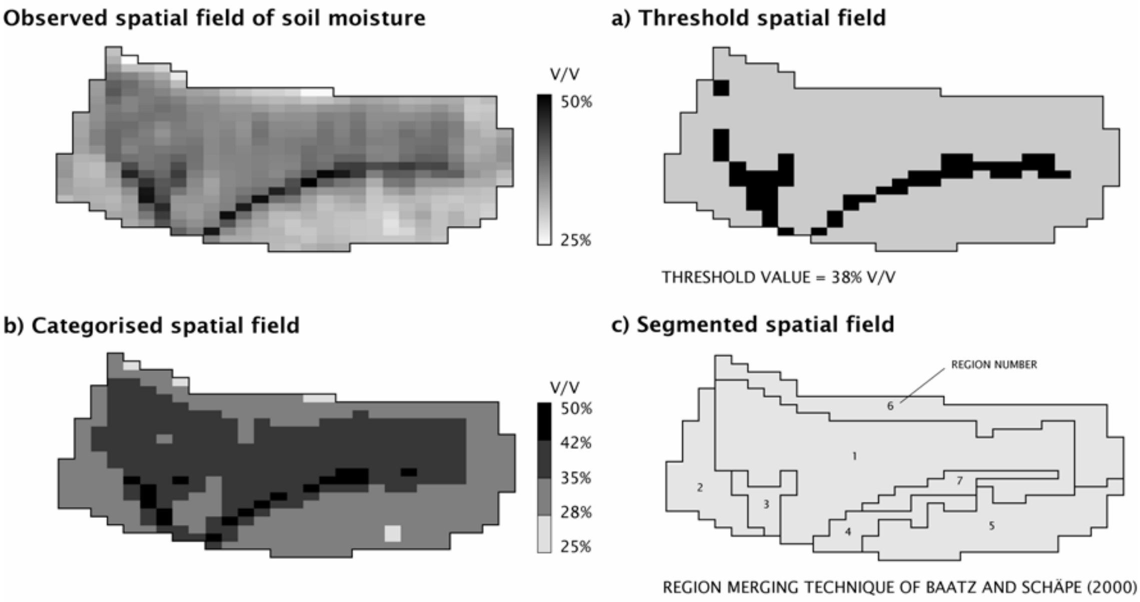

Observed spatial field of soil moisture
V/V
50%
25%
a) Threshold spatial field
THRESHOLD VALUE = 38% V/V
b) Categorised spatial field
V/V
50%
42%
35%
28%
25%
c) Segmented spatial field
REGION NUMBER
6
1
2
3
7
4
5
REGION MERGING TECHNIQUE OF BAATZ AND SCHÄPE (2000)

Figure 3.12 Identification of regions within a spatial field of soil moisture using three alternate methods: a) histogram thresholding; b) histogram slicing; and c) region merging segmentation (from Wealands et al. 2005a).

the image histogram (e.g. the threshold can be computed such that particular percentages of pixels fall between threshold values). Thresholding is useful for separating the pixels of interest from the background, but requires visual inspection to determine the optimal threshold value. Pal and Pal (1993) provide a comprehensive review of thresholding.

Categorisation segments an image by grouping pixel values into pre-determined ranges, assigning pixels to categories using information from measurement space. Thresholding is the simplest type of categorisation. After assigning new values to each pixel, they are grouped into connected regions using connected components labelling (Haralick and Shapiro 1992). This is a process whereby pixels that are connected (i.e. form a spatially contiguous region) are given a common identifier. For automated categorisation, the histogram of the spatial pattern is usually analysed and the category ranges calculated based on the characteristics of the histogram. Figure 3.12b shows a spatial pattern of soil moisture after being categorised into four groups. Notice that the regions representing some categories have ‘holes’ in them, thus producing a noisy looking segmentation. Because this method works entirely in measurement space, the resultant regions are not always spatially contiguous.

Clustering is another popular approach to segmentation that relies more on pixel values (i.e. measurement space) and less on their spatial arrangement. Clustering maps the pixel values from multiple bands into a feature space and then uses an algorithm to identify the groups of pixels with features that are compactly grouped and isolated from other features (Jain et al. 1999; Pauwels and Frederix 1999). The most common clustering method is K-means clustering (MacQueen 1967; Han et al. 2001), which is used because it is a very fast process. This algorithm makes an initial estimate of K clusters and then shuffles pixels between the clusters until every pixel in every cluster is closest in value to its cluster than to any other cluster. This is the optimal segmentation and the pixel values are updated to reflect the cluster they belong to. Depending on the clustering algorithm used, the resulting segmentation may be strong (with logical regions being grouped in the image) or weak (with many small, disjointed regions). Multi-band images provide a large number of features and are well suited to clustering approaches. However, most fields used in hydrology have only one band of data and clustering them using only measurement space information is limited.

3.4.1.4 Segmentation using pixel values and locations

The most promising segmentation approaches for simple spatial fields are those that operate with both the pixel values and their locations (i.e. in both measurement and image space). These methods use measurement space to calculate the homogeneity criteria when determining regions. Region growing is one example, in which initial 'seed pixels' are chosen and then iteratively grown (e.g. based on joining pixels that differ in value less than a specified amount) until the entire image is segmented (Chang and Li 1994). This technique is sensitive to the choice of seeds and is not widely applicable. By considering every pixel as a seed, the method of region merging has evolved and become more popular and useful. The region merging method tries to find the best possible merge between neighbouring regions (i.e. closest in value) at each iteration through the image (Woodcock and Harward 1992; Baatz and Schäpe 2000). This iterative process eventually converges when there are no more merges that could proceed without violating the merge criteria. In this method, the resulting segmentation is dependent on the criteria for homogeneity (i.e. to determine which merges are acceptable) and the order of processing (i.e. to determine which acceptable merges are performed). Other criteria can also be used, which control the size or shape of the resultant segmented regions (Baatz and Schäpe 2000).

Region merging suits a wider range of images than other segmentation techniques, having proved successful with simple grey level images (i.e. similar to most hydrological spatial fields) as well as multispectral images. Figure 3.12c shows a segmented spatial pattern of soil moisture, using the region merging technique of Baatz and Schäpe (2000). This segmentation method is adapted for use with hydrological fields later in this thesis. This result is visually the most logical grouping of all examples shown in Figure 3.12. It is also the most complex and requires more processing than the other methods, which can be disadvantageous when analysing many spatial fields (such as in image database applications). This is one of the major limitations with segmentation methods that preclude their widespread usage for computer vision tasks, although developments in computing hardware are making many image processing tasks more viable.

3.4.2 Comparisons using structural features

Fast and accurate image comparisons are essential in any type of content-based image retrieval application. One typical application is facial recognition, in which a single query image is provided (e.g. the face being photographed) and then a large database is searched for images with similar features. Comparisons for facial recognition have to manage changes in the illumination of the face, different expressions and changes due to ageing, so approaches that detect structural features in the facial image must be robust to these varied conditions. Pujol et al. (2002) used a 'valley detector' (a type of moving window operation) on facial images to reduce the greyscale images down to a set of pixels representing locations that have high local curvature (calculated from pixel values) (Figure 3.13). A similar approach is used for palmprint matching, in which the interesting features of the palm (e.g. intersections of fold lines) are identified using image processing operations and subsequently compared (You et al. 2002). This refined set of pixels (or salient features) is computed using some knowledge of the image content and context, which is rarely so clearly defined in hydrological fields.

For the comparison measure to be calculated, the salient features from the query image must be compared against the images (and their features) stored in the database. This is done using a similarity measure, such as the Hausdorff distance (Huttenlocher et al. 1993; Veltkamp and Hagedoorn 1999) that is commonly used in pattern matching

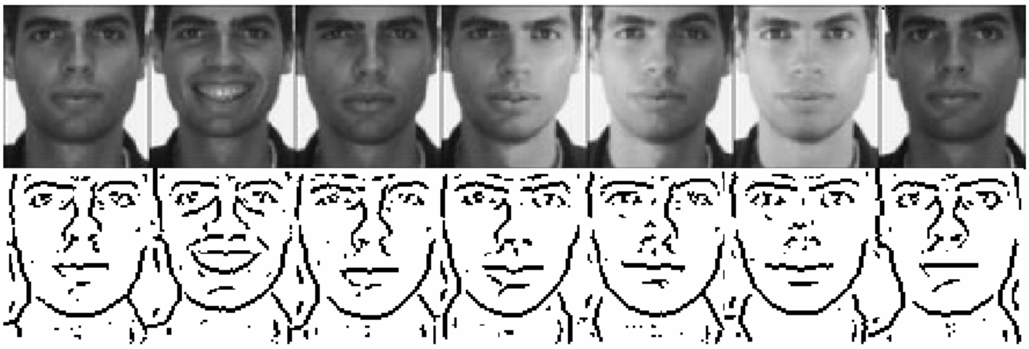


Figure 3.13 A set of seven facial images (top row) and the associated edge and 'valley' features (second row) (from Pujol et al. 2002).

research. The Hausdorff distance is a global measure of distance that describes how closely the two sets of salient features agree. If the features are closely matched, the distance will be small. There are numerous metrics for comparing two sets of features extracted from images, designed to maximise robustness to noise, shifts, scale changes and rotation (Veltkamp 2001). These comparison techniques are specifically for structural features defined in the images, so they could only be applied to hydrological spatial fields with simple, identifiable structural features (e.g. comparing observed and modelled river networks).

With images where the context of the image is not known, such as within a general image database, meaningful structural features are more difficult to define. When knowledge-driven structures cannot be defined, data-driven methods must be used to simplify an image into homogeneous regions (i.e. image objects). The regions are labelled using their characteristics (e.g. value, variance, shape measures, etc) and these can be compared individually or summarised. Because of the data-driven nature of segmentation methods, a region in one image often corresponds with several regions in another image. This causes difficulty when trying to do pair-wise comparisons between regions (i.e. it is impossible to define how the regions from different fields relate, unless labelled equivalently). Chen and Wang (2002) describe a method to determine image-level similarity measures by combining the various region-level similarities. To achieve this in an efficient way, they used fuzzy functions to represent how the region characteristics vary within an image. Comparison was then undertaken by using the fuzzy functions to represent the content of the images (i.e. a global comparison of characteristics). This made it possible to rapidly search a large database (>60,000 images) using a query image. A method that required comparison of each region would have been too inefficient for this type of application. A similar approach could be applied to hydrological spatial fields when processing speed was essential, but this is not considered to be widely applicable.

3.4.3 Comparisons for change detection

Change detection using spatial fields of the natural environment is commonly undertaken in land-use and land-cover studies, using spatial fields from remote sensing. Change detection is also undertaken using other types of images for surveillance and medical applications (Radke et al. 2005), where the changes between subsequent images need to be automatically interpreted (e.g. to trigger an alert). While there is great diversity in the applications for change detection, most studies employ a number of common processing steps and algorithms. The overall aim of change detection is to produce a 'change mask' (Radke et al. 2005), which is synonymous with an intermediate field as described in this thesis. This is produced from two images that represent the same attribute at different times or from different sources. The intermediate field should identify the elements that have changed. However, the definition of what constitutes a difference is where the comparison methods vary.

In many applications such as surveillance, images of stationary objects will change due to lighting, camera motion and even atmospheric influences. These may be regarded as insignificant changes. In the case of change detection for fields in the natural environment, insignificant changes are commonly those due to registration or processing nuances. Ideally, the comparison method should detect the significant changes, while ignoring the effects of insignificant changes, such as those caused by registration, scale differences or illumination (e.g. Figure 3.14). Radke et al. (2005) provide a comprehensive review of change detection algorithms, which describes many methods to produce a change mask that only represents the significant changes.

The most basic method presented in their review – simple differencing – is also the most widely used comparison method for change detection in land-use change and hydrology. Beyond this method, Radke et al. (2005) identify a number of pre-processing steps, including geometric and radiometric adjustment, which can be used to remove undesirable artefacts in the images (i.e. conditioning the images). During local comparison of each pixel, there are additional analyses that can be applied to evaluate change. For example, statistical significance testing can be applied between the pixel values to determine if the change is significant. Another common approach is to use a simple spatial model to evaluate the comparison for each pixel (e.g. the 'analysis window' methods described in section 3.3.5).

3.4.3.1 Applying simple spatial models during local comparison

The idea of using simple spatial models to evaluate change has been recently applied for comparing categorical fields (Hagen 2003; Güntner et al. 2004; Hagen-Zanker 2006). Land-cover change studies usually conduct an element-by-element comparison (i.e. simple differencing) and produce a contingency table, from which the assessment of accuracy is made. However, Foody (2002) and Power et al. (2001) recognised that this is prone to identifying elements effected by spatial registration errors as having land-use change, when in reality no significant change has occurred. To adjust for this situation, Hagen (2003) describes a fuzzy comparison method, which allows the location and

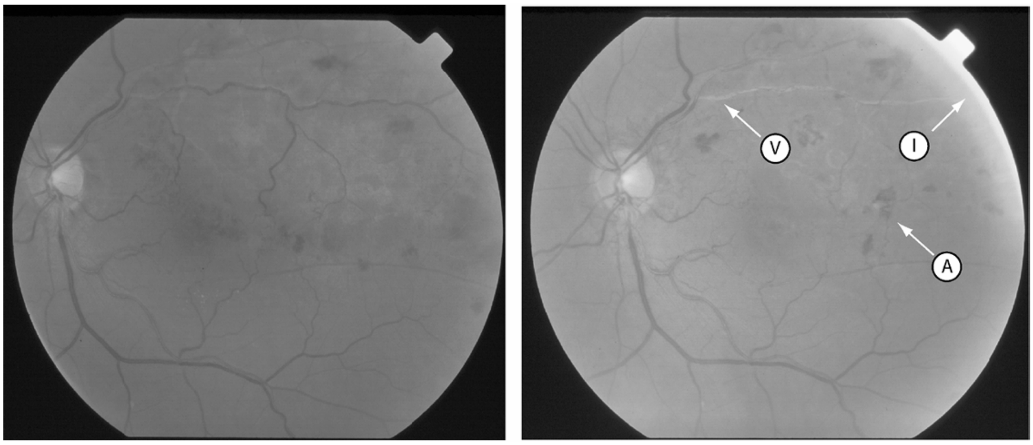


Figure 3.14 A pair of images of a human retina taken six months apart, with changes that are important to the application marked V and A. Unimportant changes, such as those due to illumination differences, are marked I (from Radke et al. 2005).

category of an element to be defined using fuzzy sets (Cheng et al. 2001). Fuzzy sets allow an element to be defined in an uncertain (or vague) manner, based on the subjective judgement of the user (or knowledge of the fields being compared). These fuzzy definitions are then applied during each local comparison (by using a simple spatial model), resulting in an intermediate measure and overall comparison measure that is ‘less strict’ (Figure 3.15).

The simple spatial model used by Hagen (2003), which calculates the similarity for each element, works with the focus element in one field and all elements within the analysis window (i.e. those defined as similar in location) from the other field. Comparisons are done for each pair of elements, with the optimum similarity being returned. The full details of this method, which is only applicable for categorical fields, are given in Hagen (2003). However, the fundamental idea of allowing elements to be uncertainly defined has been pursued in this thesis. Details of how this has been adapted and implemented for comparing continuous hydrological fields are provided in Chapter 4.

3.4.3.2 Identifying and understanding the cause of change

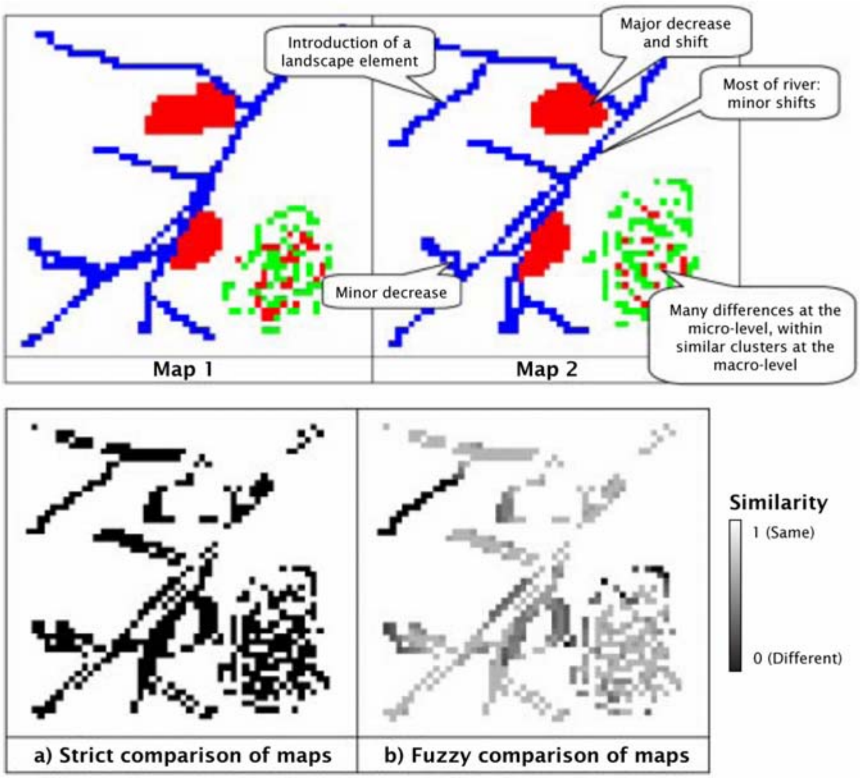

Introduction of a landscape element
Major decrease and shift
Most of river: minor shifts
Minor decrease
Many differences at the micro-level, within similar clusters at the macro-level
Map 1
Map 2
Similarity
1 (Same)
0 (Different)
a) Strict comparison of maps
b) Fuzzy comparison of maps

Figure 3.15 Two categorical land-use maps are compared using a) strict and b) fuzzy comparison methods. The strict comparison measures each pixel as either same or different, whereas fuzzy comparison is able to measure the degree of similarity (adapted from Hagen 2003).

In many applications, including hydrology, estimating the change (i.e. local errors or similarities) is the first step toward "the more ambitious goal of change understanding" (Radke et al. 2005). To understand the meaning of the changes, further analysis of the change mask (or intermediate field) is needed. In general, this involves segmenting the change field and then extracting the semantics of individual image objects, using some tools designed for specific applications. For example, in surveillance, a change mask may be analysed to interpret whether the change is due to an intruder. Explaining the meaning of features within an image is one of the great challenges for all disciplines dealing with pattern recognition and comparison. In controlled situations (e.g. industrial manufacturing), it can be possible to interpret features within images. However, when the situation is less controlled (or less controllable), there remains a dependence on human visual inspection to undertake this role. Hydrology is one of these 'less controlled' situations, which partly explains the difficulty faced when trying to explain why a certain difference exists between observed and modelled fields.

3.4.4 Comparisons utilising multiple scales

The majority of comparisons undertaken in other disciplines are performed at a single scale. The previous sections have shown that this may be at a global scale (e.g. comparing region-based characteristics), a defined intermediate scale (e.g. comparing structural features) or, more commonly, at a local scale (e.g. comparing every element). This is in contrast to the innate visual methods used by humans, in which multiscale analysis and comparison is routinely performed. In the past five years, there have been many examples from a range of disciplines (including hydrology) where the analysis of images (or spatial fields) is performed at multiple scales simultaneously. These include the multiscale segmentation and classification of images (Baatz and Schäpe 2000; Burnett and Blaschke 2003; Hay et al. 2003) and the derivation of terrain attributes using multiscale terrain models (Gallant and Dowling 2003). Despite multiscale approaches being widely used for analysis, there are few examples of multiscale comparison being used.

Within the discipline of ecological modelling, a method for comparison at multiple scales was first developed by Costanza (1989). The reasons for developing the methods were two-fold: 1) that there is no 'proper' scale at which to compare models with reality, but rather a range of scales; and 2) comparison at a single scale gives no consideration to 'near misses'. The fuzzy comparison method (Hagen 2003) reviewed in the previous section provides an explicit way of considering near misses. However, the lack of a 'proper' scale for comparison remains a valid point that is usually overlooked in favour of using the finest scale. The method developed by Costanza (1989) works by calculating a measure of fit for a range of 'upscaled' spatial fields. This measure is then plotted against the scale or summarised into a multiscale measure. The 'upscaling' is managed at each scale by increasing the size of the analysis window around each element during local comparison. Costanza (1989) calculated the fit between each 'upscaled window' by comparing the distributions of element values within the windows. For continuous fields, Costanza (1989) suggests using the RSQ measure between the 'upscaled windows', although no analysis was done using this method with continuous fields. RSQ cannot be calculated from only distribution information (i.e. it requires pair-wise comparison of elements), so it would be expected to work differently to the examples given by Costanza (1989).

Since first being presented in 1989, the multiscale method of Costanza (1989) has had limited recognition and use. Only in the past five years has it been revisited in the ecological modelling and remote sensing literature (e.g. Pontius 2002; Pontius et al. 2004a; Hargrove 2006). These recent uses of multiscale comparison have worked solely with categorical fields. Most of this work has been by Pontius et al. (2004a) and it focuses on the task of land-use change. Multiscale comparison is applied to categorical fields to determine the cause of category changes (e.g. error due to quantity, error due to location), which is useful for assessing model performance. The methods used by Pontius et al. (2004a) are not detailed here because they cannot be applied to continuous spatial fields. However, the multiscale method of Costanza (1989) appears suited for use in hydrological applications and is further developed in this thesis (using a more suitable measure of fit).

3.4.5 Potentially useful methods from other disciplines

Four major aspects of comparison that are encountered in other related disciplines have been reviewed in this section. Image processing is used for preparing images (or fields) for use in comparison. This involves tasks such as noise filtering, applying geometric corrections and deriving structural features (e.g. regions, salient points). Structural features are used for reducing the complexity of a field (i.e. for efficiency during comparison) and for providing an alternative representation that can be more readily interpreted. When comparing structural features, a range of different comparison methods tend to be combined, although their use is application specific. The most common application of comparison in other disciplines is change detection, which focuses on producing a change mask that only represents the differences between fields that are considered significant. All of these comparison methods are applied at a single scale, despite the widespread recognition that multiscale analysis is useful. Limited work using multiple scale comparisons has revealed they can provide useful information about comparison. The methods identified in this section for pre-processing and comparison are summarised in Table 3.8. There are a number of methods shaded in Table 3.8 that are considered suitable for using or adapting for hydrological purposes.

These methods can address some of the limitations with the current methods used and also emulate key aspects of visual comparison. A proposed methodology for using these methods to advance spatial field comparisons is given in the following section.

Table 3.8 Summary of methods used within other disciplines for comparison and pre-processing of spatial fields (or images). The potential uses of these methods in spatial field comparisons are described. The methods shaded are pursued further in this thesis.

Method	**Description of current use and potential use for hydrology**
Basic conditioning tasks	Used to modify element values within a field, to remove undesirable artefacts (e.g. noise) or to make the field ready for comparison (e.g. scale differences); currently used in hydrology and many other disciplines; conditioning improves the change mask produced via simple differencing
Edge-detection methods	Used for defining the edge of features in a field; unsuitable for continuous hydrological fields due to absence of sharp boundaries; potentially useful for binary categorical hydrological fields
Knowledge-driven segmentation	Used to extract predetermined features from a field; potentially used in hydrology to define features from surrogate fields (e.g. segmenting a slope map); method for integrating surrogate features into a comparison method is unclear
Data-driven segmentation (using values)	Used for identifying organisation within a data set or field, based on measurement space information (i.e. does not enforce spatial contiguity); limited value for hydrological fields due to poor spatial segmentation
Data-driven segmentation (using values and locations)	Used for breaking a field into a set of contiguous regions; no simple or existing method for labelling segmented regions; may be useful for recognising optimally homogeneous regions within hydrological fields; assigning labels (i.e. hydrological meaning) remains challenging
Identifying salient, structural features	Used to simplify a field into a limited set of features that are sufficient to describe its appearance; used in controlled cases (e.g. face recognition); generally unsuitable for hydrological fields which lack such features
Comparing region-based global characteristics	Used to summarise the characteristics of segmented regions within an image; avoids problem of ‘pair-wise’ comparison of regions with different labels; unsuitable as it removes the explicit locations of regions and their associated hydrological meanings
Simple differencing	Used for strict local comparison; intermediate field produced may identify differences that are not significant; useful as a strict comparison test, but not able to handle other situations
Applying simple spatial models for local comparison (e.g. fuzzy comparison)	Used for producing an intermediate field based on more than simple differencing; simple spatial models can make comparisons between ‘windows of elements’; fuzzy comparison method has been developed for land-use change; useful to adapt general ideas for continuous fields
Multiscale comparison	Used to describe fit at a range of scales, which can be visually analysed by plotting fit versus scale, or numerically analysed by summarising multiple scales; some methods used for continuous fields could be applied for hydrological purposes

3.5 Proposed methodology

The objective of this thesis is to advance the methods used for quantitatively comparing spatial fields in hydrological modelling. Spatial field comparisons used in hydrology are currently quite simple, so advances can be made by introducing complexity into existing methods that can facilitate additional options during comparison. From the review, three interwoven themes have been identified as being the most important when undertaking comparison: 1) importance; 2) tolerance; and 3) completeness. The following sections describe each of these themes in more detail.

The promising methods being used in hydrology and other disciplines will be adapted to include these themes into some new comparison methods for hydrology. These new methods will need to include user-defined parameters, which will enable the methods to be adaptable for different applications.

3.5.1 Importance

Defining ‘importance’ relates to controlling what is actually being compared. When comparison is undertaken only for important elements, a comparison measure that has a specific interpretation is produced. During visual comparison, importance can be controlled by the user focusing on certain parts of the spatial fields more than others and/or making comparisons between elements recognised as being important. Current quantitative comparison methods rarely use any such notion of importance, which results in comparison measures that only have general (and ambiguous) interpretations. In the other disciplines, there are a range of image processing methods that are used to define the important elements (or features) that are then compared. This has shown that by simply defining importance, many general methods can be used for quite specific comparison tasks.

In hydrology, importance is expected to be less-clearly defined than in other disciplines (e.g. facial recognition). Pre-processing methods from image processing can be readily applied to change the structure or extent of a continuous field prior to comparison, but this is dependent on the definition of importance. Importance can be based on characteristics of the field (e.g. elements in highly variable areas are most important), hydrological characteristics (e.g. hillslope elements are most important) or expert knowledge (e.g. that a certain part of the catchment is most important). Importance can also be based at different spatial scales, such as the finest scale (i.e. individual elements) or an intermediate scale (e.g. homogeneous regions segmented in the data). When combined with a comparison method (e.g. local comparison), different definitions of importance can be used to produce a range of meaningful comparison measures.

3.5.2 Tolerance

'Tolerance' is included into many of the comparisons undertaken discussed in this review. It is implicitly used during visual comparisons to ignore minor differences between spatial fields (e.g. shifts, different colours). During quantitative change detection, the change mask is manipulated (i.e. via pre- or post-processing) so that insignificant differences are tolerated. For land-use change, the user can define fuzzy similarity between locations and categories, which allows for the tolerance of mismatches. In all cases, using tolerances allows the uncertainty with which an element is defined to be explicitly considered during local comparison. Elements can potentially have uncertainty defined for value, location and time.

Uncertainty exists throughout all aspects of hydrological modelling, yet it is not accounted for in most common practices (Pappenberger and Beven 2006). Introducing tolerance (of uncertainty) during hydrological comparison can advance the comparison methods used and also promote greater consideration of uncertainty. The fuzzy similarity method (Hagen 2003) used in land-use change appears suited for adaptation to hydrological fields. While it has only previously been applied for categorical fields, the general approach can provide the basis for incorporating tolerance into local comparison for continuous fields. These tolerances also allow some consideration of the spatial (and temporal) relationships between elements, which is currently lacking in all local comparison methods.

3.5.3 Completeness

'Completeness' is a general guideline to follow during comparison and it relates to making comparisons of multiple aspects and across a range of scales with different types of comparison measures. Visual comparisons can innately achieve this, while the current quantitative methods each address a specific scale and aspect of comparison. By combining the notions of importance and tolerance into the current comparison methods (as proposed in the previous sections), many more different aspects and scales can be specifically compared. In addition to these methods, the multiscale comparison method developed in ecological modelling (Costanza 1989) can be adapted for use with continuous fields to make comparisons at all scales. When multiple aspects and scales are compared, computational demands are expected to be high. Therefore, experience with selecting the most useful methods for particular applications will need to develop. By including the user in the comparison process, it is expected that this experience will develop and make it possible to use the existing and new comparison methods for a range of hydrological situations.

The advances to spatial field comparisons expected from applying this proposed methodology are: 1) a range of comparison methods that are simple to apply, but versatile for different aspects and scales of comparison; 2) quantitative comparison methods that can be repeated and easily reported/interpreted; and 3) greater thinking by the user/hydrologist about what is compared and why.

3.6 Chapter summary

This review has investigated three topics: 1) the general processes undertaken during visual comparison; 2) the current limitations with quantitative methods used in hydrology; and 3) the promising methods being used in other disciplines for comparison. Amongst these topics, a large range of methods for potentially improving current comparison methods have been identified. These methods are useful for the hydrological community to be aware of, although many of them are only suited to specific applications or data types. The summaries given for each topic identify the most promising methods for hydrological comparisons. From this information, a proposed methodology for advancing the methods used for spatial field comparisons in hydrology has been devised and it is implemented in Chapter 4.

Chapter 4

Development of comparison methods

4.1 Chapter overview

The review of spatial field comparisons in 0 identified three major themes that are encountered in other disciplines, but are currently lacking in hydrological comparison methods. The proposed methodology describes how these interwoven themes – importance, tolerance and completeness – can potentially be implemented into comparison methods that are suitable for use in hydrology. This chapter develops these 'new' comparison methods, which encompass ideas that are new to hydrological comparisons. The chapter also describes how the new methods work with the current methods used in hydrology to develop a general strategy for comparison.

Importance is enforced in a spatial field comparison by ensuring that the fields only represent the important information. This can be achieved by using existing pre-processing methods for modifying the structure and values of continuous spatial fields. 'Hydrological importance' is expected to be defined by using aspects of the observed, modelled and/or surrogate spatial fields (e.g. homogeneous regions, terrain attributes). A simplified version of the data-driven, region-growing segmentation method (Baatz and Schäpe 2000; Definiens Imaging 2003) has been adapted here for recognising the homogeneous regions in continuous fields. A new 'region-based comparison method' for using regions of importance during comparison is developed.

A new 'tolerant local comparison method' has been developed here by adapting the fuzzy comparison method described in the land-use change literature (Hagen 2003). The new method requires an alternate definition and implementation of tolerances to be devised so that it can work with continuous spatial fields. This method produces an intermediate field of tolerant error (or tolerant similarity) from which a range of new comparison measures have been developed.

Completeness can be achieved by using different definitions for importance, tolerance and scale with different comparison measures (i.e. error and similarity) to compare multiple aspects of spatial fields. To compare all spatial scales, a 'multiscale comparison method' (Costanza 1989) has been adapted for use with continuous hydrological fields. It is not possible to compare all aspects of importance and/or tolerance, so a strategy for achieving completeness is needed. A comparison flowchart that identifies all of the possible comparison measures available when the current and new methods are combined is developed. By combining these methods, user knowledge and a set of comparison objectives, a general strategy for making complete comparisons is devised.

4.2 Preliminary information

All of the new methods aim to be simple and familiar, while still providing additional options for comparison. These factors should help to make them readily adopted by hydrologists for use in a range of situations. In all of the comparison methods, the information within two spatial fields is reduced into a single numerical measure, so there is an expected information loss. However, many of the methods also provide intermediate measures that retain the spatially-explicit information and can be visually inspected to explain more about the numerical measure.

The following sections describe some preliminary details that are used throughout this chapter: 1) the approach used for describing the algorithms and methods; 2) the synthetic spatial fields used for illustrating the new comparison methods; 3) the benchmark performance of the synthetic fields when using current comparison methods; and 4) the software tool available to the reader for experimenting with these methods.

4.2.1 Approach used for describing algorithms

Throughout this chapter, the comparison methods are detailed by describing the algorithm used to conduct the comparison. The algorithms presented in this chapter are all described using ‘pseudocode’. Pseudocode is widely used in computer programming to describe an algorithm that can be implemented on many different computing platforms. It provides a simple way of describing the processes followed for any comparison method.

When using the algorithms, the spatial fields are referred to as the source field(s) and the target field(s). In model assessment tasks, the source is usually the modelled field and it is compared with the target (i.e. the field that represents the reality or benchmark), which is usually the observed field. In other tasks, the target field may be a reference or ‘base case’. The definition of which field is the source and which is the target is critical for comparison, as all comparisons are directional (i.e. the source field is compared against the target field to describe how it differs).

Swapping the source and target field definitions can have different implications. In simple comparisons (e.g. BIAS), it will only cause the sign of the summary measure to switch (e.g. from over- to under-prediction). In more complex comparison methods, the direction can influence the overall magnitude and interpretation of the results (e.g. a tolerant comparison applies tolerances only to the target). While all efforts have been made to be unambiguous with terminology used in this thesis, in ambiguous situations the modelled and observed fields should be interpreted to be the source and target fields respectively.

4.2.2 Synthetic fields

A set of synthetic spatial fields are used throughout this chapter to illustrate the performance of the new comparison methods. A single ‘observed’ field (i.e. the target) has been used to produce nine ‘modelled’ fields, each with a different known deformation applied to it. The origin of the observed field is irrelevant for its use in this chapter, although the spatial arrangement has been caused by factors common to hydrological fields – terrain, soil type, geology and vegetation

Each of these synthetic fields has 59 x 24 elements and the elements are assumed to be related to all immediate neighbours (i.e. using eight-way connectivity). The fields are shown using a ‘dry to wet’ colour ramp in Figure 4.1 to visually reveal the spatial variability. Using a single colour ramp (e.g. black to white) can make visual discrimination of variability more difficult.

The details of the deformations applied to create each of the modelled fields are listed in Table 4.1. This table also details a hydrological modelling situation that each type of deformation represents. These synthetic fields are used throughout this chapter to test whether the new comparison methods perform as expected when applied to fields with known differences. Chapter 5 applies these methods in a real hydrological modelling situation where model performance is being evaluated.

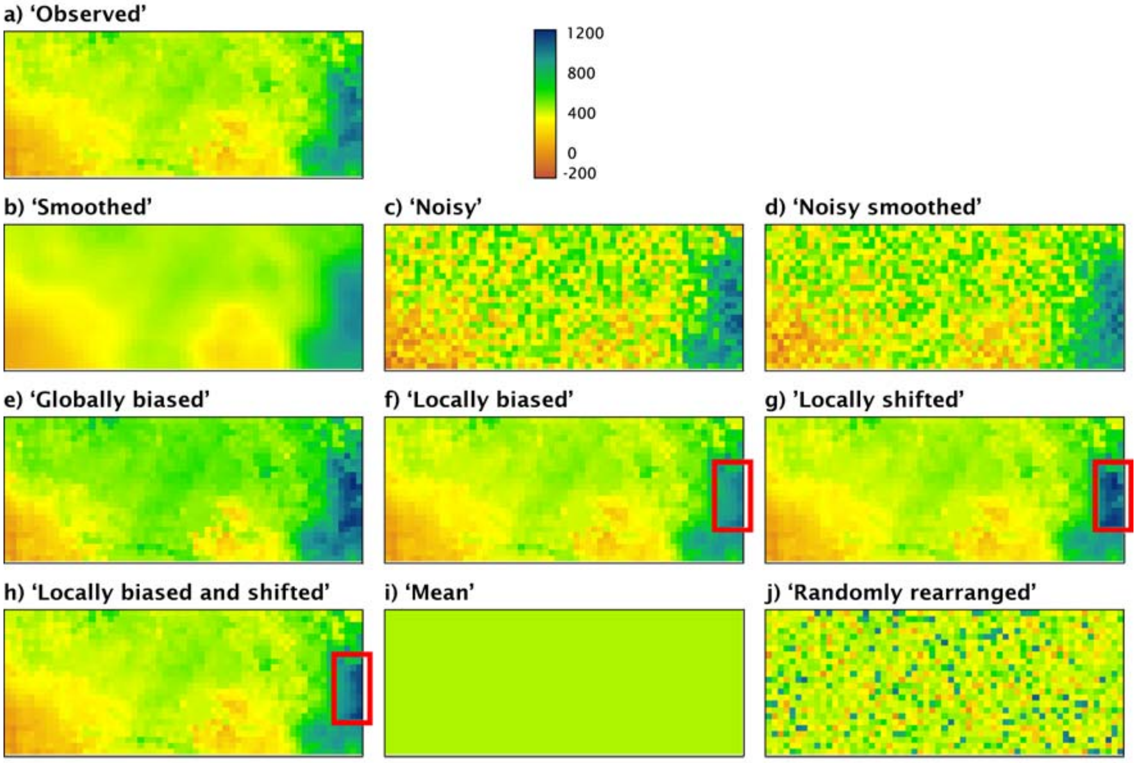


Figure 4.1 Nine synthetic spatial fields have been produced for illustrating the comparison methods. The locations altered by 'locally applied' deformations are delineated by the rectangle. All of the deformations applied to produce fields b) to j) are detailed in Table 4.1.

Table 4.1 Details of the deformations applied to the 'observed' field in Figure 4.1 to create each 'modelled' field. A hydrological situation where each type of deformation could occur is also identified.

Field	Deformations applied and hydrological situation represented
a) Observed	Treated as the observed reality throughout this chapter; representative of a typical observed spatial field used in hydrological modelling
b) Smoothed	Observed field has been smoothed using the ordinary kriging interpolation method (using the nugget parameter to control the level of smoothing applied); representative of hydrological model where small-scale variability is not represented; also representative of observed field with larger spatial support (i.e. less variability)
c) Noisy	Random noise (σ = 100) has been added to the observed field; representative of a hydrological observation that contains 10% measurement error
d) Noisy smoothed	Random noise (σ = 100) has been added to the smoothed field; representative of a model where small-scale variability is poorly represented by a non-spatial stochastic parameter
e) Globally biased	All element values in the observed field were increased by 20%; representative of a model or observation that has a systematic bias in all elements
f) Locally biased	A small region of elements in the observed field were increased by 20%; representative of a model or observation with a localised systematic error (possibly resulting from incorrectly representing a process in the region)
g) Locally shifted	A small region of elements in the observed field were shifted two elements to the right; the 'gap' left by these elements was filled by using the ordinary kriging interpolation method; representative of a model where the magnitude of values is generally correct, but spatial shifts have occurred due to model processes or data processing
h) Locally biased and shifted	A small region of elements in the observed field were shifted by two elements and increased in value by 20%, as undertaken in f) and g); representative of a model where a range of factors are producing a localised error in the model
i) Mean	The mean value of the observed field is assigned to every element, thus producing a field with no variability; representative of a model that is non-spatial (i.e. lumped) but correctly models the mean value
j) Randomly rearranged	The element values in the observed field are randomly rearranged; only one realisation is used to define the synthetic field; representative of a model or observation where the spatial arrangement is completely wrong, yet the basic statistics are correct

4.2.3 Using current comparisons methods

The current comparison methods used in hydrology are applied here to the synthetic fields in Figure 4.1. Strict quantitative methods, as well as methods requiring visual analysis are presented. These measures show how the current methods respond to the deformations introduced into the synthetic fields. As the new methods are described throughout this chapter, the new quantitative information that they add to this existing information can be identified.

4.2.3.1 Quantitative measures

Basic statistical characteristics (MEAN, SDEV, SKEW, KURT), a global comparison measure (BIAS) and local comparison measures (RMSE, MAE, RSQ, COE) are shown in Table 4.2. The basic statistics reveal that the mean is stable amongst all the 'models', apart from the globally biased model. The standard deviation is less in the smoothed model, but much higher in the noisy models. Global BIAS reveals the errors in the globally and locally biased models, although the local bias is reduced substantially by averaging. The local comparison measures (RMSE, MAE) also only respond slightly to the models with local biases and shifts. The RSQ measure identifies that most of the field is highly correlated in these instances, which correctly suggests that the error is localised. When global deformations exist in the models (e.g. bias, noise), the local error measures increase. With the correlation measures, RSQ ignores the errors in the globally biased model, but COE penalises them. The rearranged field is identical for all global comparisons, but it is the worst performing model using local comparison.

Table 4.2 The statistical characteristics (MEAN, SDEV, SKEW, KURT), a global comparison measure (BIAS) and local comparison measures (RMSE, MAE, RSQ, COE) for the synthetic fields (Figure 4.1).

Field	MEAN	SDEV	SKEW	KURT	BIAS	RMSE	MAE	RSQ	COE
a) Observed	335.4	143.8	1.7	4.4					
b) Smoothed	336.0	135.0	1.6	3.5	0.5	38.4	24.7	0.93	0.93
c) Noisy	335.4	176.4	0.9	1.7	0.0	100.0	86.1	0.68	0.52
d) Noisy smoothed	336.0	168.0	0.8	1.4	0.5	108.6	92.3	0.59	0.43
e) Globally biased	402.5	172.6	1.7	4.4	67.1	73.0	67.1	1.00	0.74
f) Locally biased	339.7	160.6	2.3	7.6	4.3	26.9	4.3	0.98	0.96
g) Locally shifted	333.6	137.5	1.5	3.5	-1.9	20.9	2.9	0.98	0.98
h) Locally biased and shifted	336.5	148.6	1.9	5.8	1.1	23.2	2.6	0.98	0.97
i) Mean	335.4	0.0	1.0	-2.0	0.0	143.8	89.1	0.00	0.00
j) Randomly rearranged	335.4	143.8	1.7	4.4	0.0	199.6	136.3	0.00	-0.93

4.2.3.2 Methods requiring visual analysis

Comparison methods requiring visual analysis are emulated by the new methods introduced in this thesis. However, to identify what information should be recognised by these methods, visual analysis of the intermediate fields from the local comparison measures in Table 4.2 is undertaken (using Figure 4.2). It is immediately apparent that the locally deformed models are identical for most of the extent, apart from the small region of deformation. The exact differences are much more difficult to perceive when visually comparing the fields (e.g. using Figure 4.1). There is a lack of spatial organisation in the residuals for the smoothed and noisy models, which suggests the errors are randomly distributed. In the globally biased, mean and rearranged models there remains structure in the residuals, suggesting systematic problems in the model.

Variograms for each of the synthetic fields are shown in Figure 4.3 and can be visually analysed and compared. They show that most of the synthetic fields have a similar spatial structure, apart from the mean and rearranged models. The rearranged field has no obvious range and appears to be comprised entirely of nugget effect, while the mean field has no variance at any lag. The other variograms reveal some degree of correlation at a lag of approximately 20 elements (i.e. half the width of the field). The magnitude

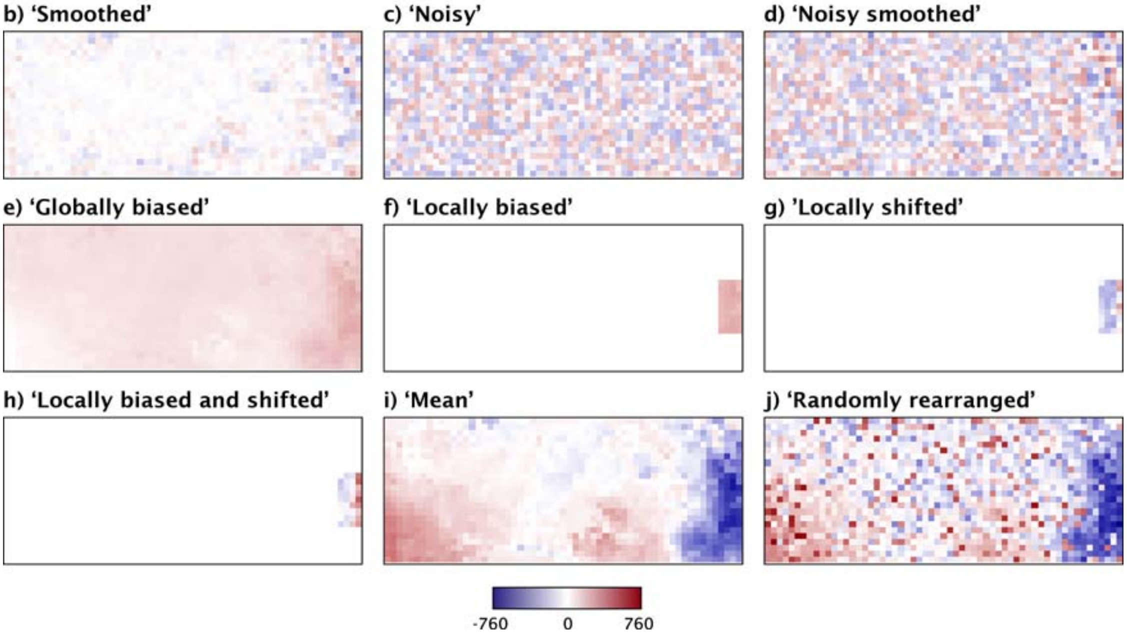


Figure 4.2 Intermediate error (or residual) fields produced during local comparison of the synthetic fields from Figure 4.1.

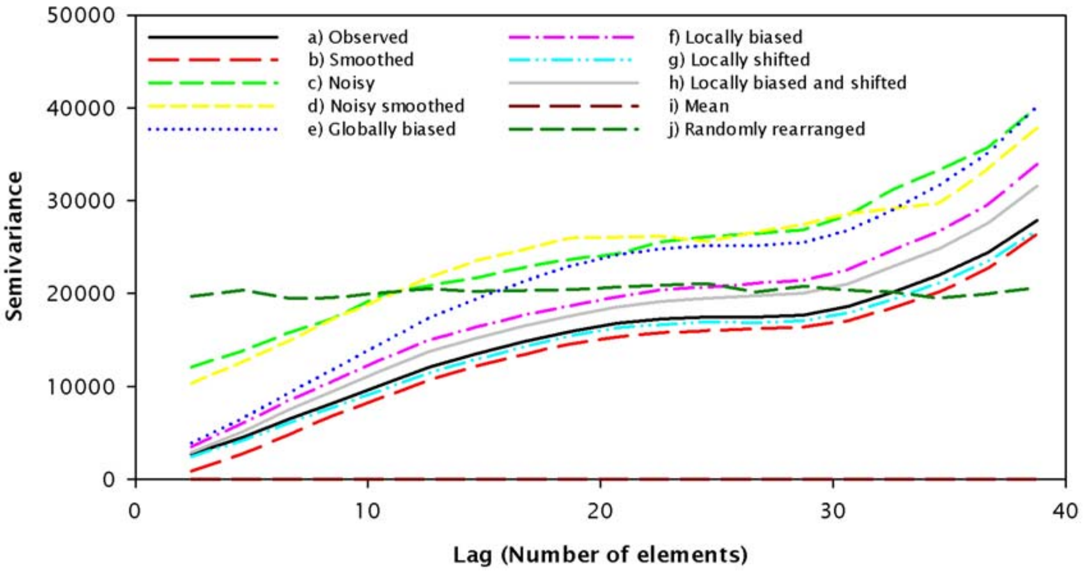


Figure 4.3 The experimental variograms for the synthetic fields in Figure 4.1 are calculated and plotted. These show the general characteristics of spatial structure (e.g. smoothly varying) for each of the synthetic fields and can be visually compared.

of semivariance varies between the models, although the similar shape in the variograms suggests similar spatial relationships exist within most of the fields. The variograms of the noisy models have a much larger nugget, while the local changes cause only minor differences to the variogram shape. The numerical differences between the semivariances at each lag could potentially be summarised (e.g. average separation between variograms), although this would only measure the absolute difference between variograms. It is preferable to identify the key characteristics (e.g. nugget, range) and provide direct information about how the variograms differ

4.2.4 Software tool for making comparisons

During development of the comparison methods presented in this chapter, a prototype software tool has been created for running all of the comparisons and producing the necessary outputs. The tool provides the user with a simple interface for defining the source and target fields, as well as a definition of analysis regions. The current and new comparison methods can each be used to produce numerical and graphical outputs.

The interested reader is referred to the more complete description of this software given in Chapter 6 and the detailed tutorial provided in Appendix A. The spatial field data used in this thesis is provided for use in this software.

4.3 Importance

Importance is defined in a spatial field comparison by using pre-processing methods to modify the elements and values of a spatial field. The pre-processing allows the user to control what aspects of the spatial fields will be compared, which can help to produce comparison measures that are not confounded by any unimportant or erroneous information (i.e. only the information that is important). When applied to hydrology, this approach allows the notion of 'hydrological importance' to be defined prior to comparison so that specific hydrological comparison measures can be produced.

This section does not attempt to fully develop a definition of hydrological importance for use in comparison. Instead, it focuses on analytical tools required to implement the concept of importance, leaving the specific definition of important to the user. Hydrological importance is assumed to relate to the elements and regions found in observed, modelled and/or surrogate hydrological fields. Therefore, this section presents: 1) the existing pre-processing methods used for making spatial fields comparable; 2) a new data-driven segmentation method for recognising homogeneous regions within continuous fields; 3) the existing, knowledge-based methods for defining regions of hydrological importance; and 4) a new region-based comparison method.

4.3.1 Making spatial fields comparable

Prior to undertaking comparison, there are pre-processing methods that can be used to make the fields more comparable (i.e. to ensure that only the important information remains in the fields). The primary focus when applying pre-processing is to remove (or reduce the impact of) any known differences introduced when producing the fields. In hydrological applications, the most common artefacts present within observed spatial fields are scale inconsistency and measurement noise. Modelled fields can have different field structures and therefore also exhibit scale inconsistency.

Noise filtering, smoothing and scaling are inter-related pre-processing methods that have been discussed in Chapter 2 and 3. They are all used to make spatial fields more comparable, by reducing the differences between the characteristic scales of the fields. These scales should be equal during comparison to ensure that no 'false differences' confound the results, although this is not always possible due to the limited methods for obtaining observed spatial fields. A number of options for upscaling spatial fields, including moving ‘analysis window’ methods and geostatistical interpolation, are currently used for reducing these differences. When random noise exists in an observed field (e.g. random measurement error), the impact of this noise can be reduced by smoothing the values. These methods tend to distribute the noise amongst neighbouring elements and increase the spatial support of element values (i.e. upscaling), which leads to the finest scale being lost.

Other pre-processing methods address factors such as spatial alignment and comparing equivalent hydrological attributes. Correct georeferencing is essential to ensure that modelled field element(s) are compared with the appropriate observed element(s) (i.e. those that are spatially coincident). This is generally managed in a geographic information system (GIS) or similar software. If the attributes being compared differ, the measurement spaces must be pre-processed so that the distributions are forced to relate (for the purposes of the comparison). Standardisation and histogram matching are both suitable methods for achieving this. This is not an exhaustive list of the pre-processing issues encountered when making spatial fields comparable, but these are the most common issues faced during pre-processing in hydrological modelling.

4.3.2 Data-driven recognition of regions

The recognition of regions within an image or spatial field is innate to humans and is used extensively during visual analysis. Regions are also a used for identifying the important structures within a spatial field, which are used to simplify it for analysis and comparison purposes. There have been many attempts to emulate this ability by using different segmentation algorithms. The most promising method for use with continuous fields representing the natural environment is data-driven region-merging segmentation. This type of method treats every element as a separate region to begin and then uses decision heuristics to iteratively merge regions together until an optimised segmentation result is obtained.

Woodcock and Harward (1992) presented an early approach at this type of segmentation, although the more recent implementation of Baatz and Schäpe (2000) is considered more suitable. Baatz and Schäpe (2000) made two important improvements:

1) a 'distributed treatment order' is used to allow regions to grow evenly; and 2) 'local mutual best merging' is used to ensure that the best merge within a local area occurs. These improvements helped to produce more visually intuitive results and resulted in a patented implementation of this segmentation method (Baatz et al. 2004). This segmentation method is used exclusively in the software named eCognition (Definiens Imaging 2003), which is increasingly used in research and industry for object-based image classifications for remote sensing.

The following sections present: 1) the algorithm used for data-driven region-merging segmentation in this thesis; and 2) examples of using the segmentation method with the synthetic fields from Figure 4.1.

4.3.2.1 The region-merging segmentation algorithm

The key components of the region-merging segmentation algorithm used in this thesis are: the general iterative process used to determine the segmentation; the decision heuristics used for determining the 'best merge'; and the order in which regions are treated. The iterative process used here (Algorithm 4.1) is a direct implementation of the method described in the eCognition User Guide (Definiens Imaging 2003). The decision heuristics used here (Algorithm 4.2) are a simplified version of the 'local mutual best merging' rules in Baatz and Schäpe (2000). These are simplified because hydrological fields only have a single band of information (unlike much remotely sensed data) and the segmentation must therefore operate on less information. The treatment order used here (Section 4.3.2.2) has been specifically developed for this implementation, although it is motivated by some 'vague' descriptions in Baatz and Schäpe (2000) and Baatz et al. (2004).

Throughout Algorithm 4.1, three attributes are maintained for each region in the spatial field – heterogeneity (i.e. standard deviation), area and perimeter. The area and perimeter values are combined to characterise the shape of each region. The heterogeneity and shape attributes are combined (using the formula in Algorithm 4.2 and the user-defined 'region shape' parameter) and used to determine the 'cost' of merging any two regions together and the 'local mutual merge' that is best. Merging progresses by using these attributes and the treatment order until no more merges can be

```
1  Define global merging parameter in range $[0,\infty)$
2  Define region shape parameter in range [0,1]
3  Initialise master list of regions using input field
4    Create region for each element
5    Initialise values for heterogeneity, perimeter and area
6    Initialise list of neighbouring elements
7  Do iteration of merging
8    Create working list of regions (using distributed treatment order)
9    While list of regions is not empty
10     Assign first region from list to A and remove from list
11     Determine neighbouring region of A with minimum merge cost (Alg. 4.2)
12     If no merge is valid, go to next region in list
13     Assign 'best merge' region from list to B and remove from list
14     Determine neighbouring region of B with minimum merge cost
15     Assign 'best merge' region from list to C and remove from list
16     While a mutual best merge does not exist ($A \neq C$)
17       Reassign A = B and B = C (i.e. step along gradient of best merge)
18       Determine neighbouring region of new B with minimum merge cost
19       Assign 'best merge' region from list to C and remove from list
20     Loop (While $A \neq C$)
21     When mutual best merge is found (A = C)
22       Merge regions A and B – update heterogeneity, perimeter and area
23       Replace regions A and B in master list with merged region
24   Loop
25 Loop if any merges were completed during iteration
26 Return master list of regions
```

Algorithm 4.1 The algorithm for the region merging segmentation method, where the 'local mutual best merge' is used at each location treated. The 'distributed treatment order' in line 8 plays an important role in allowing regions to grow evenly. Determining the neighbouring region with minimum merge cost (e.g. line 11) is detailed fully in Algorithm 4.2.

made that 'cost' less than a user-defined global merging parameter. The final segmented regions are then given unique labels for use in further processing.

The segmentation result is dependent on two user-defined parameters – the global merging parameter and the region shape parameter – and the defined treatment order. Algorithm 4.1 uses these parameters to control the size and appearance of the resulting segmented regions. The 'global merging' parameter defines the maximum permissible cost of a merge during segmentation, but this ultimately controls the size of the regions.

This parameter ranges from 0 and is boundless, due to the fact that both the measure of shape and heterogeneity are also boundless. The magnitude of the ‘global merging’ parameter is related to the magnitude of values in the field and their variance, but this relationship is non-linear and varies with each field. Therefore, the global merging parameter can only be set with some knowledge of the underlying data and otherwise via empirical methods. The ‘region shape’ parameter is a weighting parameter that controls how the heterogeneity and shape attributes combine when determining region merges. This parameter ranges from 0 to 1, with 0 producing an optimum segmentation based on the shape attribute (i.e. all regions will be close to square) and 1 producing an optimum segmentation based on heterogeneity (i.e. all regions will be optimally homogeneous).

Baatz and Schäpe (2000) introduced the ‘region shape’ parameter because regular, compact regions are most visually pleasing and desirable for image classification. In hydrology, depending on the proposed use of the regions, the shape parameter can often be left at 1, thus ensuring that regions represent the homogeneity of values most closely.

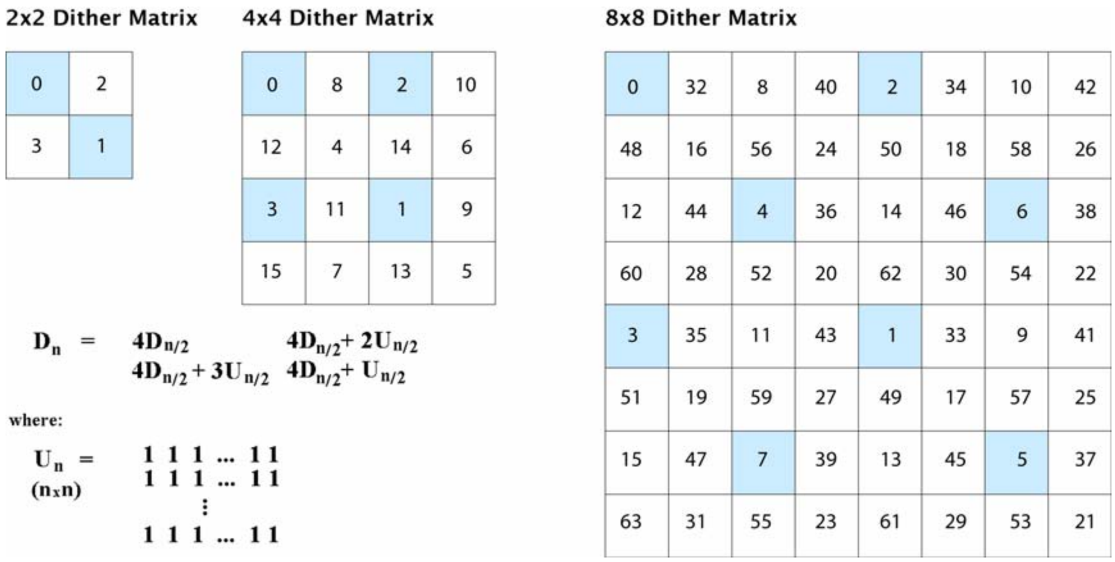
```
Assign minimum merge cost equal to global merging parameter
Initialise best neighbour as empty
Assign region shape parameter to W (i.e. weighting for region compactness)
Assign focus region to A
For each neighbouring region of A
    Assign neighbouring region to B
    Merge region A with B and assign to M
    Calculate cost of merging with this neighbour
        Note: Shp(X) = Shape of region X, Het(X) = Heterogeneity of region X
        Shp cost = [Area(M) * Shp(M)] – [Area(A) * Shp(A) + Area(B) * Shp(B)]
        Het cost = [Area(M) * Het(M)] – [Area(A) * Het(A) + Area(B) * Het(B)]
        Merge cost = [W * Het cost] + [(1 - W) * Shp cost]
    If (Merge cost < Minimum merge cost)
        Set best neighbour as region B
        Set minimum merge cost to merge cost
Go to next neighbouring region
Return best neighbour
```

Algorithm 4.2 The algorithm used for determining the neighbouring region with the minimum merge cost. This is evaluated using the user-defined parameters for ‘global merging’ and ‘region shape’ (as specified in lines 1-2 of Algorithm 4.1).

4.3.2.2 The treatment order

The order in which regions are treated during segmentation has a major influence on the way they are recognised. One solution is to allow only one ‘best merge’ to be made on each iteration, but this is slow and produces particularly large regions in the most homogeneous areas. Baatz and Schäpe (2000) suggest that a visually-intuitive segmentation has similarly sized regions throughout the spatial field. Therefore, Algorithm 4.1 treats each region once on each iteration to permit regions to grow equally, which is far more efficient but the results are heavily dependent on the treatment order. A random order can be used, but this produces non-repeatable segmentations. For repeatable segmentation results in which regions are allowed to grow simultaneously, the treatment order should follow two rules: 1) regions are processed in an order which ensures a maximum possible geographical distance from already processed regions (Baatz and Schäpe 2000); and 2) for each iteration, each region present at the beginning of the run is processed once at the most (Baatz et al. 2004).

The details of producing this type of distributed treatment order are not provided in any of the references or the patent for this segmentation method (Baatz and Schäpe 2000;



2x2 Dither Matrix

0	2
3	1

4x4 Dither Matrix

0	8	2	10
12	4	14	6
3	11	1	9
15	7	13	5

8x8 Dither Matrix

0	32	8	40	2	34	10	42
48	16	56	24	50	18	58	26
12	44	4	36	14	46	6	38
60	28	52	20	62	30	54	22
3	35	11	43	1	33	9	41
51	19	59	27	49	17	57	25
15	47	7	39	13	45	5	37
63	31	55	23	61	29	53	21

$$\mathbf{D}_n = \begin{matrix} 4\mathbf{D}_{n/2} & 4\mathbf{D}_{n/2} + 2\mathbf{U}_{n/2} \\ 4\mathbf{D}_{n/2} + 3\mathbf{U}_{n/2} & 4\mathbf{D}_{n/2} + \mathbf{U}_{n/2} \end{matrix}$$

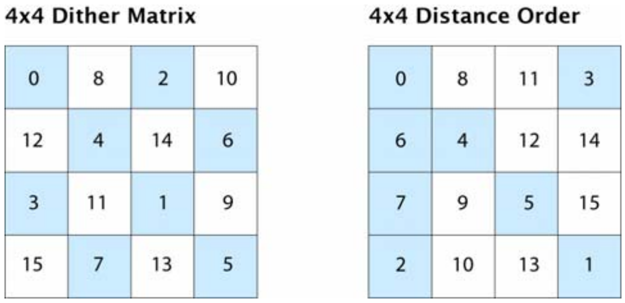
where:

$$\underset{(n \times n)}{\mathbf{U}_n} = \begin{matrix} 1 & 1 & 1 & \ldots & 1 & 1 \\ 1 & 1 & 1 & \ldots & 1 & 1 \\ & & & \vdots & & \\ 1 & 1 & 1 & \ldots & 1 & 1 \end{matrix}$$

Figure 4.4 Dither matrices of different sizes are produced by applying the recursive formula to the initial 2x2 dither matrix (Ulichney 1987). These matrices always have 2^n rows and columns. The first 2, 4 and 8 elements treated in these example dither matrices are shaded to illustrate the even distribution produced.

Definiens Imaging 2003; Baatz et al. 2004). However, it is suggested that it can be achieved by using a "dither matrix produced by a binary counter" (Baatz and Schäpe 2000, p.16). Dither matrices are used in image processing to reduce the colour palette used for display or printing. However, it is the arrangement of values within an ordered dither matrix (Ulichney 1987) that appears to be useful for defining a treatment order. Ordered-dither matrices are created by using a recursive formula based on a simple 2x2 matrix, as shown in Figure 4.4. This produces a repeating, 'checkerboard type' arrangement that is efficiently computed and exhibits a well-spaced distribution. Other algorithms, which explicitly maximise the geographical distance from previously treated elements, are unable to maintain a regular distribution once many elements have been treated (Figure 4.5).

In Algorithm 4.1, the distributed treatment order is determined using an ordered dither matrix that has an extent that is equal or larger than the spatial field being segmented. This ensures that every region is assigned a position in the treatment order during the first iteration. As each region is treated, it is removed from the treatment order until all regions have been treated once. During subsequent iterations, the positions in the treatment order are determined using the same general approach, although now there are multiple positions assigned to each region (because the regions overlap multiple elements). The first position in the order for each region is used, thus ensuring that each region is only treated once.



4x4 Dither Matrix

0	8	2	10
12	4	14	6
3	11	1	9
15	7	13	5

4x4 Distance Order

0	8	11	3
6	4	12	14
7	9	5	15
2	10	13	1

Figure 4.5 Two different treatment orders, one designed using a dither matrix (left) and the other using a simple algorithm that maximises distance from previously processed elements (right). The first 8 elements treated using each method are shaded, illustrating the better distribution obtained by using the dither matrix method.

The treatment order implemented in this thesis has only used the general ideas suggested in the literature (Baatz and Schäpe 2000; Definiens Imaging 2003; Baatz et al. 2004). Complete details of the treatment order used in their patented method have been unavailable in the literature or from the authors. The treatment order devised here is an approach that works suitably for hydrological spatial fields.

4.3.2.3 Using segmentation to identify regions

To illustrate the performance of the data-driven, region-merging segmentation method in different hydrological situations, all of the synthetic fields from Figure 4.1 have been segmented (Figure 4.6). The parameter values have been empirically set to produce 'mid-sized' regions (global merging = 500) that should be quite regularly shaped (region shape = 0.25). These parameters are kept constant for all of the fields. To illustrate the influence of the two user-defined parameters, the observed field from Figure 4.1 has been segmented using a range of parameter settings (Figure 4.7). The 'global merging' parameter ranges from 250 to 2000 and the 'region shape' is 1 or 0.25.

It is apparent from Figure 4.6 that the method is capable of producing visually-intuitive region definitions. The segmentations of the noisy fields detect some of the underlying arrangement that is also seen in the observed field, but they are greatly affected by the noise. The segmentation of the rearranged field identified only three small regions where there was some homogeneity, while the mean field had only the one region. All of the other segmentations appear to have produced logical region definitions, which is the desirable performance for this method.

The use of a shape parameter less than 1 has tended to produce more 'visually pleasing' segmentations of the observed field (Figure 4.7). This requirement for shape regularity has caused the regions recognised at each scale to have a less 'branched' appearance and has also produced larger scale regions in this example. The global merging parameter is the major control on region size and should be empirically set to produce the regions at a desirable scale. As both segmentation parameters vary with the spatial field being processed, recommended parameter values cannot be defined. However, the need to empirically determine the parameters is not considered a major limitation, as the regions must be visually inspected prior to their use.

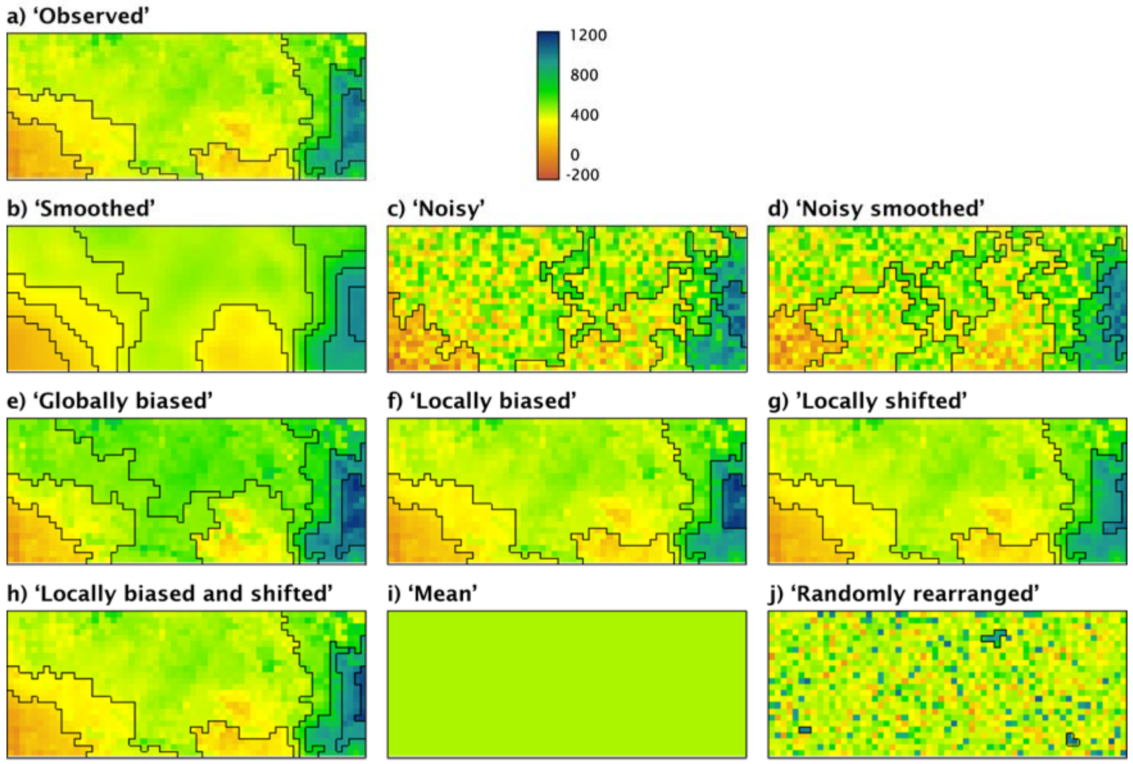


Figure 4.6 Regions determined by using region-merging segmentation for each synthetic field from Figure 4.1. The segmentations use the following parameters: global merging = 500; region shape = 0.25.

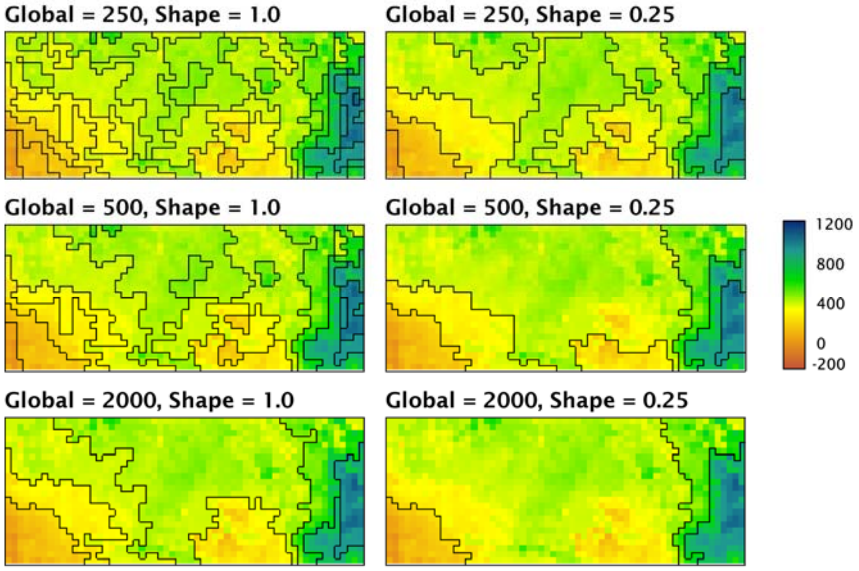

Global = 250, Shape = 1.0
Global = 250, Shape = 0.25
Global = 500, Shape = 1.0
Global = 500, Shape = 0.25
1200
800
400
0
-200
Global = 2000, Shape = 1.0
Global = 2000, Shape = 0.25

Figure 4.7 Various segmentations of the 'observed' synthetic field from Figure 4.1, showing the effects of different 'global merging' and 'region shape' parameters on the resulting regions recognised.

4.3.3 Knowledge-driven definition of regions

Knowledge-driven methods use surrogate information (e.g. terrain attributes) and/or subjective user knowledge to define regions within a spatial field. These 'artificially enforced' regions are not always apparent in the observed or modelled spatial fields, but they are used to represent regions of specific hydrological importance. Any knowledge-driven definition should aim to identify spatially-contiguous regions that include elements that have a specific hydrological meaning or respond similarly to the hydrological processes being investigated.

Some examples of regions of potential hydrological importance in different situations are listed in Table 4.3. Each example describes how the regions could be defined using surrogate fields and user knowledge. Details of these pre-processing methods are not given here because they are widely used throughout hydrology and image processing and have been discussed in 0 and 0.

The examples in Table 4.3 all make use of either surrogate or ancillary information to define importance. Because surrogate fields or user knowledge have been included into the definition of the regions, each region can be assigned a meaningful label. The labels are used to define the interpretation of any comparison measures calculated using the regions. When there are multiple contiguous regions having the same label (e.g.

Table 4.3 A list of example, knowledge-driven regions that could be used for defining hydrological importance during comparison. Some potential methods for defining each type of region are identified.

Region basis	Method of region definition
Geomorphic units	Defined by using terrain analysis methods to characterise each element within the spatial field; terrain attributes such as elevation, slope, aspect, profile curvature and Multi-Resolution Valley Bottom Flatness (MRVBF) (Gallant and Dowling 2003) are all useful; the field of terrain attributes should then be categorised and/or segmented to create spatially-contiguous regions
Soil/geology units	Defined or obtained from other data sources (e.g. remote sensing, in-situ observations); assign each element within the spatial field to the unit it is located in and processed to ensure spatial contiguity of each region
Similar climatic effects	Defined by using relationships between terrain and climatic processes (e.g. south- or north-facing slopes) or using other climatic data sets (with sufficient resolution); could be categorised into larger units with similar hydrological influences
Vegetation cover	Defined or obtained from other data sources; could be categorised into larger units that have similar hydrological responses

multiple different valleys within the extent of the spatial field), a decision must be made whether to label the regions as one non-contiguous region or as individual contiguous regions. When using regions from data-driven segmentation, meaningful region labels can only be defined based on the actual characteristics of the region (e.g. the average element value, the region size or shape). In these situations, a visual representation of the regions is used when interpreting comparison measures.

Knowledge-driven and data-driven methods can be combined together to produce specialised definitions of important regions within a spatial field. This has been undertaken with time series data in hydrology by Boyle et al. (2000). They processed a stream hydrograph and recognised 'regions' of driven and non-driven flow by using their understanding of how different hydrological processes appear in the data (i.e. user knowledge combined with data). These regions were used to compare each 'hydrologically important' region and were ultimately useful for model assessment.

Similar, specialised methods for defining regions can be developed for spatial fields using standard spatial analysis tools in GIS and the pre-processing methods discussed in this thesis. The main requirement is an understanding of how the hydrological process is revealed within the data. Morin et al. (2006) has recently used a specialised method to define rain cell regions within RADAR rainfall fields (Figure 4.8). The method grows rain cell regions from locally maximum rainfall intensity values. In this example, regions are used as the basis for characterising rainfall events rather than comparison.

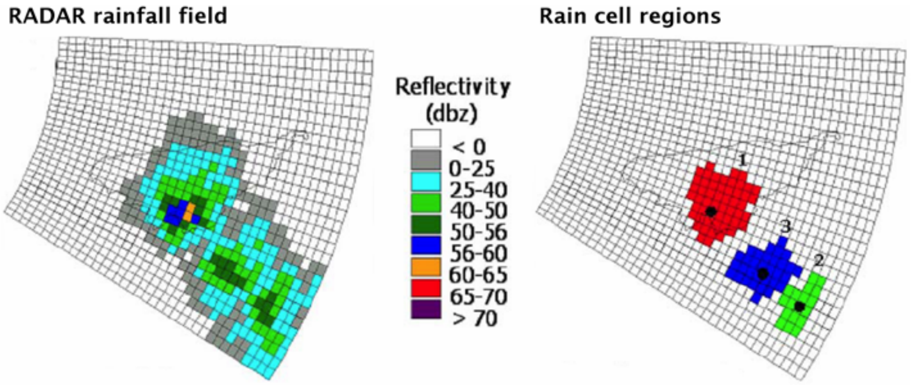


Figure 4.8 A RADAR rainfall field is processed by using both knowledge-driven and data-driven methods to identify rain cell regions deemed hydrologically important (adapted from Morin et al. 2006).

4.3.4 Region-based comparison method

The current comparison methods used in hydrology need to be adapted so that they can work with regions. Regions are defined for use in comparison for two different purposes: 1) to identify important groups of elements in the fields that should be compared separately from one another; and 2) to simplify the spatial field by defining structural features (e.g. homogeneous regions). These two purposes can be respectively thought of as local region-based comparison and global region-based comparison, because of the manner in which the elements within a region are compared.

In both of these methods, a comparison measure is produced for each region and should be interpreted relative to what the region is representing. For example, the resulting comparison measures might describe the bias in the model for simulating a particular structural feature in the observed field. Alternatively, when knowledge-driven regions have been used, the measure may reflect, for example, the model error on northern facing slopes (which may relate to evaporation issues in the model).

The following sections present: 1) the general region-based comparison algorithm; 2) a description of local and global region-based comparison measures, including a new similarity measure; and 3) an example showing region-based comparison applied to the observed synthetic field in Figure 4.1.

4.3.4.1 The region-based comparison algorithm

Algorithm 4.3 is a general algorithm that details the framework and order of processing for a region-based comparison method. When global region-based comparisons are undertaken, the variability within each region is characterised prior to comparison. When local region-based comparisons are used, any local comparison method can be applied to the elements within each region to produce the comparison measure.

Results from region-based comparisons are presented as a table of measures (one for each region) and an accompanying figure showing the regions (labelled with their meaning). Interpreting the result of a region-based comparison can be performed by contrasting the performance of each field for a particular region, or otherwise between all regions within a single field.

```
Specify source and target fields to be used
For each region
   Create weighting field (mask) to isolate region
   If local region comparison
      Multiply source field by weighting field (mask)
      Calculate intermediate measure for each source element
      Summarise error or similarity values
   If global region comparison
      Multiply source and target fields by weighting field (mask)
      Summarise source element values
      Summarise target element values
      Compare summary values for region
Go to next region
Tabulate comparison measure results for each region
```

Algorithm 4.3 The algorithm for a region-based comparison, using either global or local comparison measures. The results are produced as a table, with one comparison measure for each region.

It should be noted that this type of comparison does not compare different region definitions for each field. To compare different region definitions, such as undertaken by Power et al. (2001) with land use maps, the regions must equivalent labels (i.e. so that comparable regions can be determined). Such clear labels rarely exist for continuous hydrological fields and therefore this is not considered to be a suitable comparison method. Instead, one set of regions are enforced onto both fields being compared in these methods.

4.3.4.2 Local or global region-based comparison measures

Local region-based comparison is a simple, yet very effective, method that applies a local comparison method within each specific region. For each region, a local comparison measure is produced for each element in the ‘masked’ source field and it is summarised to describe the error or similarity within the region. When applied to every region, the relative performance within different parts of the spatial field can be assessed and contrasted (e.g. all models simulate a specific region well, but another region poorly). Any form of local comparison, including the tolerant methods introduced in this thesis, can be applied to produce this type of ‘specific local comparison’. This method is effective because it incorporates the user definition of importance to remove the confounding effects of any unimportant elements. Figure 4.9

illustrates this by showing the comparison of the set of observed regions overlain on the three different synthetic fields (from Figure 4.1) being compared.

In contrast to the specific local comparison method, global region-based comparison is a type of ‘feature based’ comparison. The region definition is enforced upon both fields being compared to change the structure of the field and to upscale the elements to an intermediate scale. A global comparison measure is then determined using the global characteristics derived from the elements within each region (i.e. the spatial arrangement is ignored). The effect of this ‘upscaling’ is illustrated in Figure 4.9 (right column), where each of the fields have been represented using the region mean. This is what would be compared during a global, region-based comparison that uses BIAS as the comparison measure. A low measure of BIAS for a region would indicate that the region is well simulated by the model. When large scale regions are used (as in this example), BIAS measures can causes large amounts of variation to be lost. This may be the purpose of defining the large scale regions as structural features and can be an effective way of reducing the impact of noise during comparison.

The global comparison methods used most commonly in hydrology are limited to either comparing the means or the variances. An alternative method, called the Kolmogorov-Smirnov test (Press et al. 1992), can be used to compare the distributions of two data sets. It uses the maximum separation found between the cumulative distribution functions of two continuous fields to describe how closely they ‘fit’ (Figure 4.10). If the distributions overlap in any way, then the separation will be less than 1. If they are identical, the separation is 0. This measure is particularly useful for global comparison methods, because it helps to use more complete information about element distributions during comparison.

This measure is used to describe similarity rather than error throughout this thesis. Therefore, it has been converted into a measure of similarity by taking the complement (i.e. one minus the separation value). This global comparison measure is herein referred to as the Kolmogorov-Smirnov similarity measure (KS).

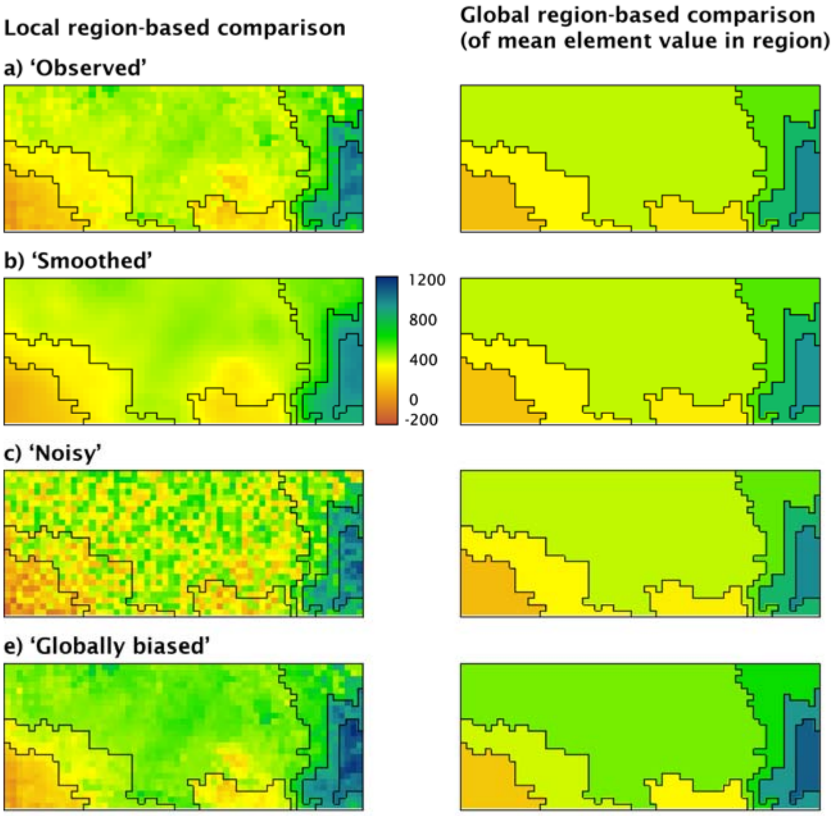

Local region-based comparison
Global region-based comparison
(of mean element value in region)
a) 'Observed'
b) 'Smoothed'
c) 'Noisy'
e) 'Globally biased'
1200
800
400
0
-200

Figure 4.9 Three of the 'modelled' synthetic fields from Figure 4.1 are compared with the 'observed' field using region-based comparison. The regions for this comparison are segmented from the observed field. Local region-based comparisons (left column) retain the spatial variation within each region during comparison. Global region-based comparisons using BIAS (right column) characterise the elements within each region and cause all variation to be smoothed out (based on the region template). When the KS measure is used for region-based comparison, only the variability of values (but not location and values) is used for comparison, thereby utilising more information during comparison.

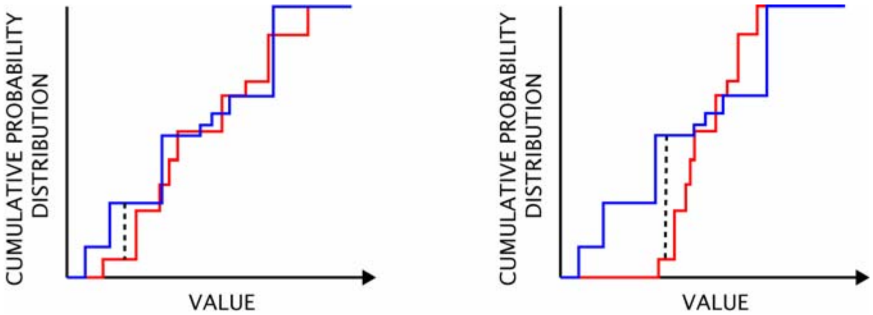


Figure 4.10 Two examples of how the Kolmogorov-Smirnov test statistic measures maximum separation between cumulative distribution functions (dotted line). The left example has a smaller separation distance and therefore receives a higher KS similarity value (i.e. 1 – separation distance).

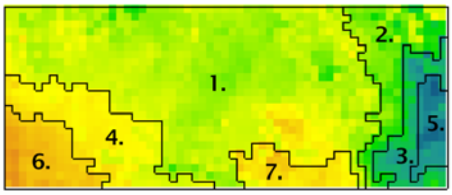
4.3.4.3 Using region-based comparison

To illustrate how region-based comparisons work and the information they produce, the 'modelled' synthetic fields from Figure 4.1 are compared with the 'observed' field using regions segmented from the observed field (Figure 4.9). This example uses MAE for local region-based comparison. Global region-based comparison is measured using BIAS and the suggested KS measure. The results for each region and the actual region definitions are detailed in Table 4.4.

Inspecting the MAE measures in Table 4.4 reveals that no individual region is found to have low error in all of the models. Nor does any model have low error for all regions. The noisy models have a regular amount of error in all regions, with no region being particular better or worse. The 'locally deformed' models only have errors in regions 3 and 5, as would be expected. The smoothed model is found as the best overall representation of the observed field (i.e. ignoring the locally deformed models). This is not as apparent when making global region-based comparison using BIAS, which finds the noisy models to be a better representation (due to the effect of upscaling). This is most noticeable in region 1 for the noisy models, where the noise is averaged over the largest region. The global KS measure finds similar performance for the smoothed and noisy models, although regions 4 and 5 are not as well 'modelled'.

When only region 5 is analysed (i.e. the highest magnitude values), the MAE finds no model that simulates the element values well. However, when the region is treated as an irregular element by using the global BIAS measure, the noisy field is recognised as being a good representation (i.e. the noise is averaged out, leaving a similar mean value). The KS measure confirms this, showing that the distributions of the values within this region are actually quite closely matched (KS=0.81), even though the local errors accumulate to suggest otherwise. All of the measures used here fail to note the perfect correlation between the globally biased model and the observations. This highlights the value of also using correlation-based measures (e.g. RSQ or COE) for the local region-based comparisons.

Table 4.4 The results of using region-based comparison with the synthetic fields from Figure 4.1. One local method (MAE) and two different global methods (BIAS, KS) have been used to produce comparison measures for each region (numbered according to the figure).



MAE	1	2	3	4	5	6	7
b) Smoothed	16.6	46.0	58.8	16.7	71.2	21.6	30.9
c) Noisy	86.8	92.1	89.3	85.6	75.9	77.8	81.9
d) Noisy smoothed	88.5	106.8	103.9	88.7	96.9	92.3	101.0
e) Globally biased	64.0	81.4	125.3	50.8	165.0	22.7	41.5
f) Locally biased	0.0	0.0	5.1	0.0	117.4	0.0	0.0
g) Locally shifted	0.0	0.0	4.7	0.0	78.9	0.0	0.0
h) Locally biased and shifted	0.0	0.0	8.7	0.0	62.9	0.0	0.0
i) Mean	29.3	79.0	291.1	81.4	489.8	221.7	128.1
j) Randomly rearranged	93.8	132.9	318.1	101.5	485.3	228.8	139.9

BIAS	1	2	3	4	5	6	7
b) Smoothed	-1.3	11.3	-2.8	1.6	-43.9	11.2	18.5
c) Noisy	-0.6	-4.0	9.6	-3.7	2.3	-7.7	22.6
d) Noisy smoothed	-1.2	12.8	-12.7	4.5	-38.9	17.4	6.6
e) Globally biased	64.0	81.4	125.3	50.8	165.0	22.7	41.5
f) Locally biased	0.0	0.0	5.1	0.0	117.4	0.0	0.0
g) Locally shifted	0.0	0.0	4.2	0.0	-61.8	0.0	0.0
h) Locally biased and shifted	0.0	0.0	8.6	0.0	18.2	0.0	0.0
i) Mean	15.5	-71.6	-291.1	81.4	-489.8	221.7	128.1
j) Randomly rearranged	16.3	-87.3	-267.7	77.2	-481.5	224.5	122.1

KS	1	2	3	4	5	6	7
b) Smoothed	0.90	0.84	0.89	0.93	0.52	0.86	0.78
c) Noisy	0.70	0.78	0.82	0.67	0.81	0.80	0.63
d) Noisy smoothed	0.67	0.76	0.80	0.65	0.75	0.73	0.66
e) Globally biased	0.29	0.55	0.40	0.41	0.38	0.74	0.53
f) Locally biased	1.00	1.00	0.96	1.00	0.52	1.00	1.00
g) Locally shifted	1.00	1.00	0.98	1.00	0.69	1.00	1.00
h) Locally biased and shifted	1.00	1.00	0.98	1.00	0.83	1.00	1.00
i) Mean	0.32	0.13	0.00	0.00	0.00	0.00	0.00
j) Randomly rearranged	0.81	0.40	0.12	0.38	0.06	0.11	0.22

4.3.5 Summary of importance methods

The importance methods discussed here are primarily dependent on pre-processing to define what is important. Importance is essential during comparison so that specific tests with specific meaning can be made. Without importance, quantitative comparison measures can only describe the overall error (or similarity) of a model, which is unclear to interpret and is very limited when trying to understand why a model performs well.

In hydrological applications, importance is usually defined based purely on measurement space (e.g. values over a certain threshold). However, there are useful methods that allow spatially-contiguous regions of related elements to be defined. These can be defined based on the data (i.e. segmentation methods) or by using user knowledge (e.g. terrain regions). Once defined, the regions can be used in the 'specific local' and 'region-based comparison methods' to compare each region. Both methods produce a comparison measure for each region, although these could be combined into a single measure using some type of weighting if required.

Importance has previously been considered in hydrological applications, such as in time series analysis where the important parts of a hydrograph can be separately used during comparison. However, the specific importance of spatially-related regions is rarely used for model assessment. These new pre-processing methods and region-based comparisons are directly useful for model assessment, as they provide performance measures with specific meaning, which may then be used for multi-criteria optimisation. Importance methods are considered less applicable to data assimilation, which is generally concerned with making corrections to the whole model extent.

4.4 Tolerance

The theme of tolerance is incorporated into comparison methods at the local scale, where is provides greater control over the local comparison of elements. Currently, strict local comparison methods are exclusively used at the local scale, with no other options for local comparison. Tolerance allows local comparisons to be more flexible and to consider the spatial (and temporal) relationships between elements that are nearby one another.

Currently in hydrological modelling, the notion of tolerance is only applied at a global scale. After a comparison measure has been calculated between two fields, the result is usually evaluated relative to some level of expected performance that implicitly includes tolerance for uncertainty. For example, the MAE measure for a soil moisture model might be calculated at 10% v/v, but it is assumed to be good performance because there is an 'implicitly tolerated' measurement error of 4% v/v. Explicitly including tolerance into the local comparison process enables a more direct application of value tolerances (i.e. the 4% v/v tolerance would be considered during comparison of each measurement, rather than globally). This approach also enables tolerances to be specified regarding the location of elements, which can allow for minor mismatches in the location of values. These ideas can also be applied globally for tolerating temporal differences between spatial fields (e.g. to permit minor timing problems during comparison).

Tolerances can be applied to either: 1) represent known uncertainties in the target (i.e. the reality) used for comparison; or 2) define the differences that are deemed insignificant to the user and apply to both source and target. Both of these uses are actually just different rationales for using the same general process, but the setting of tolerance amounts and interpretation of the results differ. In the first application, the source field is compared against an uncertain target. The tolerance amounts are defined based on the known uncertainty of the target and they allow it to change within the tolerances to optimally match the source field. This provides a 'best case' result describing the error remaining in the source field.

The second application of tolerances uses informed decisions about what differences are considered insignificant during comparison. These differences define the tolerance amounts. The same comparison process is used (i.e. source held constant while target changes to find optimal match), but it is repeated in the opposite direction (i.e. switch source and target around) to apply the tolerances to the source field also. The results of the two comparisons are averaged for each element to identify the significant differences that exist between the fields. This assumes that both of the fields are equally uncertain. Due to the fact that both fields are 'changed' in this type of comparison, the resulting measure cannot be attributed to either field directly, but it does provide a general comparison measure.

In hydrological modelling, the first case mentioned is the most common and useful. Modelled fields are usually evaluated by comparison with observed fields (which may have uncertainty in them). The uncertainty of the observed fields is tolerated during comparison so that it is not included into the measure of model error (as occurs with strict comparison). The second case is not so common for model assessment because it cannot attribute the result to either field and it allows the model to change (which should not occur when it is being assessed). In this thesis, the two direction comparison is not pursued further, although it can be easily produced using the algorithm described in this section.

A suitable comparison method that can evaluate fuzzy similarity for categorical spatial fields is used in land-use modelling (Hagen 2003) and provides the inspiration for the tolerant method implemented here. The definition of fuzzy similarity is fundamental to the comparison method of Hagen (2003), but it is not appropriate for direct use with continuous hydrological fields. This is because it only defines categorical value similarities and is only able to describe similarity, not error (which is more commonly used in hydrology).

Therefore, a new and more versatile approach for defining 'tolerance' is developed here. The new tolerances are defined in a manner that should be familiar to hydrologists and they enable tolerant comparison measures of both error and similarity to be produced. They also permit tolerance of timing differences during comparison of spatial fields or series, which has not previously been possible. The following sections detail: 1) the new approach to define tolerances for continuous spatial fields; 2) the general algorithm for the new 'tolerant local comparison method'; 3) the comparison measures that are produced from the tolerant method; and 4) examples of tolerant comparison being used with the synthetic fields from Figure 4.1. Preliminary findings from this work have been previously published in Wealands et al. (2004; 2005b).

4.4.1 Defining tolerances for continuous fields

Tolerances are defined for each component of a spatial field that is variable. These components are the time represented by the spatial field (T); the location (L) of an element value; and the element value (V). Tolerances are used during comparison to allow a source field and its elements to be compared against all 'similar' target fields and elements. This can be better understood by referring to Figure 4.11. In this figure, the source field (or element) is shown in the top row and the 'similar' target fields (or elements) are represented in the bottom row. For each different type of tolerance (and the combination of tolerances), the number of comparisons made between the source and target is described and illustrated. It shows that once all three types of tolerance are combined, there are many potential target elements that may be similar to the source elements.

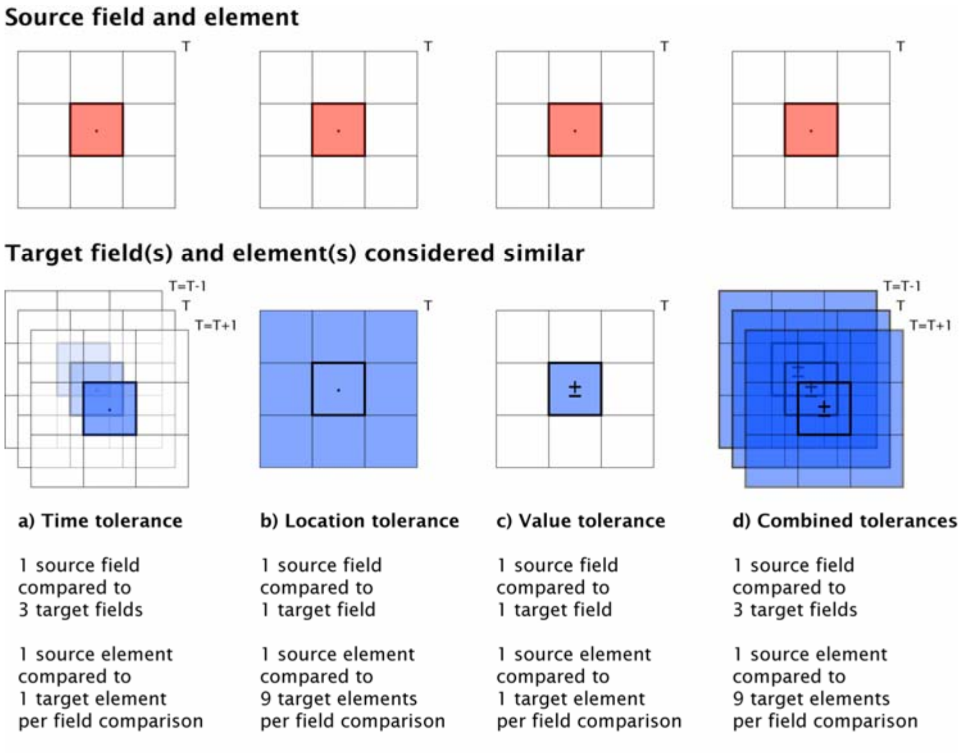

Source field and element
T
T
T
T
Target field(s) and element(s) considered similar
T=T-1
T
T=T+1
T
T
±
T=T-1
T
T=T+1
a) Time tolerance
1 source field compared to 3 target fields
1 source element compared to 1 target element per field comparison
b) Location tolerance
1 source field compared to 1 target field
1 source element compared to 9 target elements per field comparison
c) Value tolerance
1 source field compared to 1 target field
1 source element compared to 1 target element per field comparison
d) Combined tolerances
1 source field compared to 3 target fields
1 source element compared to 9 target elements per field comparison

Figure 4.11 A conceptual diagram showing how tolerance definitions change the number of fields and elements being compared. In these examples a) demonstrates a time tolerance of 1 time step; b) demonstrates a location tolerance of 1 element; c) demonstrates a value tolerance (note ± on target element); and d) demonstrates the combination of the three other tolerances. As the tolerance definition changes, so does the number of comparisons made between the source and target fields and elements. In all cases, the source field is kept constant, so the tolerances are applied to the target field.

Two different ways for defining these tolerances are developed in this thesis – crisp tolerances and fuzzy tolerances. Crisp tolerances define any differences between elements that are permitted during comparison. When elements are compared they are considered as either ‘the same’ or ‘different’ based on these tolerances. This makes it possible to measure the local error for each element (i.e. in the same units as the field). A tolerant measure of error is useful because it allows all differences to contribute to the summary measure (i.e. large errors will cause an overall larger error). Crisp tolerances should not be confused with ‘strict’ comparisons, which do not incorporate any notion of tolerance.

Fuzzy tolerances are used to define a spectrum of differences between elements that are designated as being similar to some degree. When elements are compared, they will be considered as having a defined level of similarity based on these tolerances. Unlike with crisp tolerances, fuzzy tolerances can only produce a measure of local similarity for each element. This makes them unable to describe the typical error in a model, only the recognised similarity. The local similarity values are always in the range [0, 1] and represent the spectrum from ‘no similarity’ to ‘same’.

The tolerances use the differences found between the times of spatial fields and the location and value of elements (ΔV, ΔL, ΔT). The following formulae are used to determine the differences between each component of the elements being compared:

$$\begin{aligned}&\Delta \mathrm{V} = \mathrm{V(source\ element) - V(target\ element)}\\&\Delta \mathrm{L} = \mathrm{Distance\ measured\ between\ L(source\ element)\ and\ L(target\ element)}\\&\Delta \mathrm{T} = \mathrm{T(source\ field) - T(target\ field)}\end{aligned} \qquad (4.1)$$

The crisp tolerance (CTOL) or fuzzy tolerance (FTOL) functions define how the difference for each component is translated into values describing the error (EV, EL, ET) or similarity (SV, SL, ST). These values are then combined together (via multiplication) to describe the ‘tolerated local error’ (ELOC) or ‘tolerated local similarity’ (SLOC) between two elements by using the following formulae:

$$
\begin{aligned}
&\mathrm{EV} = \mathrm{CTOL}(\Delta \mathrm{V}) \in \Re && \mathrm{SV} = \mathrm{FTOL}(\Delta \mathrm{V}) \in [0,1] \\
&\mathrm{EL} = \mathrm{CTOL}(\Delta \mathrm{L}) \in \{1, \text{undefined}\} && \mathrm{SL} = \mathrm{FTOL}(\Delta \mathrm{L}) \in [0,1] \\
&\mathrm{ET} = \mathrm{CTOL}(\Delta \mathrm{T}) \in \{1, \text{undefined}\} && \mathrm{ST} = \mathrm{FTOL}(\Delta \mathrm{T}) \in [0,1] \\
&\mathrm{ELOC} = \mathrm{ET} * \mathrm{EL} * \mathrm{EV} \in \{\Re, \text{undefined}\} && \mathrm{SLOC} = \mathrm{ST} * \mathrm{SL} * \mathrm{SV} \in [0,1]
\end{aligned}
\qquad (4.2)
$$

This describes how tolerances are used at a conceptual level (i.e. converting differences between elements into estimates of error or similarity). During implementation, it is necessary to define the actual CTOL or FTOL 'functions'. The following sections describe: 1) the simple functions designed for use in hydrological applications; 2) how to choose appropriate settings for these tolerance functions; and 3) options available for designing more specialised tolerances.

4.4.1.1 Tolerance functions

Tolerance functions are used to translate the value, location and/or time differences between elements into error or similarity values, which are then combined. Any function can be used to achieve this provided the output range of the function follows the definitions in Equation 4.2. When defining tolerances for hydrological comparisons throughout this thesis, linear decay functions and absolute differences (i.e. non-directional) are exclusively used. These are chosen because they are simple starting points and they can be specified using characteristics of spatial fields that are frequently reported (e.g. measurement error, positional accuracy).

For crisp tolerance functions, the maximum 'allowed difference' for each component of tolerance is all that needs to be defined. For the time and location components (ET and EL), this defines which other elements or fields it can be compared with. These

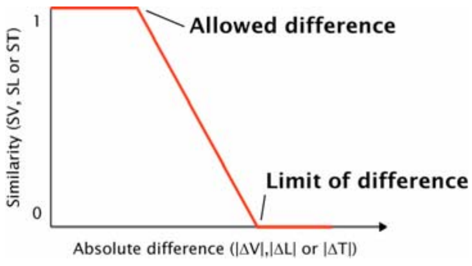


Figure 4.12 The two parameters required for defining fuzzy tolerances – maximum 'allowed difference' and 'limit of difference considered similar'.

tolerated elements are assigned a value of one; all other elements are considered to be undefined (i.e. they are not comparable). For the value component of the element (EV), the maximum allowed error is subtracted from the actual value difference (i.e. reduces the error by the tolerated amount). This is only reduced until the error reaches zero, in which case the element values are considered equal. By multiplying ET, EL and EV, the tolerated local error between each pair of elements is obtained.

Fuzzy tolerance functions require additional parameters to be specified. For each component of tolerance, the maximum 'allowed difference' is defined along with the 'limit of difference considered similar'. These parameters are denoted by the $_{\text{ALLOW}}$ and $_{\text{LIMIT}}$ subscripts for each element difference (e.g. $|\Delta V|_{\text{ALLOW}}$, $|\Delta V|_{\text{LIMIT}}$) (Figure 4.12). Any actual difference that is less than the allowed difference is assigned a similarity of one. Any actual difference that is greater than or equal to the 'limit of difference' is assigned similarity of zero. In between these values, a linear decay function is used to translate the difference into a similarity value.

Figure 4.13 illustrates the parameters that are used for either crisp or fuzzy tolerances. For each component of tolerance, the parameters have been applied to a simple situation. For example, Figure 4.13b shows an analysis window of elements surrounding the current element. The crisp tolerance designates $|\Delta L|_{\text{ALLOW}}$ = 1.5

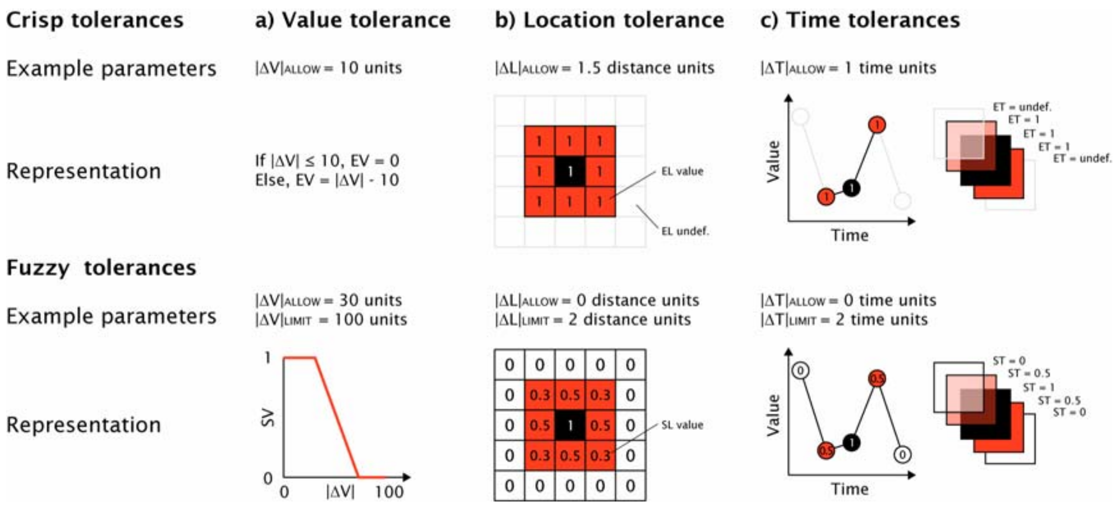


Figure 4.13 Examples of the parameters used for crisp tolerances (top) and fuzzy tolerances (bottom). For each component of tolerance, the parameters are converted into a conceptual representation showing the EV/EL/ET or SV/SL/ST values that are determined.

distance units, which allows any element directly neighbouring to be considered as having no error (and therefore being comparable). A tolerance of 1.5 units was specified so that the diagonal elements were also included. The fuzzy tolerance for location specifies $|\Delta L|_{ALLOW} = 0$, $|\Delta L|_{LIMIT} = 2$. This produces a location similarity (SL) of 1 for the same location and 0.5 for horizontal neighbours that are 1 unit away (i.e. half-way along the linear decay function from 0 to 2). All of the other representations given in Figure 4.13 help to illustrate how the two simple parameters (i.e. $_{ALLOW}$ and $_{LIMIT}$) are converted into expressions of tolerance that are used during comparison.

4.4.1.2 Choosing appropriate tolerance values

The definition of tolerances is a subjective process that should be primarily guided by knowledge of the spatial fields being compared. Tolerances should be chosen to reflect the existing uncertainty within the target field/elements; or the strictness with which the target field/elements need to be matched during comparison. By introducing tolerances, the error measured between the fields will be reduced (and the similarity increased). As the tolerances increase in magnitude, the error approaches zero (and similarity approaches one), but the actual rate of change is dependent on the fields being compared.

When tolerances are set to tolerate the uncertainty in the target field/elements, the purpose is to remove the effects of the uncertainty on the resulting comparison measure. This helps to produce a measure that better reflects only the true model errors. The tolerances can be set directly from reported characteristics of the uncertainty in the target field. For example, characteristics such as measurement error or spatial accuracy are often available for spatial fields available in hydrology. The $|\Delta V|_{ALLOW}$ parameter can be set at the measurement error, thus rendering any lesser value difference to be insignificant.

Where such values are not reported or known, an understanding of the measurement method or pre-processing used to create the target field can be used to determine appropriate tolerances. Güntner et al. (2004) describes a hydrological example where modelled field elements were derived from a 50m elevation model, while the observations were collected with 10m spacing. Because of the scale difference, location tolerances equivalent to $|\Delta L|_{ALLOW}$ = 0m, $|\Delta L|_{LIMIT}$ = 50m were used for a binary field comparison. This example applies tolerances to manage the uncertainty arising from a spatial scale difference. Similar logic can be applied to devise appropriate time or value tolerances. For example, when there are differences in the precisions of the observed and modelled values, frequent small value differences will exist. With continuous fields, these insignificant differences will all be accumulated into an error measure and they can even preclude similarity from being recognised (i.e. because the values are not equal).

The other approach to choosing tolerance values is for subjectively controlling the strictness of a comparison. If, for example, a location difference of 1 element is considered insignificant, then this can be explicitly tolerated and the measures of error (or similarity) will reflect these conditions. Using this approach allows simple hypotheses to be tested. For example, the response of a comparison measures to different tolerance conditions (e.g. with and without timing tolerances applied) can indicate the presence/absence of specific effects in certain models. The existing applications of categorical fuzzy comparison (e.g. Hagen 2003; Hagen-Zanker 2005; White 2006) generally use subjective methods for defining the similarity between values and locations, although there has been a method developed for translating expert knowledge about categorical relationships (e.g. between different land use categories) into tolerance definitions (Fritz and See 2005).

The tolerances for value, location and time do not act independently of one another. For example, a difference in location (or time) between elements (or fields) will actually appear as a value difference during comparison of continuous fields. Therefore, a value tolerance can be used to implicitly tolerate differences caused by these other factors, but doing so potentially leads to significant errors being tolerated, which is undesirable. The relationships that exist between different components of tolerance vary with each field and cannot be quantified. Therefore, when explicitly tolerating multiple types of differences (T, L and/or V) during comparison, the potential impact of all differences occurring together should be considered. For example, if value and location tolerances are used together during comparison, they effectively allow the target field to be spatially smoothed.

If using tolerances together during comparison, it is recommended to set the tolerances amounts conservatively because their combined effect may be greater than desired. However, when the tolerances are being applied individually (e.g. just a value tolerance), they may be set using known characteristics of the data (which are usually inclusive of many combined differences anyway). The ultimate goal of using tolerances is to remove the impact of insignificant or confounding differences from the comparison. Experience is needed to know what tolerances are suitable for specific comparison situations, although some empirical analysis of the data may be valuable for design suitable tolerances.

4.4.1.3 Specialised tolerances

The tolerances applied during analysis in this thesis are kept fairly simple for the purposes of illustration and understanding. However, more specialised tolerances can be easily implemented by modifying the way tolerances are defined. One of the simplest extensions to tolerances is to make them only tolerate differences in certain directions. For example, only tolerating positive value differences can ensure that only over-estimations are tolerated. Similarly, directional time tolerances might be used to ensure that only fields that are 'too late' are tolerated.

As well as considering the direction of differences, a specialised tolerance can utilise additional information to make tolerances that are adaptive (i.e. the tolerance changes with the element being compared). For example, a terrain model could be used to determine which nearby elements are located upslope and within a certain distance (i.e. to tolerate only elements that can contribute flow downslope). Alternatively, the tolerance could use a soils map and be adaptive to different soil types (i.e. have a different measurement error in each soil type).

Specialised tolerances allow any type of local comparison process to be undertaken. These are only facilitated by the generally-applicable definition of tolerances described in this section and developed in this thesis. The major reason behind this adaptability is the fact that tolerances are all applied locally, during the comparison of each individual element in the source field. The tolerances are able to use the 'raw' information about the elements to formulate an estimate of local error or similarity. While the analysis in this thesis focuses on showing the general application and utility of the new comparison methods for use in hydrological modelling, there is scope for further research and development of specialised tolerances.

4.4.2 The tolerant local comparison algorithm

The tolerant local comparison algorithm begins by defining the source and target fields. The source field is the subject of the comparison (i.e. the local comparison compares every element within the source field). This field is held constant so that the differences between it and the target(s) can be evaluated. These differences are fundamental to the evaluation of tolerances (Equation 4.1). The tolerant local comparison algorithm (Algorithm 4.4) can compare one source field to any number of target fields. To compare multiple sources, the algorithm is simply run multiple times using the same target each time. The algorithm produces both a tolerant local comparison measure using one of the summary measures in Algorithm 4.5-Algorithm 4.8 and also an associated intermediate field of tolerant local error or similarity values.

The source field is held constant in this algorithm and the time tolerance is used to determine if each target field is temporally similar. If so, it is compared against the source. When comparing these temporally similar fields, each source element is held constant. The location of each element in the target field is evaluated using the location tolerance to see whether it is spatially similar. If so, the source and target element values are compared using the value tolerance. This determines the similarity or error of the element values, which is combined with the spatial and temporal similarities to assign an overall similarity or error for the source element (based on all of the tolerances). After conducting this for any similar fields and elements, the target field that is most similar is used to produce the final summary measures.

```
Define crisp or fuzzy tolerances for value, location and time
For each target field (i.e. each time)
   Compute ET or ST between source and target fields
   If time difference is tolerated (ET=1 or ST>0)
      For each source element (i.e. the focus element)
         For each target element
            Compute EL or SL between focus and target elements
            If location difference is tolerated (EL=1 or SL>0)
               If tolerances are crisp
                  Compute EV and then ELOC (ET*EL*EV)
                  Add to list of potential ELOC values for focus element
               If tolerances are fuzzy
                  Compute SV and then SLOC (ST*SL*SV)
                  Add to list of potential SLOC values for focus element
         Go to next target element
         If list of ELOC/SLOC is not empty
            Assign optimal value (min ELOC or max SLOC) to focus element
         Else focus element is undefined
      Go to next source element
      Add to list of potential intermediate fields
Go to next target field
For each potential intermediate field
   Compute summary measure (TMAE, TCOE, TMS or TCE) for each field
Go to next potential intermediate field
Return optimal summary measure and associated intermediate field
```

Algorithm 4.4 The general algorithm for the tolerant local comparison method. This algorithm produces a summary measure (i.e. using one of Algorithm 4.5-Algorithm 4.8) and an associated intermediate field (of tolerant error or similarity values).

4.4.3 Tolerant comparison measures

The summary comparison measures produced from a tolerant local comparison method resemble those produced from a strict local comparison. The current strict local comparison measures are RMSE, MAE, RSQ and COE. Of these measures, RMSE and MAE are both direct summaries of the intermediate field of errors and describe the typical error. In contrast, RSQ and COE both relate the errors to a known, adaptive reference (i.e. based on the actual data) and produce similarity measures.

There are two general tasks for which comparison measures are needed: 1) intra-comparison (i.e. contrasting the performance of many different models all compared to the same observation); 2) inter-comparison (i.e. contrasting the performance of different models from different times/locations). For intra-comparison tasks, the distributions of errors produced during each comparison are generally assumed to be equivalent (because the models are for the same time/location and are compared to the same observation). This allows absolute measures of error (i.e. those not relative to any benchmark) to be quite safely used and contrasted. However, for inter-comparison tasks, the distributions of errors cannot be assumed to be equivalent. These tasks benefit from measures that are standardised by a known reference. The widespread recognition and understanding of the COE measure within hydrology is evidence of the benefit of these standardised measures.

Two comparison measures for summarising the intermediate error field produced from using crisp tolerances are presented here. One measure is a direct summary of the error field; the other is adjusted to a known reference. A similar set of two measures are also developed for the field of similarity values produced from using fuzzy tolerances.

4.4.3.1 Crisp tolerant comparison measures

Following the suggestions of Legates and McCabe (1999), the MAE measure is used in favour of RMSE throughout this work so that localised large residuals do not unduly influence the resulting comparison measure. The familiar algorithm for MAE is applied to the intermediate field of tolerant local errors produced from using crisp tolerances. The resulting measure is termed tolerant mean absolute error (TMAE) (Algorithm 4.5).

The tolerant coefficient of efficiency (TCOE) (Algorithm 4.6) is a relative measure of performance that uses an adaptive reference field to standardise the result (for use in inter-comparison tasks). In this algorithm, the error between the source and target fields is judged in comparison to the error found between the reference and target fields.

```
Use field of ELOC values from Algorithm 4.4
For each element in field of ELOC values
   If element has ELOC value
      Update running total of absolute ELOC values
      Increment count of elements
Go to next element in field of ELOC values
TMAE = total of ELOC values / count of elements
```

Algorithm 4.5 The algorithm for the measure of tolerant mean absolute error (TMAE).

```
Define reference field (default is mean of target field)
Use Algorithm 4.4 to produce field of ELOC values for source/target
Use Algorithm 4.4 to produce field of ELOC values for reference/target
For each element in ELOC field (source/target)
   Find spatially coincident element from ELOC field (reference/target)
   If both elements have ELOC values
      Update running total of absolute ELOC values (source/target)
      Update running total of absolute ELOC values (reference/target)
Go to next element in ELOC field (source/target)
TCOE = 1 – [total ELOC(source/target) / total ELOC(reference/target)]
```

Algorithm 4.6 The algorithm for the tolerant coefficient of efficiency (TCOE).

The default reference field is designated as the mean target field (i.e. usually the observed field). Every element in the mean target field contains the mean value of all target elements. This is chosen as the reference because it is synonymous with the observed mean, as used in the popular COE measure (Nash and Sutcliffe 1970; Legates and McCabe 1999). The mean field also represents the 'no skill' performance, in which no spatial variability has been represented. An alternative reference field can be substituted (e.g. a longer-term mean field) if desired. The TCOE measure is interpreted in the same way as a strict COE measure. A value of 0 denotes performance equal to the reference; negative values denote worse performance than the reference; positive values approaching one denote good performance.

When the relative measure is being calculated, the tolerances are applied to the comparisons between both source/target and reference/target. Therefore, if the tolerances are large relative to the variance in the field, the reference field may be found to be highly similar to the target field. In these situations, positive values for TCOE are difficult to attain. If the tolerances cause the reference to be the same as the target, the measure is undefined (i.e. divide by zero situation).

4.4.3.2 Fuzzy tolerance comparison measures

The TMAE and TCOE measures developed for crisp tolerances cannot be directly applied to the intermediate field of similarity values output from fuzzy tolerant comparisons. With these comparisons, the simplest summary measure is the tolerant mean similarity (TMS) (Algorithm 4.7). This describes the average similarity found per element in the field. It should not be interpreted as being the proportion of the field elements that are similar, because many elements can have partial similarity.

As with the crisp tolerances, a relative measure termed tolerant similarity efficiency (TSE) is developed. Algorithm 4.8 differs slightly from the TCOE algorithm, although the mean target field is again used as the default reference field. The interpretation and tolerance issues described for TCOE apply to this measure also.

```
Use field of SLOC values from Algorithm 4.4
For each element in field of SLOC values
   If element has SLOC value
      Update running total of SLOC values
      Increment count of elements
Go to next element in field of SLOC values
TMS = total of SLOC values / count of elements
```

Algorithm 4.7 The algorithm for the tolerant mean similarity (TMS) measure.

```
Define reference field (default is mean of target field)
Use Algorithm 4.4 to produce field of SLOC values for source/target
Use Algorithm 4.4 to produce field of SLOC values for reference/target
For each element in SLOC field (source/target)
   Find spatially coincident element from SLOC field (reference/target)
   If both elements have SLOC values
      Update running total of SLOC values (source/target)
      Update running total of SLOC values (reference/target)
      Increment count of elements
Go to next element in SLOC field (source/target)
TSE = [total SLOC(source/target) - total SLOC(reference/target)] /
       [count of elements - total SLOC(reference/target)]
```

Algorithm 4.8 The algorithm for the measure of tolerant similarity efficiency (TSE).

4.4.4 Using tolerant comparison methods

The tolerant comparison method is applied here to make comparisons between the synthetic fields from Figure 4.1. Tolerant comparisons are undertaken using both crisp and fuzzy tolerances to five different synthetic fields. Each field that is compared has a known deformation from the 'observed' field. The tolerances are defined differently for each comparison and attempt to tolerate the introduced deformation. To evaluate the ability of the tolerant methods, both strict and tolerant methods are analysed here. For each comparison, the basic quantitative measures (TMAE or TMS) and the intermediate fields are analysed.

4.4.4.1 Crisp tolerance comparisons

The quantitative results of the five comparisons made using crisp tolerances are listed in Table 4.5. The associated tolerance definitions and intermediate fields are detailed and shown in Figure 4.14. For these fields, the MAE and TMAE measures can be directly contrasted to see the impact of introducing tolerance during each comparison.

Example b) compares the smoothed field and allows a location difference of 2 elements to tolerate the scale difference introduced by smoothing. The location tolerance causes a significant decrease in the typical error found in the model. The intermediate field reveals this to occur across the whole field and the location tolerance has been particular useful for removing the large errors (i.e. by finding a better matching element nearby). Example c) compares the noisy field and allows value differences of 100 units to account for the 'measurement error' that was introduced. The strict comparison has MAE of 86.1, which is reduced to 15.7 after the tolerances were locally applied. If the same value tolerance were applied globally (i.e. after comparison, as is currently performed in hydrology), the fields would be considered similar (when in fact an average error of 15.7 still exists).

Table 4.5 The comparison measures produced during strict and crisp comparisons of the synthetic fields from Figure 4.1.

Field	MAE	TMAE
b) Smoothed	24.7	4.5
c) Noisy	86.1	15.7
f) Locally biased	4.3	1.7
g) Locally shifted	2.9	0.2
h) Locally biased and shifted	2.6	0.1

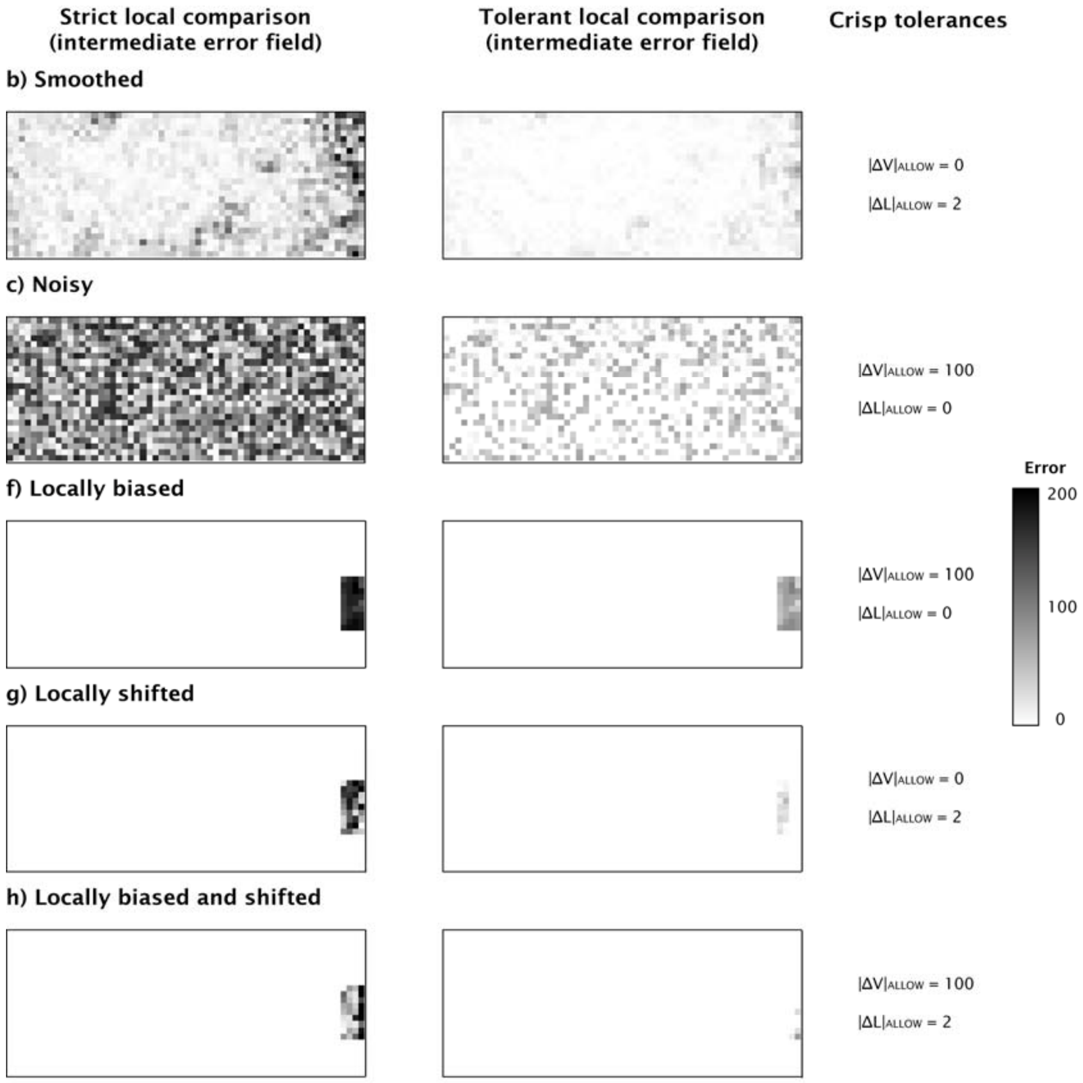


Figure 4.14 The intermediate fields produced during strict and tolerant comparison of the five synthetic fields from Figure 4.1. The effect of applying the crisp tolerances (as specified) can be seen in the right column of fields.

Example f) is very similar to c), but in this case the majority of the field has no error. As with c), if the value tolerance was applied globally it would have ignored the fact that there actually remains an error of 1.7 in the model. Example g) shows the influence of location tolerances during comparison. The 'locally shifted' field is compared while allowing for location difference of 2 elements. Because the deformation actually shifted some elements to the right, these element do not appear as errors in the intermediate field. However, the elements that replaced the shifted elements do appear as errors when compared under tolerance. Example h) allows both value differences of 100 units and location differences of 2 elements to tolerate the deformations in the 'locally biased and shifted' field. This reduces the local errors to almost zero everywhere.

4.4.4.2 Fuzzy tolerance comparisons

The quantitative results of the five comparisons made using fuzzy tolerances are listed in Table 4.6. The associated tolerance definitions and intermediate fields are detailed and shown in Figure 4.15. In this situation, the TMS (i.e. average similarity) value can be compared against the strict similarity measure, which is the percentage correct (PCOR). However, with continuous fields the PCOR generally finds few 'perfectly matching' values and often gives quite low results. The RSQ and KS measures of similarity (i.e. one local, one global measure) are also listed to provide some basis for discussion.

Example b) compares the smoothed field while tolerating a location difference limit of 4 elements and value difference limit of 100 units. The average similarity is 0.82, which suggests high similarity throughout the field. By combining a value and location tolerance together, every element is actually found to be similar to some extent. The least similar element on the right hand side of the field has similarity of 0.24, even though the actual residual for this element is greater than the value tolerance. This is due to a more similar value being found nearby. In the crisp tolerance example, the element is allocated the residual between it and the nearby value. With fuzzy tolerances, the location difference also contributes to the final estimate of similarity. Therefore, even if an identical value were found nearby, the local similarity would be less than one because of the location penalty.

Table 4.6 The comparison measures produced during strict and fuzzy comparisons of the synthetic fields from Figure 4.1.

Field	PCOR	RSQ	KS	TMS
b) Smoothed	0.00	0.93	0.93	0.82
c) Noisy	0.00	0.68	0.81	0.30
f) Locally biased	0.97	0.98	0.98	0.97
g) Locally shifted	0.97	0.98	0.99	0.98
h) Locally biased and shifted	0.97	0.98	0.99	0.99

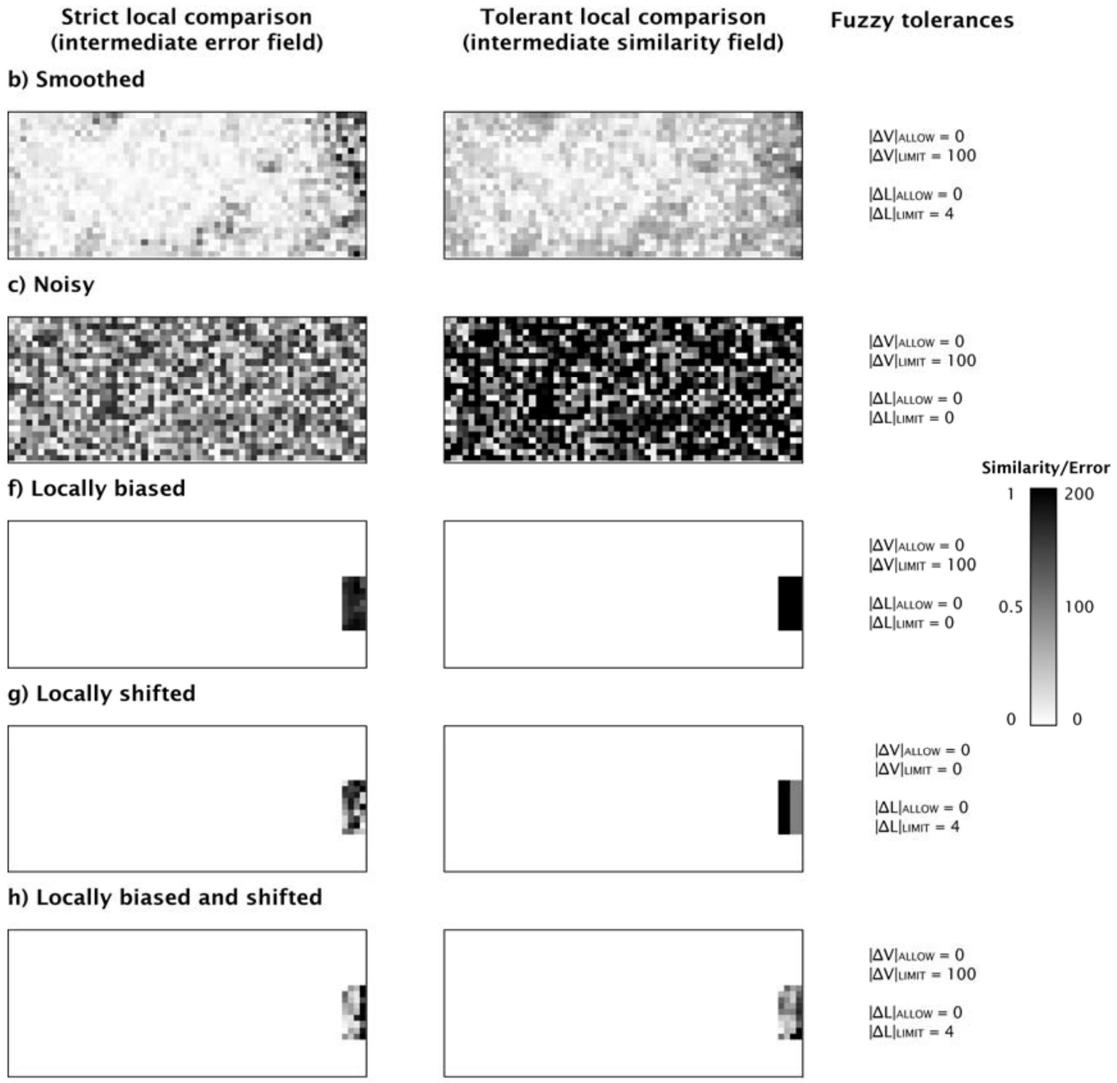


Figure 4.15 The intermediate fields produced during strict and tolerant comparison of the five synthetic fields from Figure 4.1. The effect of applying the fuzzy tolerances (as specified) can be seen in the right column of fields..

Example c) recognises any value difference less than 100 to be similar while comparing the noisy field. Contrasting the RSQ correlation measure with the TMS, it appears that the noisy field represents 68% of the variance in the ‘observed’ synthetic field, yet the elements only have similarity of 0.30 on average (under these tolerances).

In examples f) through h), the influence of tolerances on local deformations can be seen. In contrast to the error field shown in the left column of Figure 4.15, the tolerant field for f) shows that all of the elements in the locally biased region have no similarity. By including a location tolerance in g) finds one part of the local region to have similarity of 0.5, while the rest has no similarity. This is because matching element values are found within the tolerance, but the locations are not coincident and are therefore penalised. Example h) shows that when the tolerances are combined, more extensive searching for similar elements is allowed and the similarity improves.

4.4.5 Summary of tolerance methods

The tolerant comparison method and the approach described for defining tolerances permit the user to consider alternative aspects of spatial field comparison. This allows similarity to be evaluated between elements that are not coincident in time, space or value, based on the user defined rules. In practice, the tolerances can be considered to define the uncertainty for time, location and/or value of the target fields, which controls what each modelled field (or element) is compared against (refer Figure 4.11). In general, tolerances are expected to be applied separately to ensure that their combined effects are not too great. This also helps to make the resulting measure have a clear meaning (e.g. model error when location differences are tolerated).

In hydrological modelling, tolerances should be applied during model assessment to allow the observed fields (i.e. the targets) to represent all possible representations of reality (i.e. to change within the tolerances). This enables the modelled fields (i.e. the sources) to be evaluated against the ‘best possible realisation’ of the observed fields, which thereby makes the results of the comparison identify the ‘best case’ model error (or similarity). Strict local comparisons, as are currently undertaken in hydrology, use only one realisation of the observed field (i.e. assumes target field is strict) and may therefore be including observation errors into the measure of ‘model error’. Tolerances can also allow the user to investigate alternative comparison approaches and their impact (e.g. if timing errors were tolerated, would the optimal parameters change). For data assimilation, tolerances are expected to be useful for ensuring that only the significant local differences in the model are identified. Algorithm 4.4 produces an intermediate field of the tolerant error (or similarity), which represents the significant differences between the model (i.e. the subject of the corrections) and the uncertain observations. This field would be more effective for determining corrections to the model states, as it is less affected by the uncertainty of the observed fields.

4.5 Completeness

Completeness is a theme that should be sought during any comparison. It is highlighted here in response to the inability of any single comparison measure to comprehensively compare all aspects of spatial fields. Instead, completeness during comparison can be achieved by using a range of definitions for importance (e.g. data-driven and knowledge-driven), tolerance (e.g. ranging from strict to specialised) and scale. The different comparison measures (e.g. error and similarity) produced using these definitions are used to evaluate multiple aspects of comparison for model assessment tasks. Multiple 'objective functions' such as these are recognised as being essential for use in model assessment tasks and the methods for performing such assessments are widely researched in hydrology (e.g. Gupta et al. 1998; Boyle et al. 2000; Madsen 2003; McCabe et al. 2005).

The new methods described in this chapter are essential for comparing multiple different aspects of spatial fields. Importance definitions provide a 'focus control' for comparison; tolerance definitions provide a 'strictness control'; and the choice of error or similarity measure controls whether to use a 'penalty or reward' approach. Combining these methods with the current comparison methods allows some different scales to be investigated (e.g. global, local, region-scale), but there is no comprehensive treatment of scale. Therefore, a 'multiscale comparison method' developed for ecological modelling applications (Costanza 1989) has been adapted for use with continuous hydrological fields.

This section details the resources available to the user for achieving completeness during comparison. These resources include: 1) a multiscale comparison method that is developed and explained; and 2) a flowchart identifying how the current and new methods work together to produce comparison measures for different aspects of spatial fields. Using these resources, it is possible to devise a comparison strategy for different hydrological applications. While this section does not explain how to devise this for every possible hydrological situation, it does identify the general objectives of a comparison strategy and suggests how to achieve them.

4.5.1 Comparison of multiple spatial scales

The methods used for comparing spatial fields are most commonly undertaken at the finest scale (i.e. the resolution of the spatial field). However, during visual inspection comparisons are made at a variety of scales. A multiscale comparison fundamentally works by generating upscaled representations of the spatial fields; conducting local comparison for each of the 'upscaled' fields; and summarising the results for each scale. Conducting multiscale comparisons allows analysis of the variation (or persistence) of differences across scale. They also implicitly allow for some differences in the value and location of elements to be tolerated. This occurs because the exact definition of each element is being relaxed via the upscaling process.

The most critical feature of a multiscale comparison is the upscaling method used. Upscaling usually involves the averaging of multiple elements within a larger area to increase spatial support (and can also change the spacing and extent). In hydrology and other disciplines, scaling of specific hydrological measurements to different temporal and spatial scales is actively researched (e.g. Blöschl and Sivapalan 1995; Dungan 2001; Western et al. 2002). For multiscale comparison of specific hydrological fields such as soil moisture, the scaling methods described in Western et al. (2002) could be used to create the upscaled fields.

For more general application, the approach to multiscale comparison suggested by Costanza (1989) can be applied. This method uses a moving analysis window of increasing radius (R) to represent each element at a range of scales specified by the user. As the analysis window radius increases, the spatial support of each element is effectively increased and produces a form of upscaling. A new multiscale algorithm based on this approach is detailed in the following section, followed by an analysis of its use with the synthetic fields from Figure 4.1.

4.5.1.1 The multiscale comparison algorithm

The existing multiscale method of Costanza (1989) cannot be directly implemented for hydrological fields. This method was developed for comparing categorical fields from ecology and uses a specialised global measure for determining the category error between each 'analysis window'. To make this method more applicable for continuous fields, two different global measures for comparing the analysis windows are used – absolute bias (|BIAS|) and Kolmogorov-Smirnov similarity. Any measure could be used here, although the measure should compare the analysis windows as though they are actually 'upscaled element'. Therefore, the spatial arrangement of elements within each upscaled element should be disregarded during these comparisons, which is achieved by using |BIAS| and KS. The method used by Costanza (1989) for defining each analysis window has also been modified to account for the smoothing problems that can occur when treating elements on the edge of the spatial extent. These effects cause biases in the measure produced by Costanza (1989) as a result of treating 'edge elements' less frequently than elements in the centre of the field.

The adapted multiscale comparison method is detailed in Algorithm 4.9. Three different parameters must be set for this method: 1) the scales to be compared; 2) whether to use absolute bias or KS similarity to comparing the analysis windows; and 3) the scale weighting parameter used to combine the performance from all scales into a single measure. These three parameters control how the resulting intermediate and summary measures should be interpreted. During multiscale comparison, the size of the analysis window always begins at a radius of zero (i.e. local comparison) and is incremented until the analysis window covers the whole field (i.e. global comparison). Measures are produced for a range of intermediate scales that can be adjusted using the radius parameter (i.e. to reduce processing time). In practice, some knowledge about the spatial structure of the observed field should be used to parameterise a multiscale comparison. For example, the variogram of a field could inform the user about the maximum scale of organised features.

```
1  Specify maximum radius and radius increment
2  Specify scale weighting parameter K
3  For R = 0 to R = maximum dimension of field step by radius increment
4     For each source element
5        Characterise all source elements within R (mean or distribution)
6        Characterise all target elements within R (mean or distribution)
7        Compare characteristics to produce comparison measure (|BIAS| or KS)
8     Go to next element
9     Determine overall summary measure for this scale (mean |BIAS| or mean KS)
10    Determine weight for this scale = e^(-K * R)
11    Keep running sum of weights
12    Keep running sum of weighted summary measures (i.e. weight * measure)
13 Go to next scale (i.e. R = R + radius increment)
14 Plot each overall comparison measure against scale
15 Multiscale measure = sum weighted summary measures / sum weights
```

Algorithm 4.9 The general algorithm for a multiscale comparison method. The comparison measure used for each scale (lines 5-7) defines how to interpret the resulting measure (i.e. error or similarity). The method produces an intermediate plot (measure versus scale) and a multiscale comparison measure of error (MSME_K) or similarity (MSMS_K).

The decision to use either |BIAS| or KS for comparison determines whether error or similarity is reflected in the intermediate and summary measures. Figure 4.16 illustrates the plots and their general interpretations that result from using each type of measure. With continuous fields, it is common for error to decrease with scale (i.e. from MAE down to |BIAS|) and for similarity to increase with scale (i.e. from typically low PCOR up to higher KS). The mean |BIAS| or mean KS values plotted for the intermediate

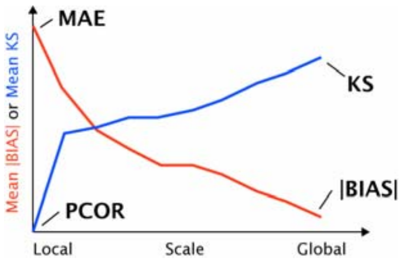


Figure 4.16 A conceptual plot of how the error (red) or similarity (blue) measures can vary with scale. For errors, the plot is bounded by MAE at local scale (R = 0) and absolute BIAS at global scale (R = maximum extent of field). For similarity, the plot is bounded by PCOR at local scale and KS similarity at global scale.

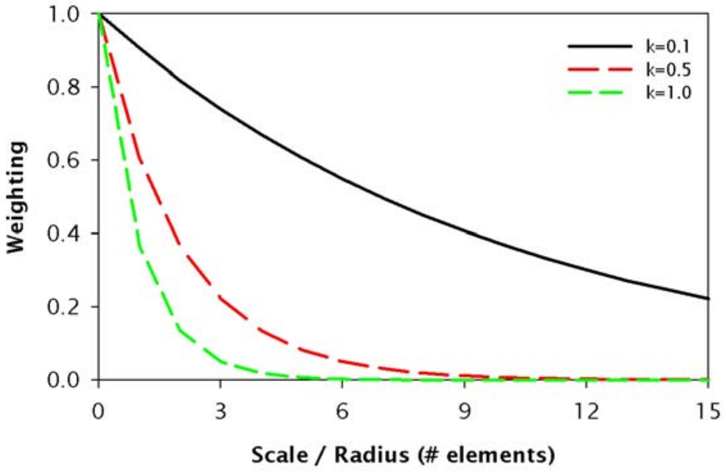


Figure 4.17 The scale weighting parameter (K) controls how influential the comparison measures made at each scale are when producing the multiscale comparison measures ($MSME_K$ and $MSMS_K$).

scales can reveal characteristic scales where the performance improves suddenly. These features may be related to scale differences between the fields or other related characteristics (e.g. noise).

The scale weighting parameter (K) (Algorithm 4.9, line 10) is used to control the influence that measures from each scale have on the summary multiscale measure (e.g. $MSME_K$). This approach uses an exponential decay function to determine the weighting and was originally developed by Costanza (1989). As Figure 4.17 shows, when K = 0 each scale contributes evenly to the summary measure and the multiscale measure is simply the average value across all comparison scales. As the value of K is increased, the contribution of coarser scales is reduced and local scales are more influential in the final measure. For example, at K=0.5 only the scales up to a radius of about 4 play any major role in defining the multiscale measure. This weighting parameter and the specification of scales to use in the comparison combine to determine what the multiscale measure is actually testing.

4.5.1.2 Using the multiscale comparison method

To illustrate the utility of the multiscale comparison method, some synthetic fields from Figure 4.1 are compared to the observed field. The smoothed, noisy and rearranged fields are used to represent situations where the spatial support differs between the fields being compared. The multiscale method is applied using the |BIAS| and KS measures and the intermediate plots (Figure 4.18) and multiscale measures are analysed. The numerical results of the multiscale comparisons are listed in Table 4.7, along with the results of the scale-specific local and global comparisons. All scales up to R = 20 elements (i.e. approximately half the field width) contribute to the resulting measures.

The numerical results in Table 4.7 reveal that the multiscale measures ($MSME_0$ and $MSMS_0$) produce overall summary measures that are between the local and global measures. These have all been produced using K = 0, where all scales are weighted equally. In the top plot in Figure 4.18, mean |BIAS| versus scale is shown and the most prominent changes occur over the first 3 scales in all cases. The noisy field shows this most clearly, with the value reducing from 86 down to 17 after the first scale. This is a logical finding because the increase in scale causes more of the random noise to be averaged out. At larger scales, the noisy and smoothed fields have minimal error. The randomly rearranged field improves in error over the first few scales, again due to the basic smoothing occurring. However, the error found in the rearranged field does not rapidly decrease, thus identifying the persistent errors present across all of the intermediate scales. Table 4.7 identifies that this error reaches 0 at global scale, where all notion of spatial arrangement is removed.

The bottom plot in Figure 4.18 represents mean KS versus scale and appears to be simply a mirror of the top plot, although closer inspection reveals differences. With the similarity measure, the smoothed field is found to be most similar at all scales (apart from global, where the rearranged field is considered identical). The KS measure compares the distribution of values in the upscaled elements, so the noisy field does not get averaged out as much and has lesser similarity. Of interest in this plot is the greater improvement in the performance of the random field once the scale increases. At a scale of 15, the random field is actually considered as similar as the noisy field, because sufficient elements are being compared for the distributions to begin to agree.

Table 4.7 The bounding values and multiscale measures ($MSME_0$ and $MSMS_0$) for the error and similarity-based multiscale comparisons plotted in Figure 4.18.

Field	MAE	$MSME_0$	|BIAS|	PCOR	$MSMS_0$	KS
b) Smoothed	24.7	3.01	0.5	0.00	0.82	0.93
c) Noisy	86.1	10.05	0.0	0.00	0.69	0.81
j) Randomly rearranged	136.3	69.16	0.0	0.00	0.64	1.00

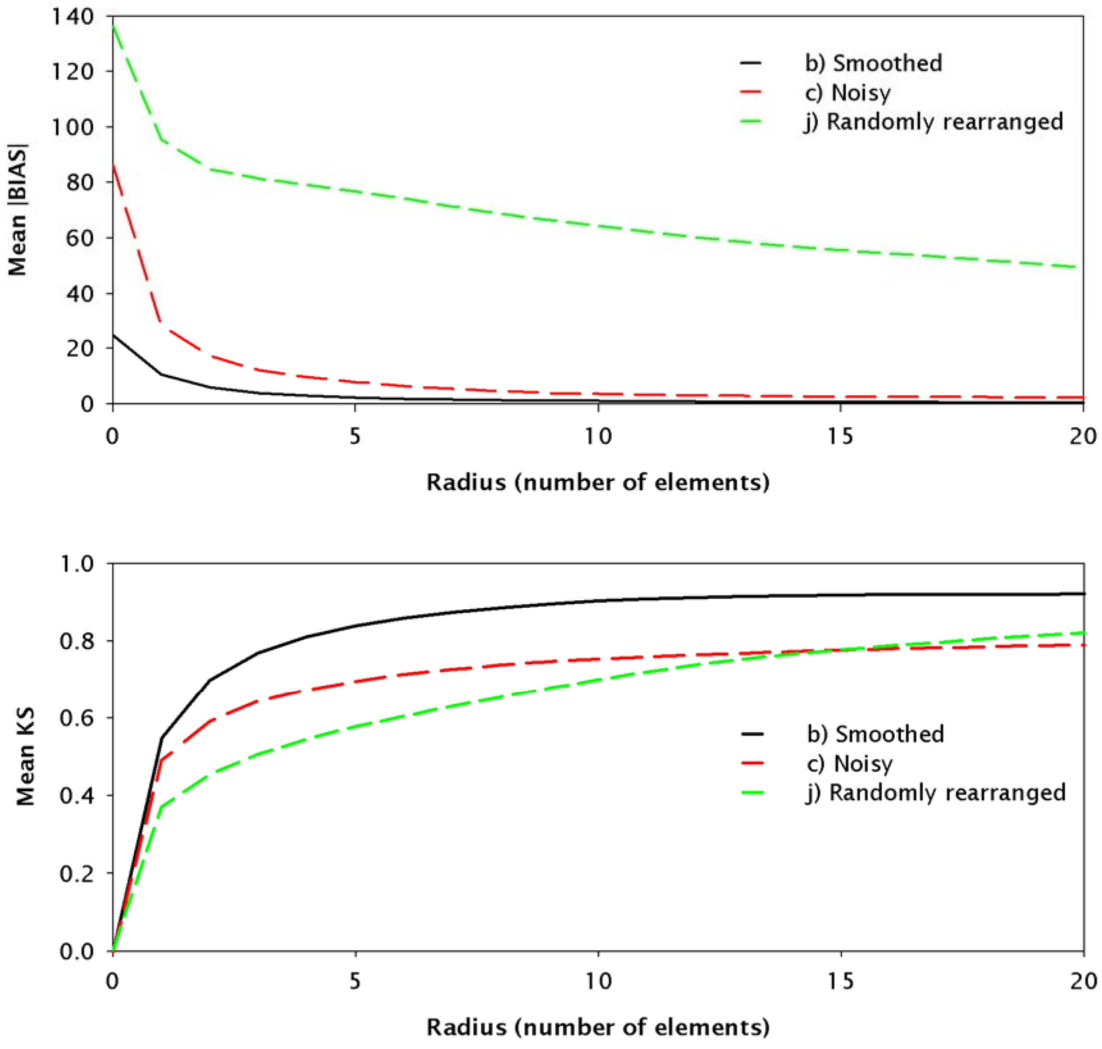


Figure 4.18 Two intermediate plots showing the multiscale comparison measures of error (top) and similarity (bottom) versus scale. The scale ranges from R = 0 (local scale) to R = 20 (approximately half global scale). The full range of scales is not shown so that details in the finer scales are visible. The global scale values (i.e. where the plots finish) are listed in Table 4.7.

An additional use of multiscale comparisons, for diagnostic purposes during analysis, is illustrated in Figure 4.19. In this example, the smoothed and noisy fields are compared using multiscale comparison but with a slightly different approach – only one of the fields is upscaled while the other is kept at its finest scale. This is used to discover if better agreement exists between the fields when represented at different scales. The plot shows that when only the smoothed field is upscaled, the error increases consistently and there is no coarser scale with a lower error. If only the observed field is upscaled, an obvious improvement in error occurs at R = 2 elements. This indicates that the observed field has a scale difference from the smoothed field (i.e. it contains finer scale features). When both fields are upscaled, the effect of the scale mismatch is still visible in the plot but is less apparent. With the plot for the noisy field, when it is upscaled there is an improvement in error, but it produces a less obvious minima than seen with the smoothed field. This provides evidence that upscaling would make the fields more similar, but the gradual increase in error at the coarser scales suggests this is not caused by a clear scale difference.

The multiscale measures are most useful when the intermediate plot of ‘measure versus scale’ can be visually analysed to detect characteristic scales. When analysis of the plot cannot be undertaken (e.g. during an automated process), analysis can be based on a multiscale measure that summarises the key scales of interest, which are usually the first

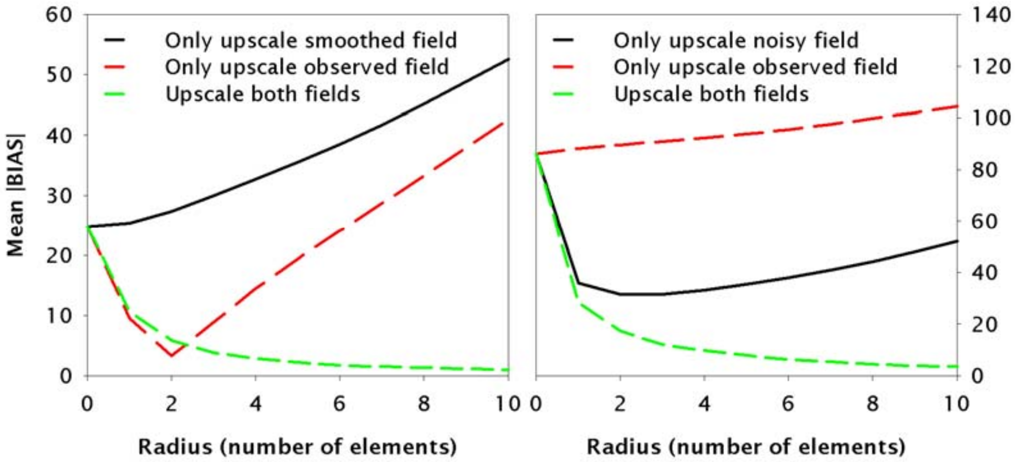


Figure 4.19 Two intermediate plots showing the mean |BIAS| measure versus scale for comparisons of the smoothed (left) and noisy (right) fields (Figure 4.1). These plots are used for diagnosing scale differences between fields. The result of upscaling both fields is shown as a reference.

few scales. Once the field has been upscaled to R > 5, the results are usually quite regular and of limited additional interest over a global scale comparison. These measures are particularly suited to situations where a single comparison measure is required (e.g. when multi-criteria assessment is unavailable). The multiscale measure condenses comparisons of multiple different scales together and thereby implicitly tolerates differences between the fields (via the upscaling). This produces a single numerical measure that is indeed useful, but it does not replace the richer information available in the individual contributing comparisons.

4.5.2 Combining methods in the comparison flowchart

As there is no single method capable of completely comparing spatial fields, the combination of current and new methods is used to achieve completeness during comparison. The widely-used current methods provide measures of strict error and strict similarity at local and global scales, but when the new pre-processing and comparison methods are added to these current abilities, an extensive suite of methods for comparing many aspects of spatial error or similarity are created. The comparison flowchart presented in Figure 4.20 shows how these combinations occur and is supplemented by information in Table 4.8. The comparison flowchart begins with modelled and observed spatial fields; applies pre-processing methods to them; uses a range of different methods to produce a range of comparison measures; and combines these measures into an overall model assessment. The final step is not the focus of this thesis; the determination of the suitable comparison measures is.

In this flowchart, the shaded boxes represent where the new pre-processing or comparison methods have added a branch to the flowchart. The bottom row of the flowchart shows 14 groups of comparison measures that are made available by combining the methods (Table 4.8). Of these 14, only 3 groups can be achieved with the methods that are currently used in hydrology. The KS measure creates the group of global-scale similarity; the new multiscale measure directly creates 2 new groups; introducing tolerances directly creates 2 new groups; and introducing importance (via regions) contributes to 6 new groups. This tally shows that by combining existing methods with some new, versatile methods, many more aspects of comparison can be measured and at a number of new scales.

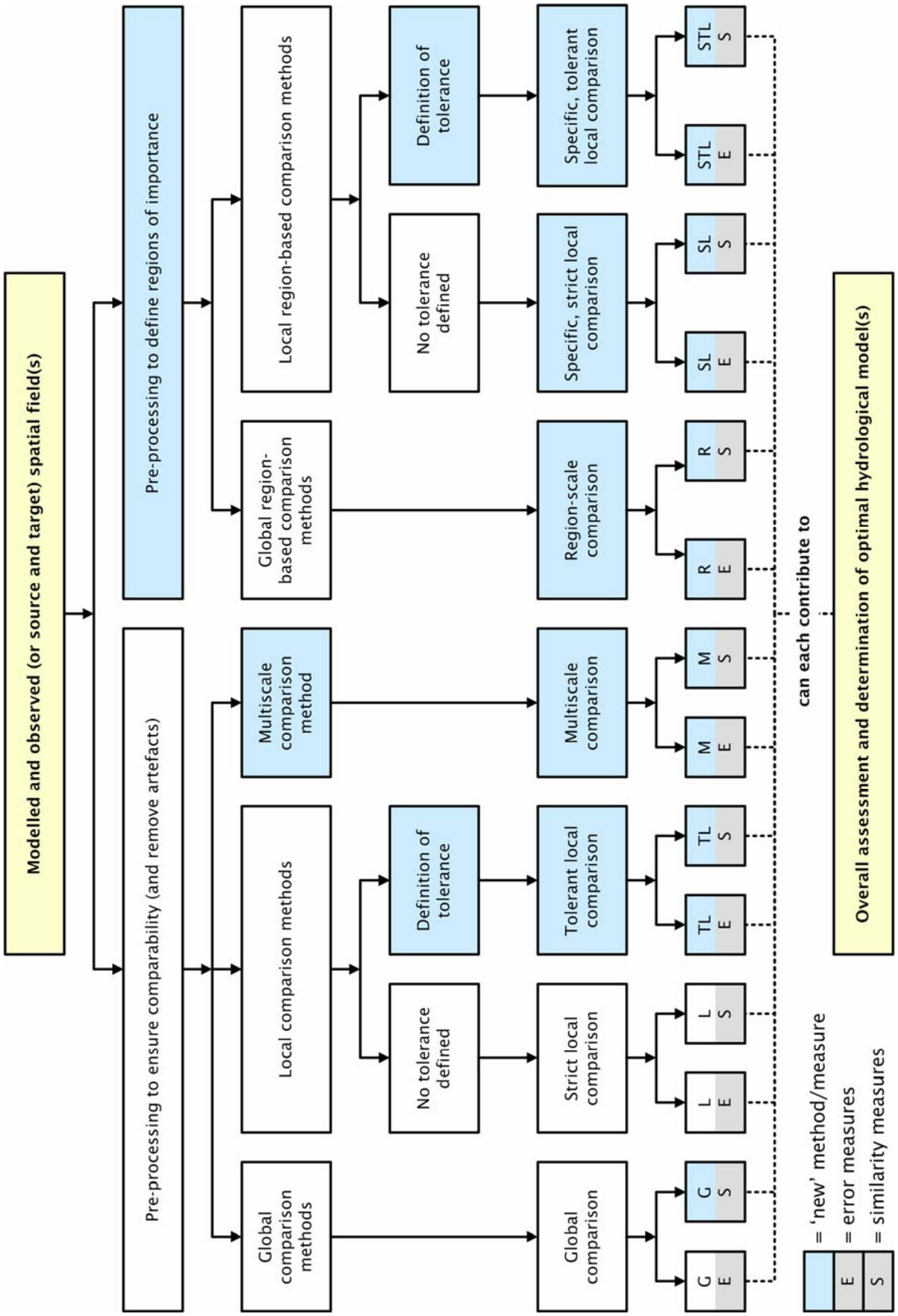

Modelled and observed (or source and target) spatial field(s)
Pre-processing to ensure comparability (and remove artefacts)
Pre-processing to define regions of importance
Global comparison methods
Local comparison methods
Multiscale comparison method
Global region-based comparison methods
Local region-based comparison methods
No tolerance defined
Definition of tolerance
No tolerance defined
Definition of tolerance
Global comparison
Strict local comparison
Tolerant local comparison
Multiscale comparison
Region-scale comparison
Specific, strict local comparison
Specific, tolerant local comparison
G E
G S
L E
L S
TL E
TL S
M E
M S
R E
R S
SL E
SL S
STL E
STL S
can each contribute to
Overall assessment and determination of optimal hydrological model(s)
= 'new' method/measure
E = error measures
S = similarity measures

Figure 4.20 The comparison flowchart shows how the current and new comparison methods interact. The acronyms (bottom row) refer to the general groups of comparison measures available (Table 4.8).

Table 4.8 Details of the acronyms listed in the comparison flowchart Figure 4.20. Each acronym refers to a general group of comparison measures, within which there may be one (or more) particular measures that have been discussed in this thesis. Each group of measures has a particular scale of application.

Group of comparison measures	Quantitative comparison measures discussed in this thesis	Scale of application
G E – Global error	BIAS |BIAS|	Global
G S – Global similarity	KS	Global
L E – Local error	RMSE MAE	Local
L S – Local similarity	RSQ COE PCOR	Local
TL E – Tolerant local error	TMAE	Local to intermediate
TL S – Tolerant local similarity	TCOE TMS TSE	Local to intermediate
M E – Multiscale error	MSME_K	All
M S – Multiscale similarity	MSMS_K	All
R E – Region-based error	BIAS per region	Intermediate
R S – Region-based similarity	KS per region	Intermediate
SL E – Specific local error	RMSE for specific region MAE for specific region	Specific local
SL S – Specific local similarity	RSQ for specific region COE for specific region	Specific local
STL E – Specific tolerant local error	TMAE for specific region	Specific local to intermediate
STL S – Specific tolerant local similarity	TCOE for specific region TMS for specific region TSE for specific region	Specific local to intermediate

The flowchart provides a clear map of the options for comparison. It can be used to navigate down through the steps of processing to determine what comparison measures are available for use in model assessment. It can also be used to navigate up from a particular comparison measure to see what decisions or steps are needed to calculate it. This is particularly useful when the measure is unfamiliar and the comparison method requires user parameterisation.

4.5.3 Comparison strategy for hydrological modelling

The flowchart shows the full range of comparison options available. Many of the new methods are dependent on user knowledge to parameterise and direct them. There is a great diversity of measures represented in this chart and the associated table, although they are not all required in every situation. For example, when the spatial arrangements in the fields are deemed unimportant, using global measures of error and similarity are probably sufficient.

When such diversity of measures is presented, the difficult question arises – "What comparison measure(s) should I use for assessing my hydrological model?" The simple and non-committal answer is – "The suitable comparison measure(s) depend on the specific application and the purpose of the assessment." While this is true, the following section is provided to give some more detailed guidance.

Based on the review of comparison methods (both qualitative and quantitative) and the measures made available via developments in this chapter, there are some major objectives that should be achieved during comparison. The user is responsible for selecting the comparison measures to use for assessments and these measures should: 1) make comparisons across a range of scales; 2) measure only the significant errors and significant similarity between elements; and 3) make comparisons that answer specific questions about model performance.

4.5.3.1 Objective 1: cover a range of scales

Continuous spatial fields contain local details, intermediate features and global characteristics that are all important to compare. Current methods are applied at the extremes of scales and fail to recognise similarity between intermediate scale features, mainly because these intermediate scales for comparison are poorly defined. A comparison strategy should aim to make comparisons at the extreme scales and also at one or more intermediate scales.

Intermediate scale comparison measures can be calculated using region-based comparison methods (where there are specific features of interest) or tolerant local comparisons with a location tolerance specified (where there are scale differences of known uncertainties). An intensive comparison of all scales can also be achieved using the multiscale methods.

4.5.3.2 Objective 2: measure both significant errors and similarity

The fundamental approaches to comparison used by error and similarity measures are different and they have quite different interpretations. At least one measure of each type should be used. For these measures, the user should make efforts to ensure that only the differences (or similarities) that are significant (for the purposes of the comparison) are being considered.

Comparison measures of error and similarity are available for all scales of comparison. Controlling these measures to only use significant differences (or similarities) can be achieved using tolerant methods (at local scale) or via pre-processing.

4.5.3.3 Objective 3: make specific tests of model performance

When all elements in the fields are compared, the resulting comparison measure is obviously a response to the error in all elements. In many applications, specific elements are more important than others. Using user knowledge to identify these important elements or regions is beneficial and can be used to test specific parts of the fields. The user should incorporate any useful knowledge into the comparison by defining regions of importance. Their choice depends on the specific objectives and uses of the model. These specific tests are then undertaken using either region-based or region-specific local comparison measures.

A range of different importance definitions should be used if they can be defined. For example, only saturated areas (i.e. those above a threshold) may be compared for assessing a model of flood response. In other situations, both the value and shape can be considered to identify importance. For example, high rainfall areas that are elongated (linear features) may be used when assessing frontal precipitation, while circular regions may identify convective precipitation. These tests will always depend on user knowledge of the phenomena being assessed, although there are a number of ‘threshold processes’ in hydrology that lend themselves to this type of approach.

4.5.3.4 Summary of comparison strategy

By selecting comparison measures that address each of these objectives, the completeness of the overall comparison can be established. Ultimately, the onus remains with the user to select the measures and parameters that are used for assessment (as is current practice). It is expected that by providing new concepts and measures that encourage greater user input, users will give more thought to comparisons and consequently make more rigorous and insightful assessments. This strategy and the decisions required in devising it are illustrated in the following chapter for use in a real hydrological modelling situation.

4.6 Chapter summary

Throughout this chapter, the three major themes from the review – importance, tolerance and completeness – have been developed into comparison methods that can be used with hydrological spatial fields. These new comparison methods have each been tested using synthetic fields. The response seen in each of the measures (to the deformations introduced) was quite clear and matched the expected performance. These measures can now confidently be applied in 'less controlled' situations (e.g. for comparison of modelled fields with in-situ observed fields) to evaluate their utility for making specific hydrological comparisons.

The combination of the new methods with the current methods has expanded the range of available comparison measures greatly. These measures can be used to compare many diverse aspects of spatial fields, including the representation of intermediate scale features (i.e. regions) and spatial relationships between elements (i.e. by using tolerances). Such aspects were not able to be quantitatively compared prior to developing these methods.

The diversity of the comparison measures made available throughout this chapter introduces greater flexibility to model assessments using spatial field comparison. A general strategy for choosing suitable comparison measures is presented, but the details of the strategy are expected to vary with application and purpose. The following chapter devises a comparison strategy that uses these comparison measures to evaluate a real hydrological modelling situation.

Chapter 5

Hydrological application of quantitative comparison methods

5.1 Chapter overview

The assessment of spatially-distributed hydrological models can involve a range of processes including calibration, optimisation, testing and understanding. All of these processes can utilise the results of spatial field comparisons, particularly when the number of models being assessed preludes use of visual analysis. The current comparison measures used in hydrology compare only a limited number of aspects of spatial fields, but new methods introduced in the previous chapter enable a greater range of spatial field aspects to be compared (e.g. the comparison flowchart). The goal in this chapter is to evaluate the use of these comparison measures for hydrological model assessment.

The new comparison measures are applied to an example data set from hydrological modelling and their performance (for model assessment) is evaluated. There is no 'correct answer' to model assessment – it is an entirely subjective process. However, a model assessment conducted by experts (i.e. those familiar with the model and data) does provide a useful benchmark against which an automated method can be evaluated. This chapter introduces the distributed soil moisture modelling work undertaken at Tarrawarra (Western and Grayson 2000), for which an existing 'expert assessment' using spatial fields is available. Based on the approach used by the experts and the comparison measures available, a strategy is devised for conducting the comparison (i.e. the measures and parameters to use).

The comparison strategy is applied to produce a range of measures, which form the basis of the model assessment processes. These measures are then analysed to: 1) recognise the optimally performing model(s) (or parameters); and 2) to identify where/when the model performs well (or badly). The first type of analysis simulates what is undertaken during calibration or optimisation. The second is used to assist with model understanding, which is usually only possible with visual inspection of the spatial fields. The analysis results from the assessment are evaluated using the expert assessment and visual inspection of the fields. The chapter concludes with a discussion about the relative merits of each group of comparison measures.

5.2 Hydrological modelling example

In this chapter, a comparison strategy is devised and applied for assessing the modelled spatial fields from a real hydrological modelling situation. Due to the absence of modelling or field observation in this thesis, an example has been chosen that is considered representative of many different hydrological modelling situations. The example chosen – the Tarrawarra project – has a variety of modelled spatial fields and independently produced observed spatial fields (i.e. not just one model and one observation); a temporal extent covering a range of climatic conditions; a spatial extent that includes significant topographic variation; and observed spatial fields that contain both random and organised features (i.e. not too simple for modelling).

The Tarrawarra project (Western et al. 1999b) focuses on the 10.5 hectare Tarrawarra catchment located in the Yarra Valley, approximately 60 km east of Melbourne, Australia (Figure 5.1). This catchment has been the subject extensive research and review in the past decade and is well known in the hydrological community (Western et al. 1998; Western et al. 1999b; Western and Grayson 2000; Western et al. 2004). The catchment is mildly undulating, covered in pasture and features two southerly facing drainage lines, but these are not perennial streams.

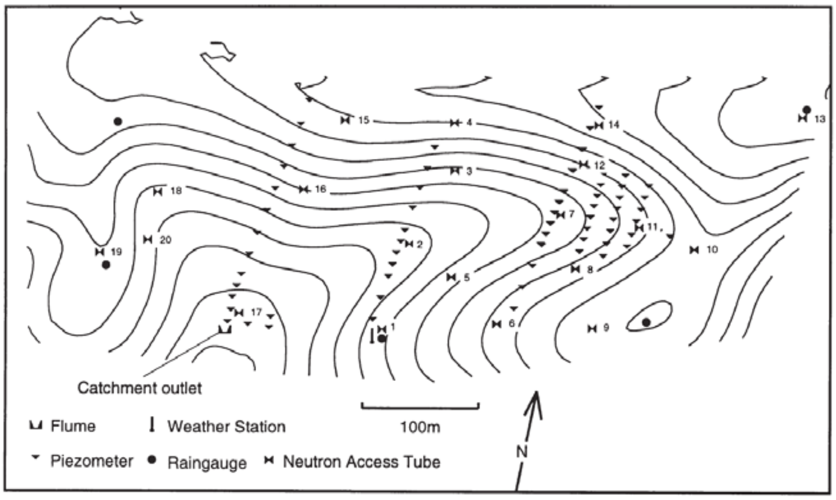


Figure 5.1 The Tarrawarra catchment, showing topographic features and gauging equipment (from Western and Grayson 2000).

In the original research by Western et al. (1999b) they investigated the spatial organisation of hydrological processes, particularly soil moisture. Soil moisture is a key controlling variable in the hydrological cycle and also a key state variable in many hydrological models. It is highly variable in both space and time and is influenced by many factors (e.g. topography, climate, land use), which together produce complex spatial arrangements of soil moisture.

Fully-distributed hydrological modelling has been undertaken at Tarrawarra by Western and Grayson (2000) using the Thales model (Grayson et al. 1995). Ten parameterisations were used to produce modelled spatiotemporal series (with daily time step) for a 14 month period (September 1995 to November 1996). Within this period, observed fields of 'daily' soil moisture were collected (using in-situ measurement) on 13 occasions and provide the 'reality' against which the modelling is assessed in this chapter. In the original work, Western and Grayson (2000) used additional data for model assessment, but only the spatial fields are used here.

5.2.1 Observed fields at Tarrawarra

The spatial observations made at Tarrawarra were produced from regularly-spaced measurements (on a 10m x 20m grid) made using time domain reflectometry probes to measure volumetric soil moisture (% v/v) in the top 30cm of soil. These were collected on 13 occasions over the 14 month period and cover different seasons and catchment conditions (dry, wetting up, wet, drying down). Each measured moisture value had a reported measurement error of σ=1.7% v/v, while the standard deviation of values across the catchment ranged from 2.3% v/v (in dry conditions) to 4.9% v/v (in wet conditions).

TDR measurements are essentially point measurements (i.e. small support) and are more likely to be affected by small scale variability than an integrated measure (e.g. remote sensing derived soil moisture pixel value) would be. Western et al. (1999b) note that the difference between a measurement made for each 10m x 20m area (as opposed to a point representing that area) would cause a minor reduction in the overall variance of the observed field (i.e. slightly smoother appearance). Therefore, the point measurements are assumed to represent the larger support of 10m x 20m. This is a convenient assumption so that the measurements can be represented as a regular-grid element spatial field, as is common in hydrology.

The observed fields are shown in Figure 5.2 and highlight the different spatial arrangements of soil moisture that occur throughout the year. This data illustrates the hypothesis put forward by Grayson et al. (1997) regarding the behaviour of soil moisture in temperate catchments. They found that the spatial arrangement varies between two 'preferred states', each dominated by different hydrological fluxes. During summer months, the vertical fluxes dominate, thus limiting the lateral redistribution of water. In winter, these lateral and surface fluxes are more dominant and produce a 'topographically controlled' arrangement, with wet drainage lines and drier slopes (due to evapotranspiration).

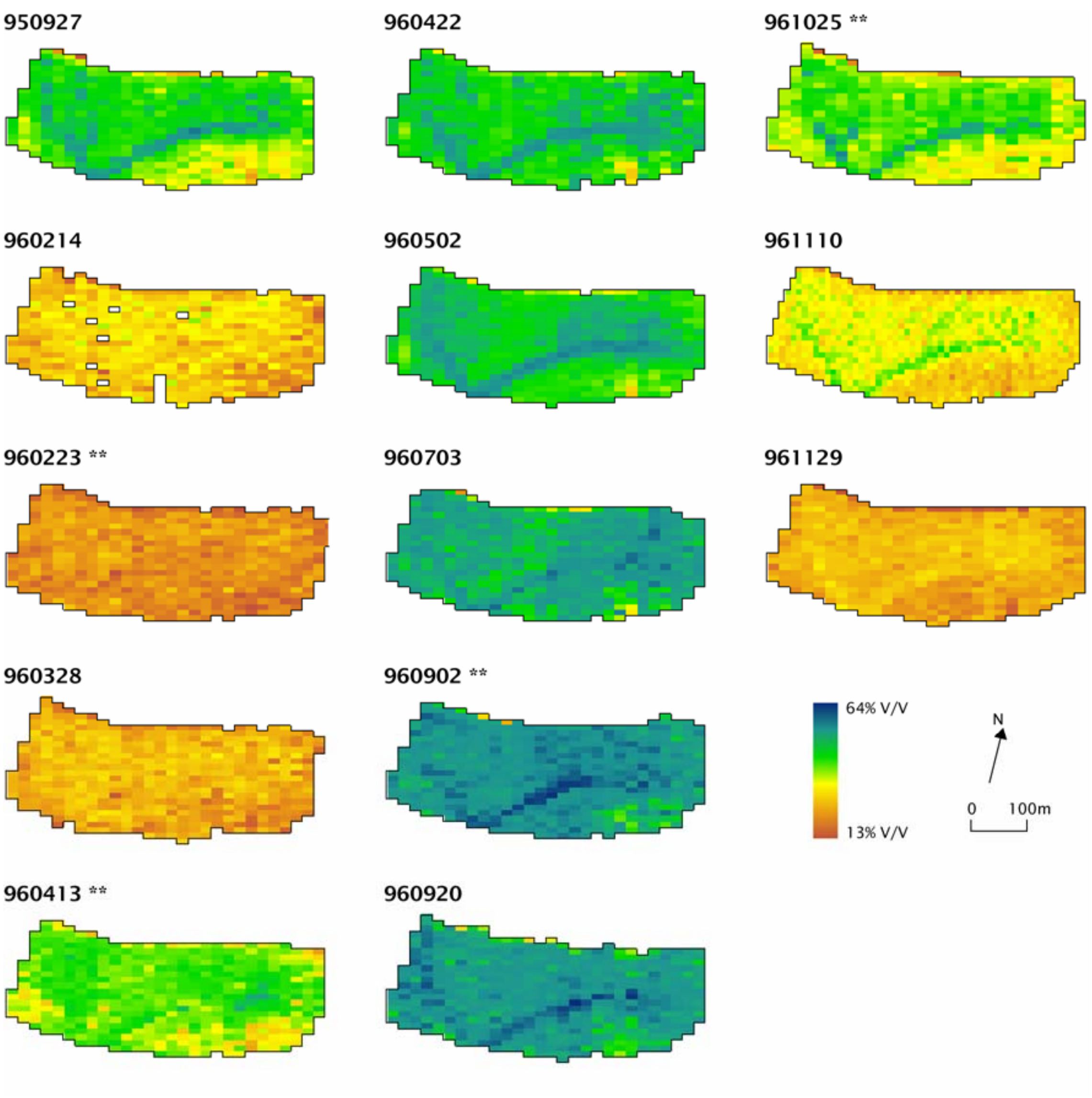


Figure 5.2 The spatial fields observed during the Tarrawarra project. Each field is labelled by the date of observation (in YYMMDD format) and all fields are represented in the same colour scheme (using dry to wet colours). The driest months are November through March (i.e. in the southern hemisphere). The four fields marked with asterisks are used throughout this chapter and are representative of the variation seen throughout the year in the observed fields.

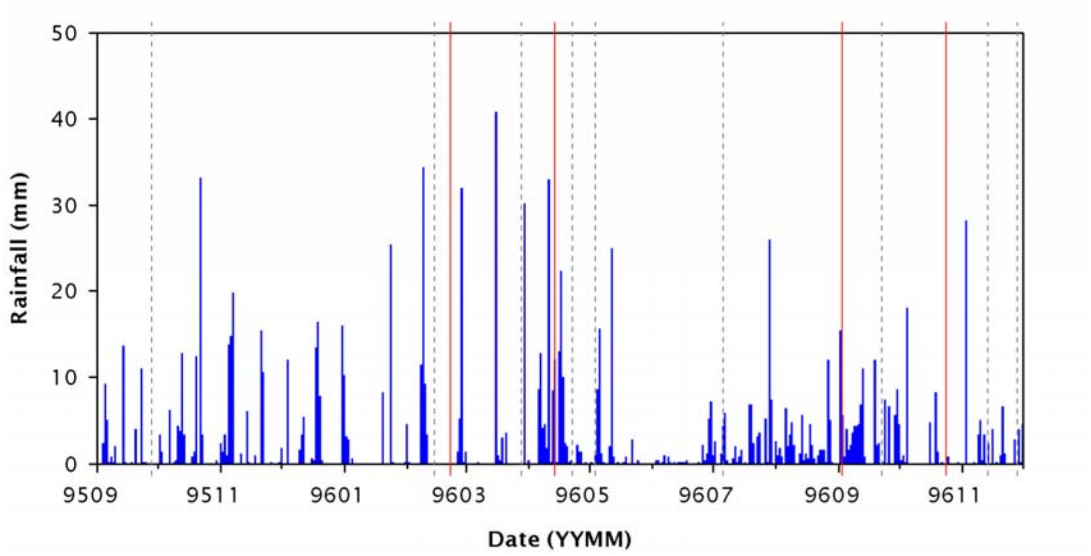


Figure 5.3 The daily rainfall record during the Tarrawarra project. The occasions on which spatial fields were observed are shown with solid vertical lines (grey). The dates for the fields selected for use in this chapter are shown with solid vertical lines (red).

Inspection of the rainfall record for the period of observation (Figure 5.3) and the spatial fields reveals four particularly interesting hydrological situations that a spatial model should correctly represent. The fields on these occasions show different levels of organisation. These four occasions are used in this chapter for model assessment so that the quantity of analysis is not overwhelming for the reader:

- **960223** Observed when there has been no rain for 12 days and occasional storms over past months. This field represents the catchment in a ‘dry’ state (i.e. the end of summer) and has a uniform appearance with minor random variations.
- **960413** Observed during a period of autumn storms that have come after a particularly dry summer. This field represents the catchment ‘wetting up’. Some organisation of drier and wetter regions is apparent.
- **960902** Observed at the end of winter when there has been extensive rainfall. This field represents the catchment in a ‘wet’ state. Clear organisation and some general features are apparent.
- **961025** Observed at the start of spring and there has been no rain for 4 days. This field represents the catchment ‘drying down’. Very organised field, with drainage lines being particularly apparent and some aspect effects.

5.2.2 Distributed modelling at Tarrawarra

The distributed modelling for Tarrawarra was undertaken with the Thales model (Grayson et al. 1995). Thales operates on a computational network based on streamlines and elevation contours (i.e. irregular elements). Thales produces an estimate of the soil moisture stored at different depths for each irregular model element (i.e. the value is an areal average) for each day (i.e. the timestep used in this application) during the modelling period. The irregular element fields were converted into regular element fields to ensure the field structure was directly comparable with the observed spatial fields detailed in the previous section. There is a scale inconsistency between the modelled (areal average) values and the observed (point sample) values that is discussed and managed throughout the comparison process.

The application of the model at Tarrawarra used rainfall and potential evapotranspiration as the forcing data. The model incorporated the following hydrological processes: saturated subsurface lateral flow; saturation excess overland flow; exfiltration of soil water; runon infiltration of overland flow; deep seepage; and evapotranspiration. These processes were included because field observations suggested they may be present. The soil profile used consists of two layers, of which only the top layer is laterally transmissive. While limited detail of the model algorithm is provided here, full details of the differences between model parameterisations are given. For complete details of the modelling algorithm, the reader is referred to Western and Grayson (2000).

In the version of Thales used to produce the modelled fields, a number of spatially-variable model parameters were set for each model element to define the soil properties and soil/plant interactions. These include: saturated hydraulic conductivity (K_{SAT}); deep seepage (K_{DEEP}); total soil depth (D_{SOIL}); and depth of laterally transmissive upper soil layer (D_{UPP}). A global parameter defining the presence/absence of the spatially-variable evapotranspiration process (SVET) was also defined. SVET is influenced by the solar radiation falling on the terrain, which is controlled by slope and aspect. Ten different model parameter sets (Table 5.1) were used to produce ten different model runs (or realisations) for the Tarrawarra catchment. Unless otherwise specified, the spatially-variable parameters are set as constant for all elements within the field.

The ten different model runs each represent a different combination of hydrological processes (e.g. evapotranspiration, deep seepage) and catchment states (e.g. soil parameters). These parameter changes combine to produce modelled fields having different spatial arrangements of soil moisture values. Based on previous analyses by Western and Grayson (2000), models runs 1, 3 and 4 are found to have very minimal differences from the other runs. Therefore, they are not used any further in this chapter. This reduces the amount of data being analysed here down to four different occasions (representing vastly different spatial arrangements) on which there are seven (rather than ten) modelled fields and one observed field (Figure 5.4).

Table 5.1 The ten model runs produced for Tarrawarra using different combinations of model parameters for Thales (adapted from Western and Grayson 2000). Of these ten runs, seven runs (shaded) are used in this chapter. Western and Grayson (2000) found run 5 to be the 'best' model, although the differences between models were noted as 'quite subtle'.

Run	**K_{SAT} (mm/h)**	**K_{DEEP} (mm/h)**	**D_{SOIL} (mm)**	**D_{UPP} (mm)**	**SVET**	**Description of parameter changes**
1	20	0	600	400	Off	Default parameters are set from available field measurements
2	40	0	600	400	Off	Doubled K_{SAT} to increase rate of water flow through soil
3	60	0	600	400	Off	Tripled K_{SAT} to further increase rate of water flow through soil
4	20	0	600	400	On	Same as run 1, but with SVET enabled
5	40	0	600	400	On	Same as run 2, but with SVET enabled
6	60	0	600	400	On	Same as run 3, but with SVET enabled
7	40	0	Dependent on terrain position	400	On	Soil depth varies spatially and is distributed by using a regression relationship between the limited soil depth observations and terrain attributes
8	40	0	600	1) 400 2) 200	On	Depth of upper soil layer varies spatially and is distributed by using a field of soil types containing: 1) duplex; and 2) gradational soils (Figure 5.5)
9	26.7	0	600	600	On	Single, laterally transmissive soil layer is used (rather than two layer model) and K_{SAT} is adjusted to be equivalent with run 5
10	40	0.013	600	400	On	Deep seepage (i.e. loss of moisture to groundwater) is introduced and calibrated so that overall catchment runoff is correct

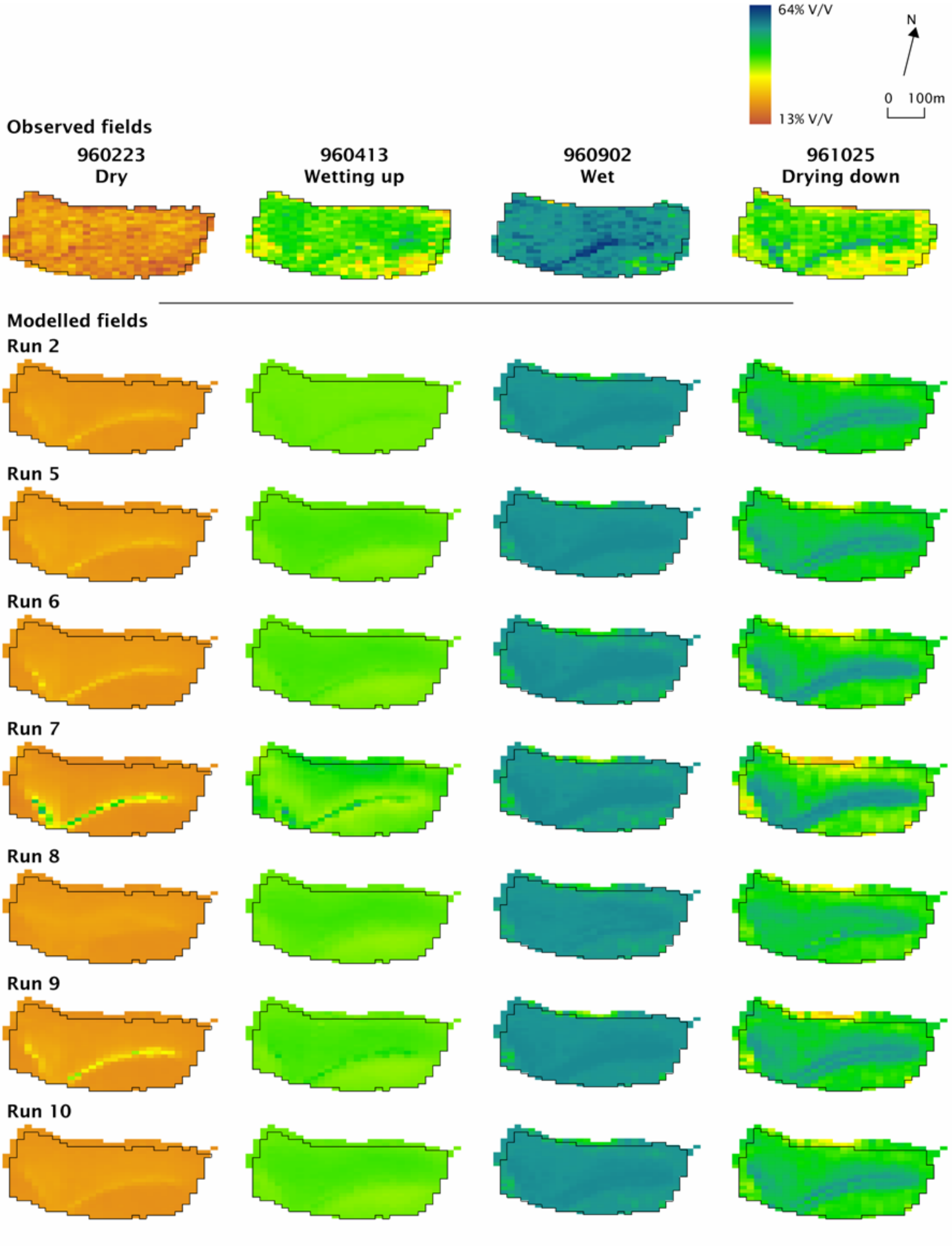


Figure 5.4 The observed fields and the temporally-coincident modelled field from the seven different spatiotemporal series. This figure can be used by the reader for visual comparison and assessment of model performance. It is also useful for visually confirming the error or similarity that is measured by different comparison measures. The outline overlain on each field shows the comparison extent for each occasion, based on the intersection of the modelled and observed fields (which varies slightly because of the varying extent of the observed fields on each occasion).

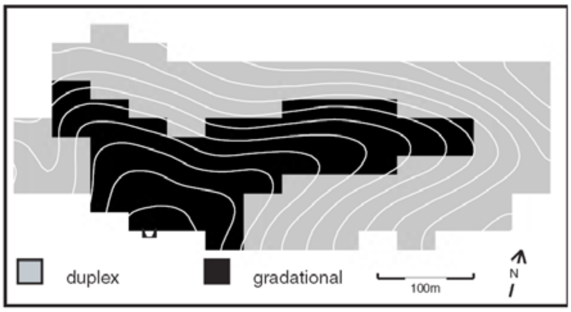


Figure 5.5 The spatial field of soil types used to define the depth of the upper soil layer in model run 8. Gradational soils are found in the gullies and duplex soils are observed to occur in all other places in the model extent (from Western and Grayson 2000).

5.2.3 An expert model assessment for Tarrawarra

In Western and Grayson (2000), the hydrological model that is used in this chapter was assessed using qualitative visual comparison of the fields (Figure 5.4), visual analyses of the intermediate error fields and basic local comparison measures (BIAS, RMSE). This assessment is considered to be an expert opinion because it was undertaken by the developers of both the modelled and observed spatial fields.

The detailed findings of Western and Grayson (2000) are presented here and subsequently used for two purposes in this chapter: 1) to identify the types of comparisons that experts would make in this situation (i.e. a comparison strategy); and 2) to define the 'expert findings' about model performance that may also be revealed by the quantitative model assessments undertaken in this chapter.

Western and Grayson (2000) made a number of notes when comparing the modelled and observed spatial fields at Tarrawarra (often referred to as spatial patterns). Overall, they judged model run 5 to be the 'best' model, but also noted that the "differences between the runs were often subtle and no one run was the best on every occasion" (Western and Grayson 2000, p.228).

They then proceeded to compare model run 5 to the observed fields for different dates. They found that model run 5 "performs well during dry and wet periods" but "during the transition periods in autumn (wetting up) and spring (drying down) there are some differences between the simulated and observed patterns" (Western and Grayson 2000, p.234). More specifically, Western and Grayson (2000, p.234) make the following notes:

- On the dry occasion (960223), the observed field was "well simulated in terms of the average moisture and the spatial pattern".
- On the wetting up occasion (960413), "the model predicted the average soil moisture well," but there was no "evidence of the significant lateral redistribution" seen in the observed patterns.
- On the wet occasion (960902), "extensive areas are saturated" and "the soil moisture pattern is well predicted".
- On the drying down occasion (961025), "the model performs relatively poorly" with "the effect of lateral redistribution over predicted" and the existence of "strong aspect bias in the soil moisture errors".

After comparing the fields from model run 5 to the observed fields, Western and Grayson compared the other model runs to run 5 (i.e. model to model) to identify the spatial changes caused by the different model parameters (i.e. a controlled comparison) (Western and Grayson 2000, p.235-238). Any differences that were noted about the temporal dynamics of the models are not listed because they would require spatiotemporal comparison methods (i.e. comparing two model series), which is a challenge that remains for future research. The differences that were noted from comparing the modelled spatial fields are listed here.

- Run 2 used double the initial value for K_{SAT} to "most closely represent the effects of lateral flow" in the observed field. Because the SVET process was not active the run fails to represent the "strong aspect effect [that] was apparent in the observed data" on the wetting up, wet and drying down occasions.
- Run 6 had "little difference compared with run 5" on the wetting up occasion, but the upper parts of the hillslopes were drier and the drainage lines wetter in both the wet and drying down fields.

- Run 7 had a more topographically controlled arrangement compared with run 5 and this "generally led to similar or poorer simulations of the [observed] patterns." In the wetting up field "the hilltops were too wet" and in the wet field they "were too dry".
- Run 8 contained "an [unrealistic] band of high soil moisture at the soil unit boundary" caused by the discontinuity of lateral flow between the gradational and duplex soil types (Figure 5.5).
- Run 9 was more influenced by lateral flow on the dry occasion than run 5 and this factor made it inconsistent with the observed field.
- Run 10 exhibited slightly lower soil moisture (than run 5) in the drainage lines on the wetting up and drying down occasions and on the ridge tops on the wet occasion.

These comparisons were made by experts using both visual and basic quantitative methods. They compare different aspects of the hydrological fields – average moisture content; the representation of important hydrological features (e.g. hilltops, drainage lines); differences between terrain features (e.g. steep slopes, north/south facing slopes); and localised regions of large errors (e.g. unrealistic band of high soil moisture at soil unit boundary) – and compare at a range of scales (global, local and intermediate).

They also note both similarities and differences, making it a complete comparison according to the definition in Chapter 4. This comparison relies heavily on linguistic descriptions (e.g. "generally similar", "relatively poorly") of comparison, which are difficult to relate to real quantities and cannot be replicated by another set of experts. Therefore, any quantitative comparison measures evaluated against these expert opinions are done so with these factors in mind. Ideally, a single dataset that was assessed by multiple experts would have been used to evaluate the automated methods.

5.2.4 Emulating automated model assessment

This chapter undertakes the different processes of model assessment in a manner that emulates automated model assessment. When automated procedures (e.g. calibration, optimisation) are used in hydrological modelling, they require the observed (or reference) field and other parameters (e.g. performance criteria) to be defined prior to running the procedure. This requires the observed field to be well-understood, so it must be available for analysis prior to use (e.g. visually, quantitatively). Once the automated procedure is parameterised, it can be run on any number (i.e. often hundreds) of models. This overall process does not permit the visual inspection of the modelled fields at any stage. Instead, each model is represented only by the results of the automated procedure.

During this chapter, visual inspection of the modelled fields (i.e. those being compared to the reality) is not used for model assessment. When discussing and evaluating the results of each comparison measure, the modelled fields are used to visually confirm and evaluate the measures. The reader is encouraged to make additional visual analyses (using Figure 5.4) to evaluate whether the results from the various comparison measures are indeed logical (i.e. it is subjective judgement).

5.3 Model assessment

In this section, seven different models of the soil moisture processes occurring at Tarrawarra are assessed by making comparisons of the modelled and observed spatial fields. This assessment is representative of how more complex model assessments, such as those involving Monte Carlo simulations (e.g. Bates et al. 2004), could also be conducted. The measures available for use in this assessment are those listed in the comparison flowchart (Figure 4.20). As described in the previous section, the assessment is applied here in a manner that simulates automated procedures, so visual inspection, comparison and analysis is not used.

The remainder of this section details the comparison strategy used for conducting the assessment of Tarrawarra; uses the comparison measures to compare modelled and observed fields for each of the four occasions; and presents a manual analysis of the comparison measures. The other sections in this chapter conduct the evaluation of the model assessment (i.e. using the expert assessment and visual inspection) and discuss the suitability of the comparison strategy used (i.e. the strengths and weaknesses of each group of measures).

5.3.1 Details of comparison strategy

The comparison strategy devised for assessing the models of Tarrawarra is guided by the general aspects of spatial fields that were compared during the expert model assessment. Table 5.2 details the strategy that has been devised for the quantitative comparison, including the measures; their parameters or pre-processing; and a description of what each measure is expected to assess about the models (based on their operation). This strategy is applied to the Tarrawarra dataset and the performance of the various measures (i.e. describing how the modelled and observed spatial fields compare) is evaluated. A discussion of this strategy is provided later in the chapter, identifying any issues with the way the measures were applied or how they operated.

5.3.1.1 Rationale for parameters chosen

The multiscale error measure is calculated for radius values up to 6 elements. This is set by using a brief analysis of how upscaling affects the variability seen in the observed fields. As Figure 5.6 shows, beyond R = 6 elements, any comparisons would simply be replicating a global comparison. The scale weighting parameter, K = 0.1, is set so that the larger scales (which often appear similar) are given slightly less weighting than the local scales (where most variation occurs) in the final $MSME_{0.1}$ measure.

The allowed value difference ($|\Delta V|_{ALLOW}$) is set to the reported measurement error of the observations (σ = 1.7% v/v). This is a conservative value, as it only represents one standard deviation of the error (i.e. 68% of errors under this amount if normally distributed errors). The maximum allowed value difference ($|\Delta V|_{LIMIT}$) is set at 5.1% v/v (i.e. 3σ) for the tolerant similarity measures.

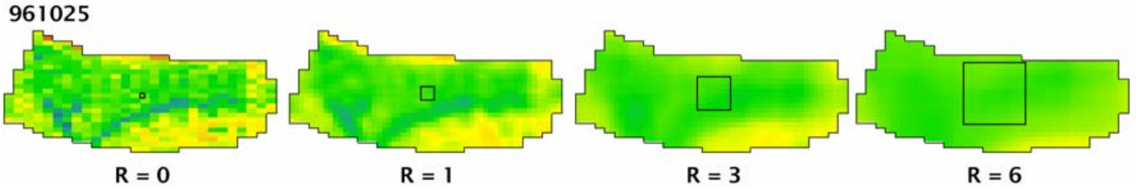


Figure 5.6 A brief analysis of upscaling an observed field (961025) reveals that almost all localised variability is removed once R = 6 elements. The boxes overlain on each field represent the size of the analysis windows used at each scale.

Table 5.2 The list of comparison measures used in the comparison strategy for model assessment of the Tarrawarra models. The parameters (or pre-processing required) and a description of what the measure is expected to assess about the models is given for each measure. The data-driven region-based measure (shaded) produces a variable number of results for each occasion (i.e. dependent on number of important regions identified in observed field).

Group	Measure(s)	Parameters/Pre-processing	Description
G E	BIAS	-	Over- or under-estimation of model
G S	KS	-	Similarity of element value distributions between model and observation
L E	MAE	-	Strict error per model element
L S	RSQ	-	Amount of observed variability that is represented by the model

M E	$MSME_K$	K = 0.1
		R = 0 to 6 elements
			Summary of model error over scales where variability in the observed field is still apparent

TL E	TMAE	$\|\Delta V\|_{ALLOW}$ = 1.7% v/v $\|\Delta L\|_{ALLOW}$ = 10 m	Significant error per model element
TL S	TCOE	$\|\Delta V\|_{ALLOW}$ = 1.7% v/v $\|\Delta L\|_{ALLOW}$ = 10 m	Performance of model relative to 'feature-less' field; used for inter-comparison
TL S	TMS	$\|\Delta V\|_{ALLOW}$ = 1.7% v/v $\|\Delta V\|_{LIMIT}$ = 5.1% v/v $\|\Delta L\|_{ALLOW}$ = 10 m $\|\Delta L\|_{LIMIT}$ = 30 m	Significant similarity of model elements
R S	KS	Segment observed field; remove unimportant regions; parameters are empirically determined	Similarity of model to observed region; reflects whether region exists in model (in some form)
STL E	TMAE	Define slope regions from terrain model (classify slope map and connect regions); use tolerances previously defined	Significant model error per slope region; describes model performance for drainage lines, hillslopes and ridge tops
STL E	TMAE	Define aspects regions from terrain model (classify aspect map into north/south and connect regions); use tolerance previously defined	Significant model error per aspect region; describes model performance for slopes facing north and south (i.e. different incident solar radiation)

The allowed location difference ($|\Delta L|_{ALLOW}$) has been set to 10m due to the positional uncertainty introduced when creating the observed fields. Each point-based observation was assumed to have a 10m x 20m support, but these observations were not necessarily located at the centre of each regular-element in the final field. A half-element distance is the maximum error that should exist in the location of these measures. The maximum location difference considered to be similar ($|\Delta L|_{LIMIT}$) is arbitrarily set to 30m to allow some spatial relationships to be considered. This produces a location similarity of 0.5

when the element values differ in location by 20m. The location and value tolerances are combined together during comparison, making them capable of tolerating situations where both differences occur. This has the effect of allowing the observed field to be 'adaptively smoothed' and may address any scale inconsistencies that exist.

5.3.1.2 Pre-processing required to undertake comparisons

Data-driven segmentation of the observed field on each occasion is required to identify any interesting regions (or features) that should be represented by the modelled fields. This is an interactive process in which the segmentation parameters (i.e. scale and shape controls) are modified to produce a segmentation that recognises the key features in the observed field (if such features exist). The region shape parameter is generally kept constant (i.e. at one) unless 'branched' regions are being found. Any unimportant

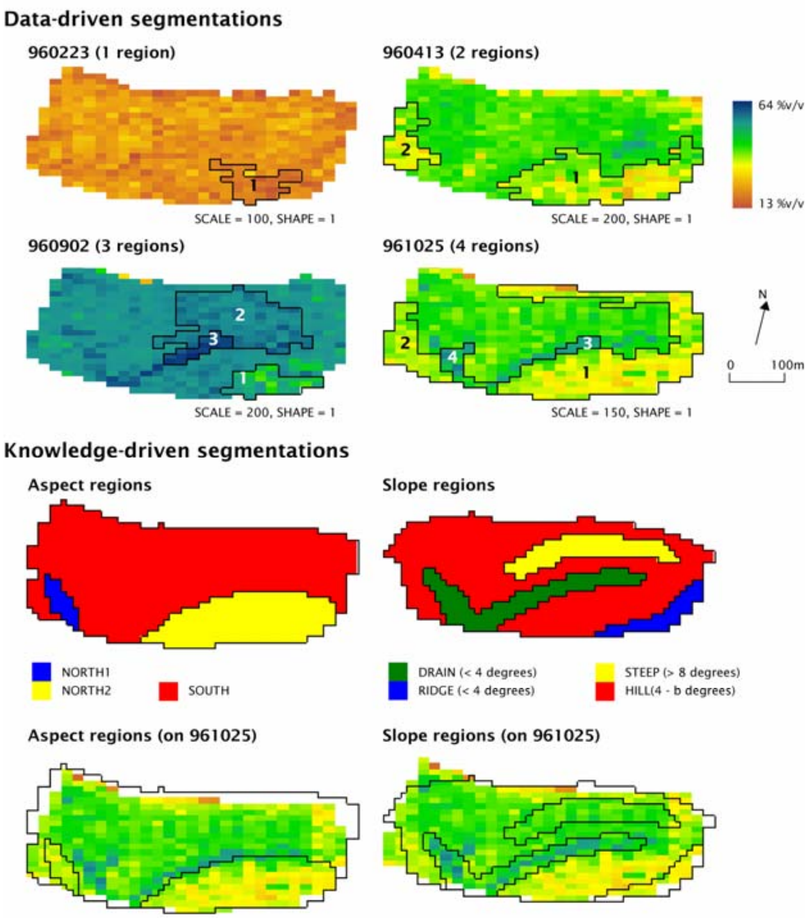


Figure 5.7 Data-driven and knowledge-regions defined for Tarrawarra. For the data-driven regions, only the interesting regions have been retained (numbered for identification in results). The knowledge-driven regions are derived from terrain analysis and any very small regions have been removed.

regions that are recognised are removed to leave only the important ‘features’ (as decided by the user). The data-driven regions used for Tarrawarra are shown in Figure 5.7.

Knowledge-driven segmentation of the catchment is undertaken using terrain attributes of slope and aspect to define the regions of hydrological importance to be used for region-specific comparisons (Figure 5.7). The slope attribute of the terrain model for Tarrawarra is particularly effective at identifying the drainage lines and ridge top (i.e. low slopes) and the hillslopes (i.e. the steeper slopes). The derived slope field has been categorised into low, mid and high slope regions, and these have been made into contiguous regions and suitably labelled. Aspect has been derived from the terrain model and adjusted to account for the field rotation relative to true north. The aspect values were categorised into elements facing north (i.e. greater incident solar radiation) and south. Again, the regions were made contiguous and labelled. The data- and aspect-driven segmentations have some generally similar features, which must be a response to aspect-related arrangements in the observed fields of soil moisture.

The situation investigated at Tarrawarra is a relatively simple one, as there are a minimal number of ‘features’ defined for each occasion. In more complex catchments, non-contiguous regions with the same label (e.g. many different drainage lines) may be recognised, but the user can decide to group these together or treat them individually.

5.3.2 Analysis of comparison measures

The comparison strategy outlined in the previous section is applied here to assess the models against the observed fields. The comparison strategy identifies eight measures that apply to the whole field extent; two measures that apply to each of the seven knowledge-driven regions (i.e. four slope, three aspect); and one measure that applies to a variable number of different data-driven regions (i.e. maximum four), depending on the occasion.

These results for all of the comparison measures (sorted into groups according to comparison flowchart in Figure 4.20) are listed on the following pages in Table 5.3-Table 5.5. The major points of analysis from these tables are listed in Table 5.6. The optimum result(s) for each measure on each occasion are highlighted in bold in the tables, providing an indication of the 'best' model runs (i.e. the runs that perform strongly across multiple different measures). However, assessment such as this should consider the relative importance of each measure (i.e. the number of 'hits' for a model does not necessarily reflect the best). As was noted during expert assessment, "differences between the runs were often subtle and no one run was the best on every occasion" (Western and Grayson 2000, p.228). Therefore, it is the task of the user to define 'optimal performance' for their purposes, but the way to do this is not the focus of this analysis (or this thesis). Instead, the focus of this analysis is to highlight how the different measures identify the optimum performance and whether these results can be supported by expert and/or visual evidence.

Table 5.3 Results from using strict global, local and multiscale error and similarity measures (applied to whole field) for model assessment at Tarrawarra. The comparison measures and parameters used are defined in the comparison strategy (Table 5.2). The optimum result for each measure on each date is bolded. Coloured shading of any results indicates they are related to findings listed in the analysis table (Table 5.6).

			Statistics of observed field		**Global comparison (G)**		**Strict local comparison (L)**		**Multi-scale comparison (M)**
Field			**MEAN %v/v**	**SDEV %v/v**	**BIAS %v/v**	**KS**	**MAE %v/v**	**RSQ**	**MSME$_{0.1}$ %v/v**
960223			20.8	2.3					
	DRY	Run 2			0.6	0.55	1.8	0.02	1.0
		Run 5			0.9	0.57	1.9	0.05	1.1
		Run 6			0.8	0.57	1.9	0.03	1.1
		Run 7			1.4	**0.67**	2.7	0.01	1.8
		Run 8			**0.5**	0.60	**1.7**	**0.09**	**0.9**
		Run 9			1.0	0.61	2.2	0.02	1.4
		Run 10			0.6	0.59	1.8	0.06	**0.9**
960413			35.1	3.5					
	WETTING UP	Run 2			-0.6	0.41	2.8	0.13	1.8
		Run 5			0.1	0.56	2.4	0.21	1.3
		Run 6			**0.0**	0.56	2.4	0.20	1.3
		Run 7			0.5	**0.80**	3.2	0.00	2.0
		Run 8			-0.1	0.56	2.5	0.17	1.4
		Run 9			0.1	0.61	**2.3**	**0.27**	**1.2**
		Run 10			0.0	0.56	2.5	0.18	1.4
960902			42.5	3.8					
	WET	Run 2			-0.3	**0.70**	2.7	0.16	1.7
		Run 5			**-0.2**	**0.70**	**2.6**	**0.19**	**1.6**
		Run 6			-0.9	**0.70**	2.8	0.17	1.8
		Run 7			-1.0	**0.70**	2.8	0.18	1.8
		Run 8			-0.9	0.69	2.7	0.16	1.7
		Run 9			-0.3	**0.70**	2.7	0.16	**1.6**
		Run 10			-0.7	**0.70**	2.7	**0.19**	1.7
961025			35.1	4.4					
	DRYING DOWN	Run 2			5.8	0.33	6.1	0.29	5.8
		Run 5			6.2	0.30	6.4	0.34	6.2
		Run 6			5.4	0.50	5.6	0.39	5.5
		Run 7			4.8	**0.63**	5.3	**0.40**	4.9
		Run 8			**4.6**	0.44	**5.2**	0.31	**4.7**
		Run 9			6.1	0.37	6.4	0.33	6.1
		Run 10			5.0	0.42	5.4	0.35	5.1

Table 5.4 Results from using tolerant local comparison (applied to whole field) and also region-based similarity measures for model assessment at Tarrawarra. The comparison measures, parameters and regions used are defined in the comparison strategy (Table 5.2). The regions defined on each occasion vary (i.e. DRY REG 1 ≠ WET REG 1) and can be better understood using Figure 5.7. The optimum result for each measure on each date is bolded. Coloured shading of any results indicates they are related to findings listed in the analysis table (Table 5.6).

		Tolerant local comparison (TL)			**Region-based (R) comparison of important observed regions**			
Field		**TMAE %v/v**	**TCOE**	**TMS**	**REG 1 KS**	**REG 2 KS**	**REG 3 KS**	**REG 4 KS**
960223								
DRY	Run 2	0.1	-0.63	**0.99**	0.00			
	Run 5	0.1	-0.88	**0.99**	0.03			
	Run 6	0.1	-1.84	0.98	0.03			
	Run 7	0.7	-20.84	0.91	0.03			
	Run 8	**0.0**	**0.18**	**0.99**	0.03			
	Run 9	0.3	-7.54	0.95	0.03			
	Run 10	**0.0**	-0.03	**0.99**	0.03			
960413								
WETTING UP	Run 2	0.3	-0.16	0.93	0.31	0.25		
	Run 5	**0.2**	0.28	0.96	0.40	0.28		
	Run 6	**0.2**	0.27	0.96	0.40	0.28		
	Run 7	0.6	-1.17	0.89	**0.49**	**0.32**		
	Run 8	**0.2**	0.23	0.96	0.41	0.28		
	Run 9	**0.2**	**0.37**	**0.97**	0.41	0.28		
	Run 10	**0.2**	0.24	0.96	0.41	0.28		
960902								
WET	Run 2	**0.1**	**0.10**	0.97	0.17	**0.34**	0.00	
	Run 5	**0.1**	0.30	**0.98**	0.17	**0.34**	0.00	
	Run 6	0.2	-0.40	0.95	0.54	**0.34**	0.00	
	Run 7	0.2	-0.50	0.96	**0.55**	**0.34**	0.00	
	Run 8	0.2	0.00	0.97	0.50	0.32	0.00	
	Run 9	0.2	0.00	0.97	0.25	**0.34**	0.00	
	Run 10	**0.1**	**0.10**	0.97	0.38	**0.34**	0.00	
961025								
DRYING DOWN	Run 2	2.1	-5.54	0.65	0.07	0.16	0.57	0.59
	Run 5	2.3	-6.28	0.60	0.06	0.16	0.57	**0.65**
	Run 6	1.8	-4.62	0.69	0.13	0.29	0.14	0.53
	Run 7	1.7	-4.38	**0.71**	**0.32**	**0.65**	0.00	0.35
	Run 8	**1.5**	**-3.73**	0.70	0.11	0.16	0.14	0.41
	Run 9	2.3	-6.28	0.60	0.10	0.27	**0.71**	0.59
	Run 10	1.6	-3.94	0.70	0.10	0.29	0.43	0.53

Table 5.5 Results from using specific tolerant local comparison measures for model assessment at Tarrawarra. The comparison measure, parameters and regions used are defined in the comparison strategy (Table 5.2). Abbreviated terms are used to describe each of the hydrological regions defined from the terrain model and these can be better understood using Figure 5.7. The optimum result for each measure on each date is bolded. Coloured shading of any results indicates they are related to findings listed in the analysis table (Table 5.6).

			Specific tolerant local comparison (STL) of slope-based regions				Specific tolerant local comparison (STL) of aspect-based regions		
Field			**DRAIN TMAE**	**RIDGE TMAE**	**HILL TMAE**	**STEEP TMAE**	**NORTH 1 TMAE**	**NORTH 2 TMAE**	**SOUTH TMAE**
960223									
	DRY	Run 2	0.2	0.1	0.1	**0.0**	**0.0**	**0.0**	0.2
		Run 5	0.2	0.1	0.1	**0.0**	**0.0**	**0.0**	0.2
		Run 6	0.4	0.1	**0.0**	**0.0**	**0.0**	**0.0**	0.2
		Run 7	4.7	**0.0**	0.1	**0.0**	**0.0**	0.2	0.9
		Run 8	**0.0**	0.1	**0.0**	**0.0**	**0.0**	**0.0**	**0.1**
		Run 9	1.6	0.1	0.1	**0.0**	**0.0**	**0.0**	0.4
		Run 10	0.0	0.1	**0.0**	**0.0**	**0.0**	**0.0**	**0.1**
960413									
	WETTING UP	Run 2	0.2	1.1	0.3	0.2	0.4	**0.4**	0.4
		Run 5	0.2	**1.0**	**0.1**	**0.0**	0.4	**0.4**	0.3
		Run 6	0.2	**1.0**	**0.1**	**0.0**	0.4	**0.4**	0.3
		Run 7	0.5	1.8	0.5	0.1	1.5	0.7	0.7
		Run 8	0.3	**1.0**	**0.1**	**0.0**	0.4	**0.4**	0.3
		Run 9	**0.1**	**1.0**	**0.1**	**0.0**	**0.3**	**0.4**	**0.2**
		Run 10	0.2	**1.0**	**0.1**	**0.0**	0.4	**0.4**	0.3
960902									
	WET	Run 2	**0.5**	0.3	**0.1**	**0.0**	**0.0**	**0.1**	**0.3**
		Run 5	**0.5**	0.2	**0.1**	**0.0**	**0.0**	**0.1**	**0.3**
		Run 6	**0.5**	0.2	0.2	0.2	0.4	**0.1**	0.5
		Run 7	**0.5**	0.3	0.2	0.1	0.6	0.2	0.5
		Run 8	0.6	**0.1**	**0.1**	**0.0**	**0.0**	**0.1**	0.4
		Run 9	**0.5**	0.2	**0.1**	**0.0**	0.4	**0.1**	0.4
		Run 10	**0.5**	**0.1**	**0.1**	**0.0**	0.2	**0.1**	**0.3**
961025									
	DRYING DOWN	Run 2	1.1	4.6	2.4	0.7	2.7	3.7	1.8
		Run 5	1.1	4.5	2.7	1.3	2.6	3.4	2.2
		Run 6	1.9	3.3	2.0	0.4	1.0	2.5	1.8
		Run 7	2.5	**1.8**	1.8	**0.2**	**0.8**	**1.9**	1.8
		Run 8	**0.7**	3.1	**1.7**	1.3	2.3	2.4	**1.4**
		Run 9	1.2	4.0	2.6	1.6	2.4	3.3	2.2
		Run 10	1.0	3.5	1.8	0.7	1.5	2.5	1.5

Table 5.6 List of major findings from Table 5.3-Table 5.5. Each finding from the analysis is related to certain measures, dates and runs, so these are identified and the finding is described. The colour shading of the findings relates back to the tables of results.

Table	Measure(s)	Date(s)	Run(s)	Description
5.3	ALL	960223	8, 10	Performance of all models quite similar with different measures, but these runs are best overall
5.3	KS, RSQ	960413	7	Contrasting performance suggests globally representative of values, but spatial arrangement is wrong
5.3	ALL	960413	9	Best overall on this occasion
5.3	ALL	960902	ALL	Nearly identical performance, apart from BIAS; all are globally good, but locally mediocre
5.3	BIAS	960902	ALL	Models are consistently under-predicting soil moisture
5.3	BIAS	961025	ALL	Models are consistently over-predicting soil moisture
5.3	ALL	961025	7, 8	Best overall performance
5.3	MSME$_{01}$	ALL	ALL	Provides good overall summary of scales that reflects the best and worst models in each case
5.3	KS	ALL	ALL	Values generally above 0.50 unless there is a major bias in the models; useful to combine with BIAS measure for global performance of model
5.4	ALL TL	960223	8	Best performance; only model better than ‘feature-less’ observed mean field (i.e. TCOE)
5.4	TMAE	960223	7	Local errors much higher than other models, despite tolerances; reflects significant local errors
5.4	ALL TL	960223	2-6 8-10	Models are very similar to observed when minor tolerances are allowed, but the lack of features in observed makes performance better than observed mean field difficult to achieve (i.e. negative TCOE)
5.4	REG 1	960223	ALL	Region is poorly represented, possibly due to variance of region not being represented in models
5.4	ALL TL	960413	9	Best performance, although all models perform similarly and most are better than mean field (i.e. positive TCOE)
5.4	REG 1	960413	7, 2	Best represented in run 7 and worst in run 2; this region is aspect related, so poor performance in run 2 is likely due to the absence of SVET in model
5.4	ALL TL	960902	ALL	Little to no error remains in the model once minor differences are tolerated
5.4	REG 1	960902	6, 7, 8	Best representations of the ‘drier hillslope’ region
5.4	REG 2	960902	ALL	All models represent ‘larger saturated’ region in mediocre manner (KS $\approx$ 0.34)
5.4	REG 3	960902	ALL	No models represent the ‘small saturated’ region in the observed field

Table	Measure(s)	Date(s)	Run(s)	Description
5.4	TMAE	961025	ALL	Significant model errors exist in all runs, possibly due to effects of over-prediction
5.4	REG 1, 2	961025	7	Best represents the ‘lower moisture on northern slope’ regions
5.4	REG 4	961025	2, 5, 9	Best represent the ‘smaller drainage line’ region
5.4	TCOE	ALL	ALL	Models on 960413, 960902 are the best matching in terms of spatial arrangement; measures for 960223 are influenced by lack of organisation in observed field; 961025 is poor performance due to bias
5.5	DRAIN	960223	ALL	Quite variable results for this region, suggesting different representations in this part of the field in each model; runs 7 and 9 are worst
5.5	ALL STL	960413	7	Performs worst for most measures, suggest fundamental problem with spatial arrangement
5.5	ALL STL	960413	2-6 8-10	Very similar performance for all these models, although model run 9 is slightly better on all measures
5.5	RIDGE	960413	ALL	Worst performing part of fields with all runs, particularly run 7
5.5	ALL STL	960902	ALL	Similar results for all models and no particularly large model errors in any region; suggests comparable appearance and arrangements that match observed field within tolerances
5.5	DRAIN, HILL, SOUTH	961025	8	Best performing model for these regions in this highly organised observed field
5.5	NORTH 2	961025	7	Best representation of larger northern facing hillslope

5.4 Evaluation of model assessment

The analysis table (Table 5.6) identifies a variety of results from the different dates, runs and measures that are interesting for identifying the optimal model and for understanding why a certain model is optimal. The differing appearance of the four observed fields (i.e. ranging from having random noise in the dry period, to highly organised features in the drying down period) makes it challenging to recognise one optimal model, as was noted by Western and Grayson (2000). Also, the differences between the measures for each model are often minimal, so the optimal result is often only marginally better.

The expert assessment made by Western and Grayson (2000, p.228) was that "run 5 was judged to be the best", although "the winter and spring periods were emphasised in [the] comparison because [of the] strong spatial organisation" present. Throughout the results and analysis tables in this chapter, run 5 was rarely found to be the optimal model, except where many of the models performed similarly. In the spring field (961025), run 5 was actually the worst performing run with the strict global and local measures. In contrast, this analysis found runs 7 and 8 to both perform optimally on many occasions, although run 8 is the more consistent performer. Run 7 produced a number of poor results (e.g. the error in the drainage line in the dry period) that suggest it may be less optimal across all dates. The overall optimal model, based on this quantitative analysis, is considered to be run 8, although run 5 does not perform much worse. The model that also shows promise is run 7, although there are some major errors detected in this model that preclude it from being optimal.

The more detailed evaluation that follows addresses three different groups of measures: 1) the global, local and multiscale measures applied to the entire field; 2) the strict and tolerant measures applied to the entire field; and 3) the measures produced for different regions of importance. For each group, some of the major findings from the analysis are investigated further and contrasted with expert and visual evidence. There are many other minor findings from the analysis, but for the sake of brevity these are not discussed here.

5.4.1 Using global, local and multiscale measures

The strict global and local measures listed in Table 5.3 represent the information that is most widely currently used in hydrological modelling for assessment. In addition to these measures, the new multiscale measure (Algorithm 4.9) is evaluated here. All of these measures apply to the whole field, thereby making them suitable for quantifying statements about the average or typical soil moisture errors, or general similarity.

5.4.1.1 Evaluation in the dry period (960223)

In the dry period (960223), the observed field is featureless apart from apparently random noise. This is best simulated by runs 8 and 10, although at the local scale (MAE, RSQ) there are errors of 1.7% v/v and the models general fail to explain any of the variance in the observed field (i.e. low RSQ). The MAE value is of the same magnitude as the measurement error in the observations, suggesting that this may be an insignificant error (although it is not tolerated with these strict measures). The expert assessment suggests that the average soil moisture and soil moisture pattern were "well simulated" on this occasion, although the global and local measures do not really reflect this (i.e. some bias and low fit). The MSME provides a convenient summary of the scales (as it is designed to) and actually reveals that runs 8 and 10 are most similar if considering the 'important scales' (i.e. up to R = 6). Visual inspection of Figure 5.4 confirms this to be logical, as these runs are the only ones without a linear drainage feature (which would cause larger errors and be influential across intermediate scales also). The measures also correctly identify run 7 as being worst on this occasion, although run 7 does have the best fit between distributions (i.e. highest KS) due to the greater variability it simulates (i.e. because of variable soil depths).

5.4.1.2 Evaluation in the wetting up period (960413)

As the catchment begins wetting up in the autumn (960413), run 9 is the strongest performing model across all scales (i.e. lowest MSME). On this occasion, runs 5, 6, 8, 9 ane 10 are all quite similar, with runs 2 and 7 being poor. All runs have quite large typical error and relatively low RSQ, suggesting an inability to model the local details. Of interest here was the particularly high KS similarity (KS = 0.80) for run 7, yet RSQ of zero. This translates to high global similarity, but no agreement between high/low values at a local level. Visual inspection of Figure 5.4 supports this finding, although there does appear to be some local agreement between elements. As RSQ is often overly influenced by mismatches between extreme values, the high values in the drainage line may be the cause of overall low RSQ here, but this could not be ascertained from the quantitative analyses alone. The expert assessment found good prediction of the average moisture content on this occasion, which is revealed in the low BIAS values. However, these measures cannot determine whether the 'significant lateral redistribution' in the observed field was simulated or not, due to their summary of the entire field. The low RSQ and high MAE values (approximately 2.5% v/v) suggest poor agreement of the general spatial arrangement.

5.4.1.3 Evaluation in the wet period (960902)

All of the global, local and multiscale measures were highly similar, suggesting that all models perform equally on the wet occasion. BIAS provides some discrimination between the measures, with runs 2, 5 and 9 all performing with the least amount of under-prediction. The similar results obtained here concur with visual inspection, which reveals it to be difficult to detect any differences between models. The MAE values for this occasion are of similar magnitude to the previous date, suggesting that the model fails to represent the detailed strict local arrangement. The expert assessment, which suggests "the soil moisture pattern is well predicted", is not supported by the local measures because low RSQ and high MAE are found. The expert assessment is likely to be more tolerant of minor errors, which may be causing this mismatch between the quantitative and qualitative assessments.

5.4.1.4 Evaluation in the drying down period (961025)

The observed field in the drying down period has the most spatial organisation and appears the most challenging to correctly simulate. This is one of the reasons the expert assessment focused on this period and the wetting up period during model assessment. The results here are quite variable, with runs 7 and 8 each being optimal for different measures. Run 8 is found to be the best for the global, local and multiscale error measures. Run 7 performs more strongly with the similarity measures. The expert assessment found the model to be generally quite poor on this occasion, particularly with failing to capture the variations with terrain aspect correctly, but this cannot be assessed with these generalised measures. The BIAS and MAE measures show there to be a large amount of over-estimation in all model runs, which reflects the generally poor performance.

5.4.1.5 Assessment using these measures

If these measures were used in a calibration or optimisation process, the most promising model run would most likely be run 8 or 9. This is, of course, entirely dependent on how heavily weighted each date is in the automated procedure. The expert assessment weighted the wetting and drying periods more highly, yet found run 5 to have the best performance. Based on these quantitative measures, this finding is not supported, although run 5 can not be said to perform badly on any occasion. The global and local measures provide quite different tests of agreement, as do the error and similarity measures, thereby testing different aspects of comparison. The multiscale measure also appears to be a useful summary of these results, especially when it is focused onto the most important scales. It also allows the findings from global and local scales to be confirmed, which informs the user that there must be no immediately conflicting results found at intermediate scales. Closer inspection of the multiscale plot would be required to understand more about intermediate scale differences, but this is not possible in automated procedures. Overall, these measures are a useful and exact comparison, although they fail to integrate user knowledge into the analysis (apart from the definition of scales for MSME).

5.4.2 Contrasting strict and tolerant measures

The tolerant measures reveal the significant error and similarity in the model runs. When the tolerances are set appropriately, any remaining errors can be safely assumed to be important and a major departure from the observed field. The tolerances used in this model assessment allow for measurement error and minimal locational differences to exist without affecting the measure. This causes many of the models to have TMAE ≈ 0 on the dry, wetting up and wet occasions (i.e. no significant error). Runs 2 and 7 have significant model error remaining on these occasions, while run 9 is only in error during the dry period. All of these can be considered poor simulations, although the reasons why they are poor cannot be established because they are calculated across the entire spatial extent. This is also a visually intuitive result, as runs 2, 7 and 9 each contain ‘false features’ on specific occasions.

In contrast to the strict local measures, using tolerant measures makes the models containing ‘true’ errors much more apparent during analysis (i.e. they are the only ones not to be very low). They are very highly correlated with the strict measures on the dry and wetting up occasions, although when the tolerances are particularly ‘active’ in the comparison (i.e. when they are frequently used), this correlation drops substantially and the tolerant measures provide different information. This occurs on the wet occasion, where the TMAE is only 67% correlated with the MAE.

The TCOE measure provides a useful way of assessing good performance across all the occasions. Run 8 is the only model not to be considered worse than the ‘featureless’ observed mean field, against which the TCOE is standardised. However, this measure penalises biases in the model substantially (as do all the tolerant measures), which is seen particularly on the drying down occasion. All of the TCOE values are negative on this occasion due to the observed mean being a much closer estimate of the observed field than any model.

In general, the tolerant measures produce multiple optimum models more often than the strict measures, as the tolerances reduce the variability within each measure and the ability to discern between ‘slightly different’ fields. However, provided that the tolerances are wisely and conservatively set, the multiple optima should be appropriate, represent acceptable performance and importantly let the user identify and reject poor models.

5.4.2.1 Visually inspecting the effect of tolerances

To gain some insight into the localised impact of introducing tolerance, the ‘significant errors’ remaining in the models on 961025 are visually examined. Three situations are shown in Figure 5.8 to illustrate how tolerances work to reduce the error (or increase similarity) between elements. This figure shows the intermediate similarity fields from the tolerant comparisons for runs 7, 8 and 10 during the wet period at Tarrawarra. Looking at run 7, the circle marked in Figure 5.8 shows a gully region where run 7 has predicted some elements with too much moisture. When only a value tolerance is

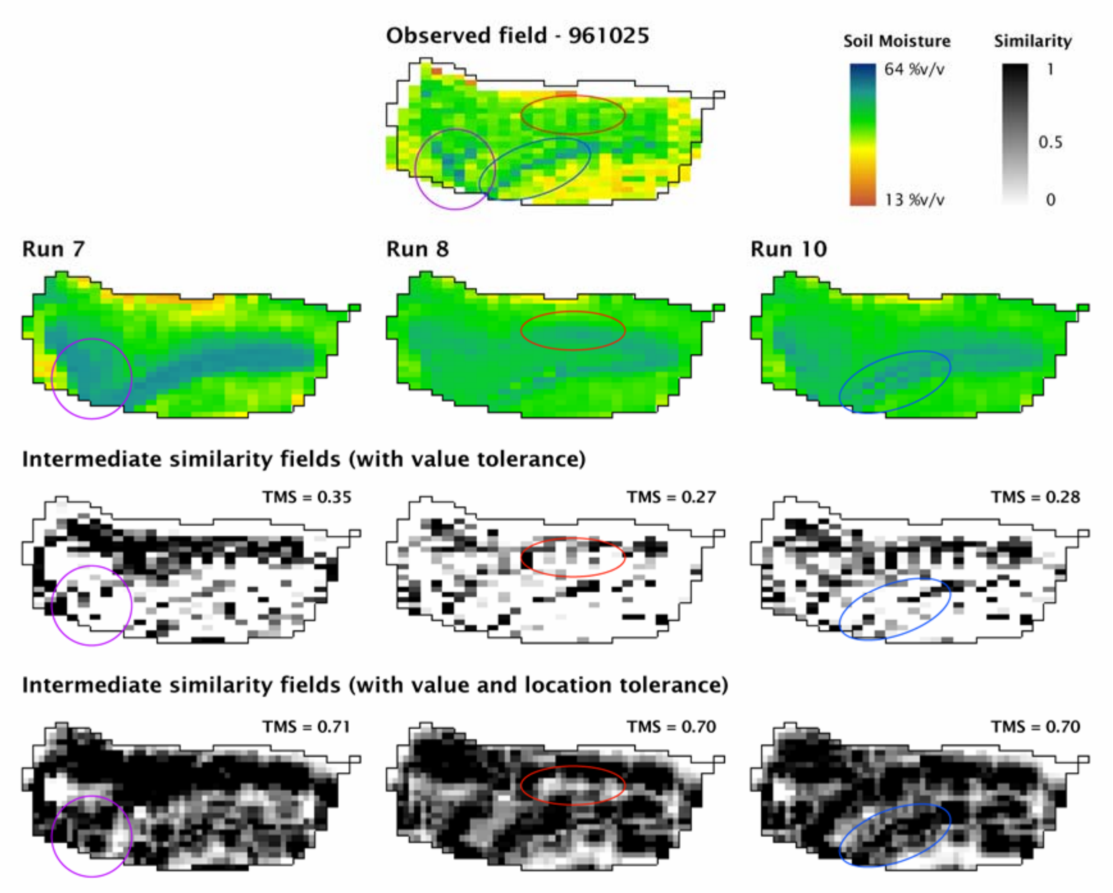


Figure 5.8 The influences of location and/or value tolerances are shown here to identify how they can produce global comparable results, but the local arrangement of similarity is quite different. This figure also illustrates how three different 'features' are assessed with and without location tolerances.

introduced, there is similarity found where the few observed cells have similar moisture contents. Introducing the location tolerance produces much more extensive similarity within this region because the 'too wet' modelled elements are allowed to search within 30m (i.e. 3 elements) for a similarly wet element. In this situation, the tolerance has the effect of allowing the wider, modelled wet area to match with the narrow, observed wet area (but with lesser similarity due to distance difference).

The ellipse shown on run 8 indicates a modelled wet region that does not exist in the observed field. This region has formed at the boundary between the two different soil types (and associated depths) specified for this model run. When there is only a value tolerance, no similarity is found throughout most of this feature. The location tolerance increases similarity, but because no elements as wet are found within the tolerance distance, low similarity is still assigned to the elements making up this feature. Run 10 looks at a situation where there is minimal agreement between the wide gully feature in the model and the observed feature. Due to only minimal mismatch in the location (i.e. 1-2 elements), the location tolerance recognises higher similarity around this feature. The end result of using a combined tolerance on these fields is that they are each judged as being almost equally similar (TMS ≈ 0.70), although the local contributions to this summary differ considerably.

5.4.3 Using importance regions

Three different types of regions are used in this model assessment, attempting to emulate the regions that are utilised in the expert model assessment. Western and Grayson (2000) made comparisons of the drainage lines, hillslopes, ridge tops and different facing slopes during the expert assessment. These regions were used because they respond to particular hydrological processes. For example, the drainage lines are where moisture collects and saturated soils are found, while the hillslopes are where the impacts of evapotranspiration are most obvious. The evaluation conducted here separately discusses each of the region types used, then suggests how the importance measures would impact on any model assessment conducted.

5.4.3.1 Data-driven regions

The data-driven regions have been recognised within the observed field on each occasion and represent ‘features’ in the observations. On the dry occasion, the only data-driven region is a small area where the moisture values are consistently lower than the rest of the field. It is not immediately apparent what this region represents, but expert knowledge identifies this as an area where the soils were particularly shallow (R. Grayson, personal communication, 17/05/2006). For evaluating the features, the KS measure is used. It finds that none of the model runs simulate this feature, although this may be a result of the KS measure only being based on a small number of elements (which limits the detail of the distributions used). For small features such as this, the BIAS measure is expected to be a more useful. In this example, using BIAS per region would have found run 7 to be the best, which is an intuitive result considering that this run uses variable depth soils. This region was not recognised during expert assessment.

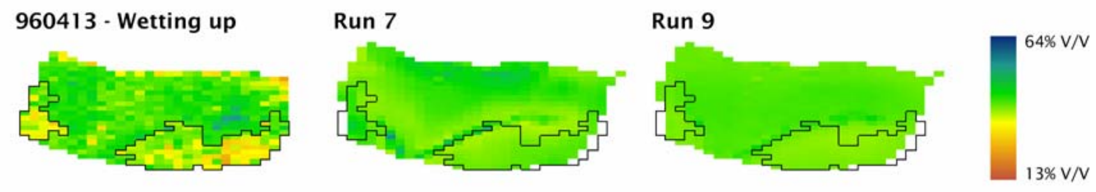


Figure 5.9 Visual inspection of data-driven regions being compared on 960413.

During wetting up, two regions of lower moisture are found. Based on knowledge of the terrain, these regions appear related to the north facing slopes. On this occasion, run 7 is found to simulate the features better than all other runs, despite performing poorly with most of the other measures. This is in contrast to the measures for the whole fields, which found run 9 to be optimal. As can be seen in Figure 5.9, the fields are not vastly different in these regions, although run 7 does correctly simulate more 'drier elements' and this produces the higher KS measure. Using a BIAS measure would have recognised run 9 to be the better run, although the amount of BIAS (approximately 2.5% v/v over-estimate) would suggest the representation of features was far from exact.

In the wet field, the regions recognised include a small area of slightly drier values on a ridge top; a small highly saturated area; and a large, but less visually distinct saturated area. Of these fields, the small highly saturated area is not represented well in any of the models. The larger saturated region is represented similarly in all runs (KS = 0.34). The drier ridge feature is best simulated in runs 6, 7 and 8, which concurs with visual inspection of Figure 5.4, but it is not particularly obvious. Again, none of these features were recognised as being important during expert assessment, yet they certainly exist.

During drying down, there is particularly strong spatial organisation and the regions recognised are obvious features. The region-based measures on this occasion reveal run 7 to simulate the drier features (regions 1 and 2) best, but the small drainage line features (regions 3 and 4) poorly. In contrast, runs 2, 5 and 9 all simulate the smaller features within the drainage features best but not the drier features. These inconsistent results between runs make it quite difficult to select an optimal run. This may be due to the drainage features being quite small and therefore very sensitive to minor errors. None of the data-driven regions recognised using segmentation were used in the expert assessment. Instead, the expert assessment focused on using knowledge-driven regions.

5.4.3.2 Knowledge-driven slope and aspect regions

Slope was a particularly useful terrain attribute for dividing up the spatial extent into process-related regions. Low slope areas could be separated into either drainage lines or ridge tops, while the mid and high slope values identified the major hillslope parts of the catchment. Each of the aspect regions overlaps the slope regions, but they change the focus of the results to be the regions that face either north or south. In each of these areas, the models have been assessed using a tolerant measure (TMAE), so that only significant differences are reflected in the results.

On the dry occasion, all of the regions are well represented, apart from the drainage lines and southern facing slopes in runs 7 and 9. These measures make clear and specific findings about why runs 7 and 9 are not good models on this occasion - too much moisture in the drainage lines (which are also south facing). This measure is in accordance with visual inspection (Figure 5.4). By using tolerances within each of the regions, any errors caused by the lesser drainage lines seen in runs 2 and 5 are reduced. This information is particularly useful for model understanding and can quantify how bad particular components of the model are.

As the catchment wets up, run 7 continues to perform the worst and all of the other models are nearly identical for all of the slope and aspect regions. Run 9 is the overall optimal, but not by any great amount. On this occasion the worst performing region is the ridge top, where the TMAE values of more than 1% v/v exist (i.e. after tolerances have been applied). This suggests the ridge top is not well simulated. The expert assessment found that “the significant lateral redistribution” in the observation was not simulated on this occasion, which is supported by the poor ridge top simulations. When the lateral redistribution of moisture is occurring, the ridge top would be much drier, yet in the models there is consistent over-estimation (i.e. too wet) in this region.

The wet occasion was assessed by the experts as having “the soil moisture pattern well predicted.” The region specific measures support this finding, as all of the different regions have similar comparable performance and the significant errors are much less than 1% v/v. The drainage lines and southern slopes contain the largest model errors on this occasion, although they are under-estimates (i.e. not as saturated as observed field).

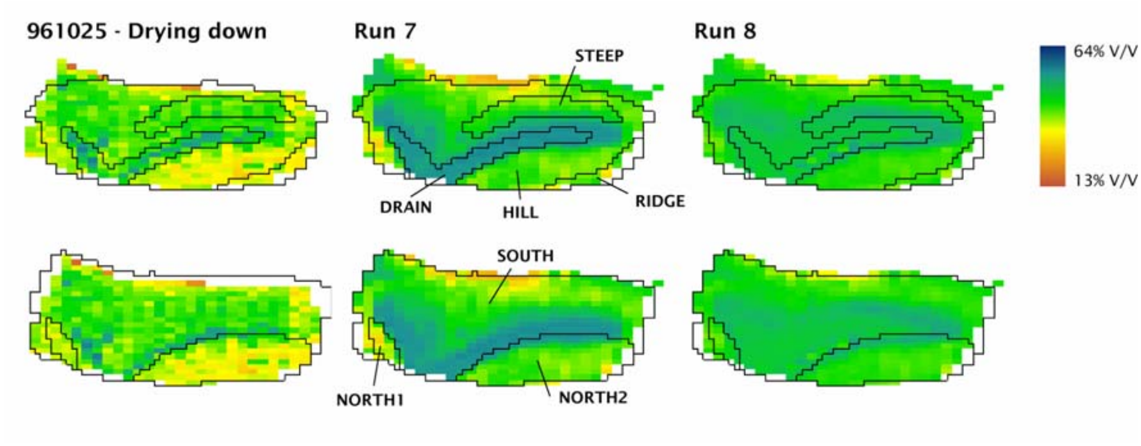


Figure 5.10 Visual inspection of the knowledge-driven regions on 961025 (drying down). Either model run 7 or 8 is the optimal model for each of the regions, although neither model performs optimally for all regions.

Western and Grayson (2000, p.234) noted that some of the observed moisture values "are likely to be slightly too high due to surface water ponding introducing some [additional] measurement errors." This is possibly the cause of the under-estimates found for each model in the drainage lines. It should be noted here that measures such as TMAE and MAE do not inform of the direction of the error. Therefore, the BIAS measure is an additionally useful summary, predominantly for this purpose.

The drying down period has the most spatial organisation and produces the most variable results for these knowledge-driven regions. In all cases, either run 7 or run 8 is the best performing model (Figure 5.10). For example, the drainage line is considered best represented by run 8, while run 7 over-estimates the moisture there. Visually, the drainage line in run 8 is not very prominent and contains the "unrealistic band of high soil moisture at the soil unit boundary" referred to in the expert assessment, but this seems to assist with matching the observed field better according to the quantitative measures. Run 7 performs much better than run 8 on the northern facing slopes. The expert assessment found that the models generally over-estimated lateral redistribution, which caused a "strong aspect bias in the soil moisture errors" on this occasion. This finding is supported by the quantitative measures, as there are much higher errors in the northern facing regions than the southern facing regions in most of the runs (particularly runs 2 and 5). This is a response to the inability of the models to correctly simulate the northern facing slopes, which are generally drier in the observed fields.

5.4.3.3 Assessment using importance regions

The use of region-based or region-specific measures allows a clear meaning to be assigned to each measure. When using multiple sets of regions, some elements may be compared multiple times as parts of different regions. If only one set of regions is used (e.g. the slope-based regions), the relative contributions of error for each parts of the field can be ascertained. When there are certain regions that perform well (or poorly), they can be directly analysed for model understanding. For more automated assessments, the relative importance of each region can be specified and their results used to select an optimum. In this example, if the slope and aspect regions are each given equal importance when combined, then runs 8 and 10 are found to be the most optimal across all knowledge-driven regions for all occasions. Run 8 was noted to contain an 'unrealistic soil moisture feature' during expert assessment, although the presence of such a feature could not be detected in this quantitative assessment (due to the elements being shared amongst other regions). Therefore, this run would not be rejected as being non-behavioural as it was in the expert assessment.

5.4.4 Overall evaluation of quantitative model assessment

Quantitative model assessment can be conducted in a more thorough manner by using a greater range of comparison measures. The tolerant measures allow clear results that are not confounded by insignificant differences to be created. The importance measures produce much more directed and useful measures that facilitate model understanding when they are logically applied. The data-driven methods can be used for feature comparison, although in hydrological model assessment this is considered to be of limited value unless the observed fields contain particularly important features that could not be related to a certain process (and thereby defined using knowledge-based methods). Finally, the multiscale measure provides a useful summary of the strict comparison measures currently used and ensures that local errors are not too influential.

These measures can be combined together using different weightings or optimisation strategies to determine the optimally performing models. Regardless of the way these measures are applied, they provide the base information on which models are assessed. The evaluation given here using expert assessment and visual evidence to confirm the findings of quantitative assessment shows that these measures are capable of emulating and quantifying some of the visual approaches used when comparing spatial fields. However, the results of the quantitative analyses do not directly agree with the expert assessment. Most noticeably, the quantitative analyses find run 8 to be the best model, as opposed to run 5 (by expert assessment). Recent discussion with the experts suggests that had more sophisticated comparison methods been available at the time, run 5 may not have been chosen as the optimal model (R. Grayson, personal communication, 17/07/2006). The problem they visually recognised in run 8 (i.e. band of higher moisture at soil boundary) led to the rejection of this model, despite the fact that it otherwise performed well. This was due to an over-emphasis on the visual appearance of the fields during comparison, which led to the potential importance of different soil types being overlooked (R. Grayson, personal communication, 17/07/2006).

The unbiased treatment by quantitative comparison methods can avoid such situations by providing clear measures describing how different aspects of spatial fields compare. Analysis of the measures then permits the rejection of poor models, while also helping to understand why other models are performing well. At the same time, visual analysis still has its benefit (e.g. feature recognition and comparison) and is not expected to be replaced. However, the quantitative comparison measures used here produce a complete, intuitive, repeatable and reportable model assessment that can benefit users greatly. The nuances with their application to hydrology are discussed in the remainder of this chapter, but overall their performance in a hydrological modelling application reveals great potential for further use.

5.5 Discussion of comparison strategy

The comparison strategy employed in this chapter for assessing the distributed model from Tarrawarra includes a large number of measures. These measures cover the three major groups of measures shown in the comparison flowchart (Figure 4.20) – strict measures, tolerant measures and region measures. Each of these groups includes a range of measures that require different parameters, pre-processing, etc. Chapter 4 gave the details of what these parameters represent and some recommendations on how to set them. This discussion focuses on the issues that arise when using each group of comparison measures for hydrological model assessment and how they can be managed.

A more general discussion about comparison strategies and making subjective decisions is given in the following chapter.

5.5.1 Strict measures

The strict measures used in the comparison strategy – BIAS, KS, MAE, RSQ – are all applied to the whole field equally and they do not utilise any of the relationships between elements within a field. These global and strict local comparisons are a useful 'first step' during comparison, as they allow the user to identify the model bias (BIAS); the overlap of the distributions (KS); the average error per element (MAE); and how well the model explains the variations in the observed field (RSQ). By adding in the multiscale measure (MSME), a more complete treatment of strict comparison across all scales (including intermediate scales) is obtained, as highlighted in the comparison flowchart (Figure 5.11). Without MSM, this group of measures would only compare the extremes of scale and could be unknowingly influenced by artefacts (e.g. noise, scale differences) in the fields. All of these measures can be calculated without the need

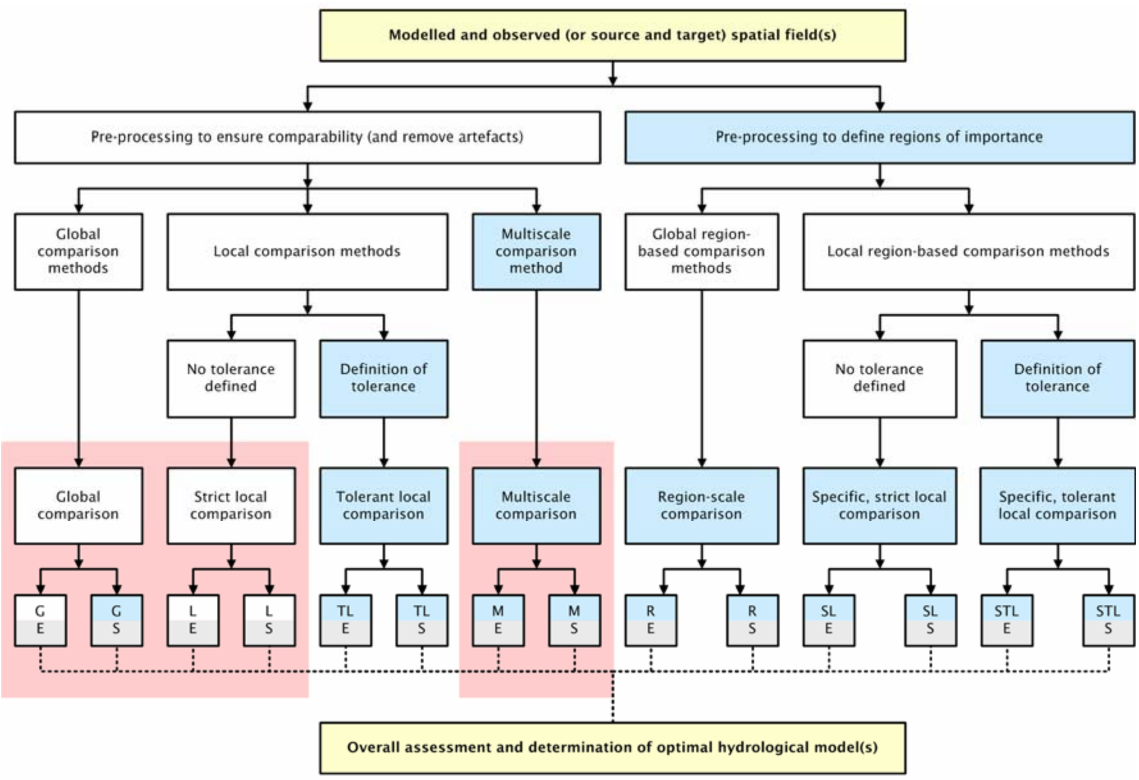

Modelled and observed (or source and target) spatial field(s)
Pre-processing to ensure comparability (and remove artefacts)
Pre-processing to define regions of importance
Global comparison methods
Local comparison methods
Multiscale comparison method
Global region-based comparison methods
Local region-based comparison methods
No tolerance defined
Definition of tolerance
No tolerance defined
Definition of tolerance
Global comparison
Strict local comparison
Tolerant local comparison
Multiscale comparison
Region-scale comparison
Specific, strict local comparison
Specific, tolerant local comparison
G E
G S
L E
L S
TL E
TL S
M E
M S
R E
R S
SL E
SL S
STL E
STL S
Overall assessment and determination of optimal hydrological model(s)

Figure 5.11 The position of the strict comparison measures within the comparison flowchart is highlighted to show how the measures fit into the overall flowchart and what pre-processing they require.

for much user control, apart from defining the range of scales used in MSME. This may be considered as a strength of the measures, but as the previous evaluation shows, with limited user input there is generally limited control over the measures and they usually fail to compare the fields as desired. These measures also treat the observed field as being fixed, thereby being unable to account for observations errors.

The multiscale measure is quite a useful measure. By using an upscaling process, it allows the effects of noise and scale differences to be minimised during comparison at the intermediate scales. These then influence the overall MSM, which is not simply the average of the local and global scales, despite the fact it often works out that way. Where models with large local error, but much lesser errors at an intermediate scale (e.g. because noise is no longer dominating) are compared, the MSM provides an important addition to strict comparison. This measure must only be calculated for the important scales, which are considered to be those up to the point where intermediate-scale variability disappears. In the example in this chapter, a simple empirical analysis of the upscaling helped to set these parameters. In larger spatial fields, the number of important intermediate scales is expected to be much higher than was necessary for Tarrawarra.

The KS similarity and global BIAS measures are both used for comparing each scale in a multiscale measure. These measures are also used on their own to describe global similarity and error. The results of model assessment at Tarrawarra highlighted a number of instances where the KS measures found high similarity, yet quite large errors or bias existed. Therefore, as with any similarity measure, KS should always be supported by an accompanying error measure. By using multiple measures, the response found can be supported or verified by the other measure and increases the confidence in interpretation. This is one of the objectives needed for completeness during comparison.

All of these measures are summaries for the entire extent of the field and their causes cannot be localised without further analysis (visually or with importance regions). Knowledge of the parameter changes between different models can be used to describe the cause of the comparison measure being better for one model than another. However, these descriptions also cannot be localised unless the parameter change only had local impacts. Therefore, the strict global, local and multiscale measures are all of limited value for model understanding.

The strict error and similarity measures are very sensitive to any location or timing errors, unlike visual comparisons. Therefore, the MAE and RSQ measures will often fail to describe what has been visually recognised. While these measures are useful for providing a strict and simple estimate of error or fit, they are all influenced by insignificant differences and are misleading. This makes them unsuitable for comparison, unless the user makes a clear assumption that there is no error in the observations

5.5.2 Tolerant measures

Tolerances are used to control the strictness of a comparison, which is desirable for removing the effects of insignificant differences on the measures (e.g. removing effects of potential uncertainties). In hydrological comparisons, tolerant comparisons are generally only used for removing uncertainty in the target (or observed) field. This is due to the purpose of comparison being to evaluate the error in the model, which encompasses the model uncertainty. If uncertainty were defined for both fields in a comparison, the resulting error or similarity cannot be attributed to either of the fields directly and is difficult to interpret and use.

Tolerances cause both the error and similarity measures to improve for all model runs in this chapter. By applying tolerances to comparison, more model runs are therefore found to produce the same measures, highlighting the fact that they have equivalent performance under the conditions specified. While the tolerances are reducing the ability of measures to discern between ‘apparently slightly different’ fields, this effect is desirable as it evaluates the fields in regard to the uncertainty that may exist within them. This is comparable to visual comparisons in which there is widespread acceptance of ‘less than exact’ matches. The tolerant methods allow this to be explicitly managed and also for different levels of tolerances to be analysed.

It is important that the tolerances defined correctly express the tolerances of the user or data. This is the only parameterisation that is needed when applying tolerances, as can be seen in Figure 5.12. It is not recommended to set tolerances too high, as the ability to differentiate between different models is reduced as tolerances increase. However, if there are sound reasons to tolerate certain differences, the tolerances should be set conservatively to permit these departures from strict agreement. For example, consider a model that is not expected to represent narrow features correctly due to the element setup. If the observed data has such features, this inability of the model will cause large errors nearby the features and produce poor results. If a location tolerance is introduced, the elements within the narrow feature are allowed to shift around within the tolerance (as needed) to produce a better match with the model. Provided the tolerance is set conservatively, the desired effect can be achieved without allowing too much dissimilarity to influence results. Tolerances could potentially be varied for each region analysed, although this requires additional parameterisation.

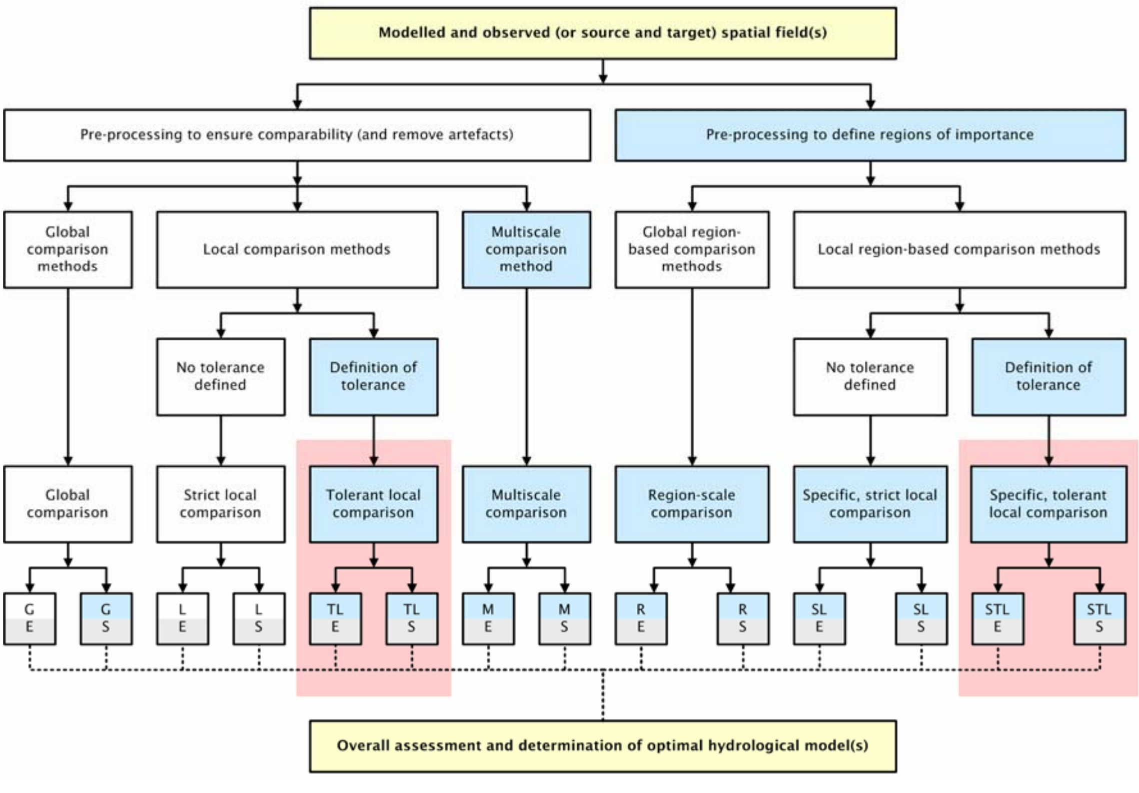

Modelled and observed (or source and target) spatial field(s)
Pre-processing to ensure comparability (and remove artefacts)
Pre-processing to define regions of importance
Global comparison methods
Local comparison methods
Multiscale comparison method
Global region-based comparison methods
Local region-based comparison methods
No tolerance defined
Definition of tolerance
No tolerance defined
Definition of tolerance
Global comparison
Strict local comparison
Tolerant local comparison
Multiscale comparison
Region-scale comparison
Specific, strict local comparison
Specific, tolerant local comparison
G E
G S
L E
L S
TL E
TL S
M E
M S
R E
R S
SL E
SL S
STL E
STL S
Overall assessment and determination of optimal hydrological model(s)

Figure 5.12 The position of the tolerant comparison measures within the comparison flowchart is highlighted to show how the measures fit into the overall flowchart and what pre-processing they require.

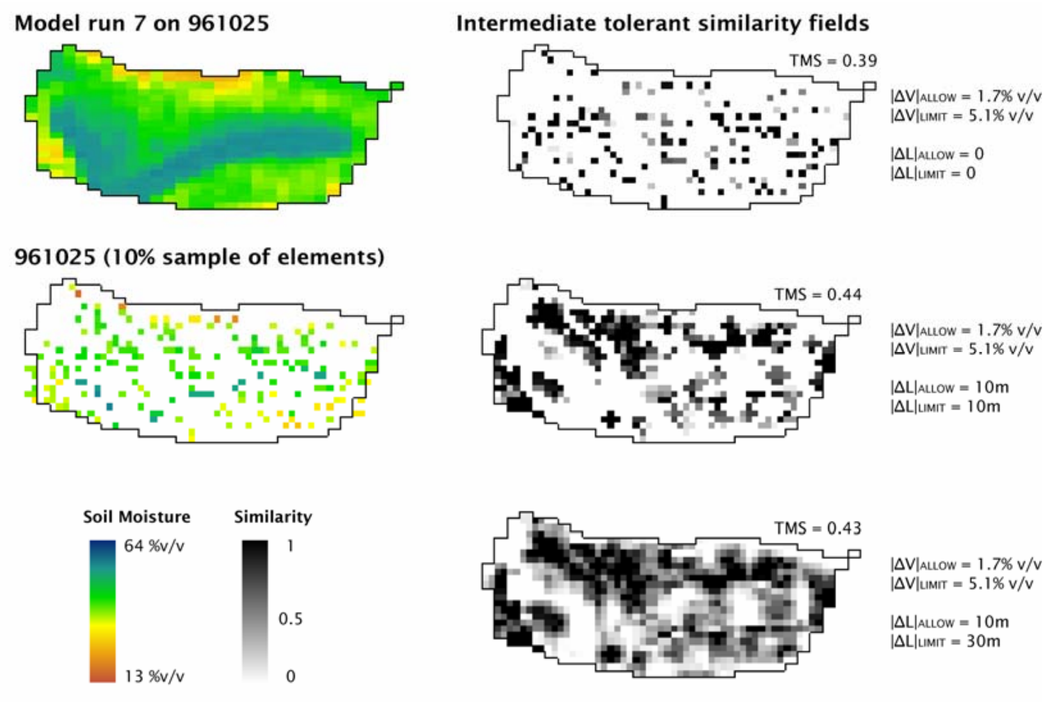


Figure 5.13 A simple example showing the situation where the observed field (bottom left) has a much more limited extent (or greater spacing) than the modelled field (top left). Here, a random 10% of elements from the observed field on 961025 are used to represent the sparse observations. The intermediate tolerant similarity fields (right column) illustrate how introducing a location tolerance allows more of the modelled field to be assessed (albeit with lesser similarity due to locational error).

Another potential use for tolerances is to permit a greater extent of a modelled field to be assessed when only limited observed data exists (e.g. sparse, but related point samples). Introducing a location tolerance can permit some locational differences and thereby allow model elements that are nearby the observed elements to be assessed, as depicted in Figure 5.13. This situation often arises in ‘data poor’ hydrological studies and provides the user with a means of accounting for relationships between nearby points and thus testing more modelled elements with the limited observations. The current method for handling this situation is to make the observed points represent a larger support by using a larger element size (i.e. so that they are spatially coincident with the modelled elements). Such an approach does not permit any notion of reduced similarity for the non-coincident elements, whereas the use of tolerances allows greater control over how the relationships are defined.

5.5.2.1 The effect of combining tolerances

The way in which value and location tolerances reduce a comparison measure varies with the fields being compared. However, there are some common responses to these tolerances that are useful to recognise. When the strict errors are mostly larger than the value tolerance, the amount of reduction in mean error will be similar to the tolerance (i.e. MAE will reduce by approximately $|\Delta V|_{ALLOW}$). When this occurs, the variance of the errors will change little. When the strict errors are mostly smaller than the value tolerance, there is a much lesser change in the MAE, but the error variance will change greatly (as many errors are now being reduced to zero). Viewing the value tolerance dimension in the surface plots in Figure 5.14 reveals this response. At location tolerance of 0, the change of TMAE starts reducing at a constant rate and then the rate decreases. However, when the errors start out small, the change of TMAE is never constant (due to all errors being less than the tolerance), as seen in the plot for 960413.

Location tolerances cause different responses depending on the arrangement of values in each field. The surface plots show that the greatest reduction in error is found by introducing a small tolerance, which allows some scope for elements to find a better match within their immediate neighbourhood. As this neighbourhood is made increasingly large, the error reduces, but will not reach zero (unless a perfect match is

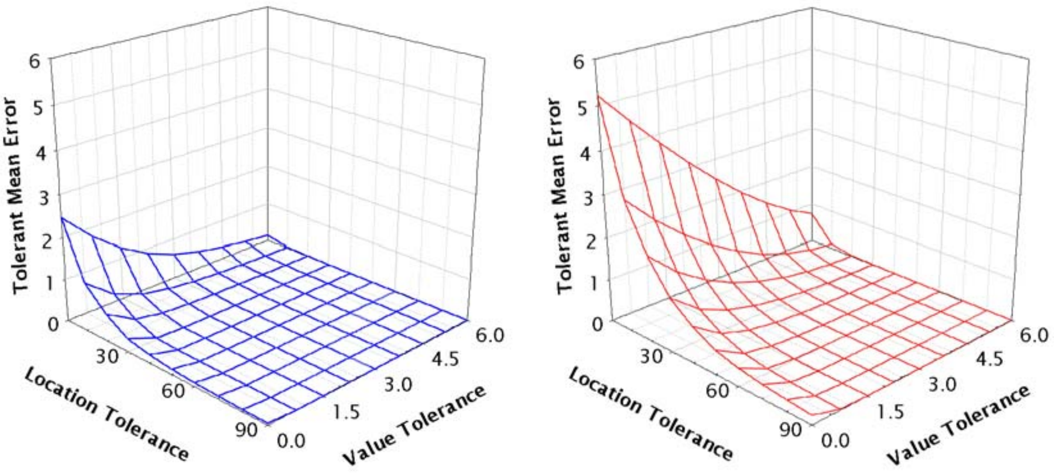


Figure 5.14 Surface plots showing the relationship between value tolerance, location tolerance and tolerant MAE on two different occasions (960413 (left) and 961025 (right)). The TMAE values are calculated between model run 8 and the observed field using different combinations of the tolerances.

found for every element). The correlation lengths of the spatial fields being compared here are approximately 40-60m (Western et al. 1998, p.30), thus showing that location tolerances smaller than this amount can still have significant impact on the comparison. However, it is by combining the location and value tolerances together that the greatest effect is seen. If both these are set quite small, they may cause the same reduction as a single tolerance that was much larger. Without combining the tolerances, the fields compared here are never found to have zero error, whereas a combination of $|\Delta V|_{ALLOW}$ = 1.5% v/v and $|\Delta L|_{ALLOW}$ = 40m would allow all errors present in these examples to be tolerated.

5.5.3 Region-based or region-specific measures

The use of importance regions is essential for: 1) enabling the comparison of 'features' within a spatial field; and 2) making specific comparison measures that can be used for both model assessment and model understanding. The definition of regions allows measures to be produced for each region, which can remove the need for visual inspection of fields to ascertain the causes of errors. The definition of regions is undertaken in pre-processing and then feeds through to the final measures, as shown in the comparison flowchart (Figure 5.15).

In all of the strict and tolerant measures, there is no consideration of whether all of the field elements are actually necessary to evaluate model performance. For example, if the model only needs to correctly simulate soil moisture content in the gullies (e.g. for saturated source area runoff predictions), then elements outside these regions may be of minimal consequence to whether the model is good. There may also be other regions within the field that are important for determining if the model is behavioural or not (e.g. if the overall spatial arrangement is logical). During model assessment, the individual regions with the highest importance can be favoured via higher weightings, determined by the specific application and purpose.

Both knowledge- and data-driven regions can be used to separate spatial fields into manageable and related regions. The knowledge-driven approach was far more successful in the model assessment conducted in this chapter, mainly because the same set of hydrologically important regions were analysed across all dates (i.e. they relate to

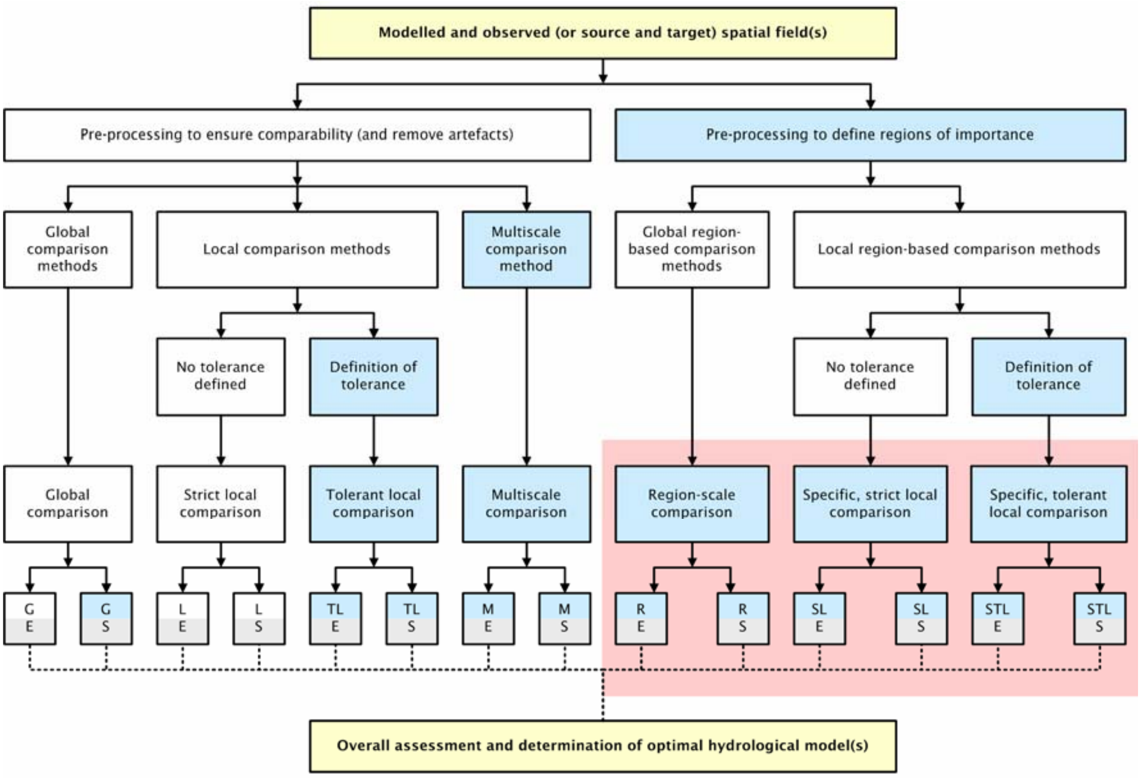

Modelled and observed (or source and target) spatial field(s)
Pre-processing to ensure comparability (and remove artefacts)
Pre-processing to define regions of importance
Global comparison methods
Local comparison methods
Multiscale comparison method
Global region-based comparison methods
Local region-based comparison methods
No tolerance defined
Definition of tolerance
No tolerance defined
Definition of tolerance
Global comparison
Strict local comparison
Tolerant local comparison
Multiscale comparison
Region-scale comparison
Specific, strict local comparison
Specific, tolerant local comparison
G E
G S
L E
L S
TL E
TL S
M E
M S
R E
R S
SL E
SL S
STL E
STL S
Overall assessment and determination of optimal hydrological model(s)

Figure 5.15 The position of the region-based and region-specific comparison measures within the comparison flowchart is highlighted to show how the measures fit into the overall flowchart and what pre-processing they require.

general aspects of the spatial fields, not the specific date being compared). In contrast, the data-driven regions were 'date specific' and were much more difficult to identify/label. This is one of the challenges when working with data-driven regions and it occurs because the regions are created using a complex algorithm that minimises heterogeneity within regions. These regions are more useful when particular features within the target field need to be tested.

The findings from comparing each region can be used to indicate possible issues with the model structure or parameters (i.e. model understanding). For example, over-prediction on the steep slopes might infer that drainage is too slow, while under-prediction may suggest it is too fast (or evapotranspiration is too high). If there is over-prediction in low slope areas, there may be insufficient drainage or evaporation occurring. If northern facing slopes are always under-predicting, then either insufficient water is reaching them or they are losing too much via evapotranspiration. Additional investigation of performance in regions of different soil type can also be useful. These results provide feedback on ‘response regions’ which can inform the user about how to improve the model, either through changes to model structure or parameters. Without these importance methods, the results from comparison are unable to be related to particular areas without visual analysis.

These are conceptually simple extensions to current methods. Clearly, the onus rests on the user to define suitable regions, which will depend entirely on their knowledge and understanding about the hydrological processes at work. As shown for Tarrawarra, many hydrological processes cause responses in specific components of the terrain and therefore can benefit from terrain analysis research for identifying suitable regions. However, terrain is not always the key surrogate field for defining suitable regions in all areas of hydrology, so users should consider other suitable surrogate fields.

When comparing the regions as features (i.e. region-based comparison), the global KS and BIAS measures are used. When the regions are quite small, KS is not very effective as it only has a limited number of elements to use to describe the distribution and generally produces low KS values (unless the regions match very closely). In contrast, the BIAS measure is more robust with small regions as it provides a directed measure of error, but it fails to describe overall similarity with larger regions. This was evident in the assessment conducted in this chapter, where KS was used for region-based assessment and produced low values for all of the small data-driven regions. It is therefore recommended to use both KS and BIAS for ‘feature comparison’, just as both error and similarity measures should be used in any comparison strategy.

Finally, the use of importance regions is different from simply analysing the correlation between an intermediate field of local errors and another surrogate field (e.g. slope). Using regions enforces spatial contiguity and therefore the regions represent spatially-related parts of the field (as recognised during visual analysis). In contrast, a correlation-based analysis and the associated scatterplot do not treat values that have both similarity in value and location as being a feature. Instead, every element with a particular slope is analysed separately, therefore removing the use of any spatial relationship between nearby values. Using regions also permits multiple sources to contribute to the definition of regions, making it far more versatile.

5.5.4 General discussion about spatial field comparison

The comparison measures evaluated here are generally used for determining the optimal model from a set of different runs. This may be to facilitate calibration, optimisation or for testing (and model understanding). In most current model assessments, a single measure (or index based on multiple measures) is used to determine the optimal model, but it is now recognised that single objective functions are rarely sufficient. Instead, approaches using multiple comparison measures are recommended for completely evaluating the optimal model. The methods evaluated and discussed here are the key to providing suitable comparison measures for spatial fields.

Comparison across a range of scales is considered an essential part of any complete comparison and this is best achieved by using a combination of different measures (e.g. global, local, multiscale and region-based). In situations where only a single measure can be used, the multiscale measure is suggested as the best ‘all-in-one’ assessment. These quantitative measures are often ambiguous in their meaning and therefore difficult to interpret. Measures that utilise the notion of importance should be used wherever possible, although when there is only limited knowledge about the situation being compared this may not be possible. Combining these with the tolerant measures allows insignificant differences to be managed during comparison and helps avoid any confounding information in the measures. The end product of these three approaches is a suite of measures that can be adapted for different scales, different meanings and different levels of strictness to provide suitable tests of spatial field comparisons. When used in model assessment they give the hydrologist control over where their model must perform well to be considered suitable.

5.6 Correlation between measures

At the most basic level, all quantitative comparison measures are based on the differences between elements. With any local comparison measure, the numerical differences between pairs of element values are calculated. The field of errors produced from this process (and variants of it) can then be transformed and summarised in various ways to produce different measures. As a result of the measures being based on the same underlying element differences, there is always high correlation between comparison measures. The tolerance and importance methods introduced in this chapter are specifically aimed at changing the field of errors produced during local comparison. In this way, they can produce measures that are based on different (or refined) local element comparisons, which subsequently measure different aspects of comparison. There will always be some correlation (or redundancy) between different measures, due to the fact that they are all comparing the same fields. However, the way that the fields are used during comparison determines what information about error or similarity is described by the measure.

When multiple, different tolerance definitions are used to produce a range of measures, there is expected to be very high correlation between the measures. Each tolerant measure is a very similar test and therefore only departs from being perfectly correlated when the tolerances results in different responses across the field (e.g. when locational shifts produce improved performance in some areas). In situations where the tolerances result in regular changes across all fields, they do not actually provide any greater ability than strict measures in discerning between models, although this would be rare in real data. It is the importance methods that generally result in the most variation between comparison measures, as they take a very different view of the underlying error field (by limiting the extents or changing the representation of elements). They produce measures that test more specific parts of the fields, which generally reveal greater variation in the relative performance of different models.

5.7 Chapter summary

The practical application of the new and improved comparison measures for model assessment has been presented in this chapter using the Tarrawarra data set. The observed and modelled fields from Tarrawarra on different occasions (e.g. spring, autumn) are considered to be representative of the range of fields that are generally used in hydrological modelling, thereby availing this example of broad relevance. This data has also been previously analysed by experts (Western and Grayson 2000), which provides a qualitative benchmark against which the quantitative performance of the measures can be evaluated.

The new measures have been applied using a comparison strategy that achieves the general objectives of comparison set out in Chapter 4 – comparison of multiple scales; significant error and similarity measures; and application-specific comparison tests. This strategy produced a large number of quantitative measures that describe how different aspects of spatial fields compare. For each measure, the optimal results (or particularly poor results) have been identified and are related back to the qualitative findings from expert assessment and also visual evidence that supports the quantitative results.

The methods applied to the whole spatial fields (i.e. not specific regions) produced results that agreed with certain aspects of the expert assessment, although the expert assessment is not necessarily correct (i.e. it is only one opinion) and the assumptions made were generally undefined. The tolerances were effective at reducing the minor value and location differences that were overlooked in the expert assessment. The measures that best aligned with the notes made in the expert assessment were the knowledge-driven region methods, which effectively represented hydrological response regions. This clearly demonstrated the value of including suitable regions via pre-processing to facilitate improved quantitative model assessments and also enable model understanding without the need for visual analysis.

Overall, the quantitative analysis using the new and improved methods found run 8 to be the optimal model. This was in contrast to the expert assessment, which rejected run 8 as being non-behavioural due to the presence of a 'soil boundary' feature. In this situation, the bias in the expert assessment appears to have weighted this factor more highly than many other positive comparisons for run 8. This may also have resulted in alternate soil parameterisations (e.g. a soils maps with gradual transitions) being overlooked. This application gives strong evidence of the importance of quantitative, adaptable and repeatable comparison measures.

Chapter 6

Discussion

6.1 Chapter overview

The quantitative comparison methods presented in this thesis require considerable thought by the user to define what is being tested by the spatial field comparison. They are both quantitative and repeatable, while still utilising expert knowledge of the user in their design. The development and application of each comparison method has been shown throughout Chapters 4 and 5. This chapter discusses the general issues with the application of new comparison methods.

The chapter begins by reviewing the general approach that is currently used for comparison (and subsequently model assessment). The developments made throughout this thesis are briefly discussed in regard to how they can prompt a change to this approach. This is followed by a discussion of the major issue raised in applying the new comparison methods – requiring the user to make subjective decisions. The new measures provide the user with simple ways to control the comparison, whereas current measures are generally ‘black boxes’ that remove the user from the comparison process. Each type of subjective decision made enables additional capabilities during comparison that are outlined here. The way that these decisions combine to devise a comparison strategy for any specific application is discussed and then illustrated using a hydrological example.

Issues relating to how the comparison measures are analysed, such as the interpretation and significance of differences between comparison measures, are then considered. Finally, the issues relating to implementing these new comparison methods (i.e. software development, computational demands) are discussed.

6.2 Changing the approach to comparison

Comparison of data sets is fundamental to model assessment and other tasks in hydrology. Many of the situations faced with comparison therefore directly flow through to the model assessment process. For example, when meaningful comparisons can only be made using qualitative methods (e.g. visual inspection of plots or fields), the model assessment is similarly qualitative. If the quantitative comparison methods only provide generalised tests of comparison, then the model assessment can only be based on such information. These two examples describe the current state of comparison and model assessment in hydrology. Some advances have been made for time series comparison and assessment of temporal models, although the spatial comparison problem in hydrology has had limited attention prior to this thesis and some recent publications (Grayson et al. 2002; Jetten et al. 2003; Güntner et al. 2004; Western et al. 2004; Wealands et al. 2005b).

Rykiel (1996, p.242), when reviewing the state of model assessment for ecological models, points out that "performance testing is fundamentally limited to showing that the model passes the validation tests devised for it." However, "modellers do not describe in an engineering sense the requirements [and] specifications of the models they build. Because there are no generally agreed validation criteria for ecological models, the best that can be done at present is for the modeller to state explicitly what the validation criteria are and leave it to the user to judge if the criteria are adequate. The most common criteria at present are the 'see how well the simulated data matches the observed data' test and the 'the model did a reasonable job of simulating…' test in which the reader is asked to agree subjectively that the match is adequate." These findings are largely applicable to most current hydrological model assessments.

The comparison measures presented in this thesis have been developed and are expected to help address this situation in the following ways:

- Users are now able to rigorously describe the measures and parameters used for comparison, rather than relying on descriptive language to describe performance. This permits comparisons to be repeated and clearly reported, although the decisions made still need to be rationalised.

- Users are now able to quantify more meaningful tests of how spatial fields compare. This can help the user to avoid qualitative comparisons for model assessment, or at the very least supplement them with numerical evidence of performance.

The measures require the user to take greater control of the quantitative comparisons made in hydrology, which is a substantial change from current practice. Currently, the user is able to utilise their knowledge, but it is rare for this knowledge to be captured into the comparison process. It is expected that having greater access to comparison methods that emulate the qualitative approaches commonly used will lead to improved model assessment.

6.3 Devising the comparison strategy

The comparison methods developed in this thesis are not designed to be applicable to all situations, but they are suitable for most typical situations in hydrology and can be readily adapted for other situations. The use of these measures together for any application is determined by the comparison strategy, in which subjective decisions are required to make each measure specialised for hydrology. The following sections discuss what the major subjective decisions are and how they come together to control the aspect of spatial fields being compared (and the associated interpretation of the measure). The overall comparison strategy can then be devised by using different combinations of these decisions to achieve the major objectives of comparing multiple scales; describing both error and similarity; and comparing specialised aspects of the spatial fields.

6.3.1 Subjective decisions

There is no such thing as a purely objective comparison measure, because a subjective decision has been made to use that measure in the first place. Also, the underlying nature of making a comparison is actually subjective, because there are so many different aspects of a field that can potentially be compared. This thesis presents quantitative comparison methods that manage the subjective decisions in a rigorous manner, rather than purporting to be purely objective. This allows the decisions to be clearly stated and the resulting measures to be interpreted appropriately. The decisions only need to be made once to parameterise the comparison method, after which the quantitative measures can be produced for many different spatial fields.

Subjective decisions can be made explicitly to produce a specialised measure, or otherwise they may be made implicitly by choosing an existing method. For example, choosing to compare fields using a BIAS measure makes the following set of implicit decisions – it measures error using global scale comparison of the whole field with no tolerance for uncertainty. When explicit decisions are made regarding the scale, focus and strictness of comparison, measures with specialised hydrological interpretation can be produced. The following sections describe the specific decisions involved when producing a spatial comparison measure.

6.3.1.1 Error versus similarity methods

Error and similarity measures (as defined in section 1.6) should both be used when investigating how spatial fields compare, as they perform the comparison in distinctly different ways. An error measure accumulates all differences between fields, with large differences having a major impact. Error measures are generally small for similar fields and large for different fields. They must be evaluated against background knowledge (e.g. expected measurement errors) to be correctly interpreted (i.e. is the performance good or bad). In contrast, similarity measures accumulate any similarities (as defined by the user) while overlooking the differences (i.e. they are not influenced by outliers). They are limited to a specific range (e.g. 0 to 1), which makes it possible to evaluate them without knowing the nature of the data. However, understanding the variability and expectations of the data will help to improve their interpretation. The subjective decision of whether to produce error or similarity measures should favour using both together, which is supported by the findings of Legates and McCabe (1999) with time series comparison. This ensures the user knows not only how closely the fields match, but also how bad the errors are when they do occur.

6.3.1.2 Scale of comparison

The scale of comparison directly changes how a measure should be interpreted. For example, a global scale comparison treats the field (or region) as a single element that is characterised and then numerically compared. In contrast, local comparison compares each element within the field (or region) and then summarises the results. This allows the variability to influence the measure, which is not necessarily desirable if performing a less detailed comparison (e.g. general feature comparison). Local scale comparisons respond to differences in variability and arrangement, whereas global scale comparisons do not. There is a default scale for most comparison methods (e.g. MAE works at local scale), but if an upscaling method is chosen this enables any (or multiple) larger scale(s) to be used. This decision is often limited by the comparison method, but using a multiple scale method is considered valuable whenever a single summary measure is needed. In most situations, using a range of methods that apply to different scales (e.g. one local scale, one intermediate (or region-based) scale, one global) is preferred so that the variation across scale can be understood.

6.3.1.3 Regions of comparison

Defining spatially contiguous regions, using the spatial fields or ancillary information (e.g. terrain model), is the most powerful way of changing the meaning of a comparison measure. This decision is commonly overlooked and the entire fields are compared, which can cause ambiguous measures to be produced (i.e. comparisons in unimportant parts obscure the comparison desired). Comparison regions must be defined and labelled using user knowledge or characteristics from the ancillary data to assign them a specific meaning. When comparing hydrological fields, the definition of suitable regions will improve with experience and knowledge of the processes being modelled (e.g. for soil moisture at Tarrawarra, the slope-based regions were most useful). A variety of spatial analysis tools (e.g. terrain analysis, segmentation) exist for translating this knowledge into the set of region definitions.

As an alternative to using ancillary data, the regions that exist within the target field can be recognised using segmentation. The region-growing method used in this thesis can be applied to any type of spatial field by adjusting the scale and shape parameters, making it particularly useful for this type of investigation. It identifies optimally homogeneous regions that can then be used to perform a type of feature-based comparison (i.e. the similarity of each observed region). In complex fields, a large number of regions may be identified. The user may analyse these individually, or group them together into a summary measure for all regions of a particular type (e.g. the average error of all hillslope regions). The important point is that the user assigns hydrological meaning by introducing additional data. These comparison methods encourage this user input and ensure it is managed in a rigorous manner.

6.3.1.4 Tolerances

Subjective decisions relating to the definition of tolerances allow the strictness of local comparisons to be controlled. Basic value and time tolerances can also work with global comparisons. Tolerances encourage the user to explicitly consider uncertainty within the fields being compared. Considering uncertainty continues to be called for throughout the hydrological literature, although there is no 'Code of Practice' for incorporating this into analysis (Pappenberger and Beven 2006).

Whenever possible, tolerances should be set based on user knowledge about uncertainty (e.g. measurement error, positional accuracy). Where this is not possible, user experience can be used to assign suitable values. The different components of tolerance (i.e. value, location and tolerance) have been shown to depend on each other, so they should be set conservatively if they are not based on user knowledge to avoid tolerating too much difference. A decision about whether tolerances are applied to just the target field (e.g. uncertainty in only the observed field) or both the source and target (e.g. uncertainty in both model and observation) is also required. When using observed fields to evaluate a model, only the uncertainty in the observation should be tolerated. This makes sure the resulting measure represents all the model error (including uncertainty).

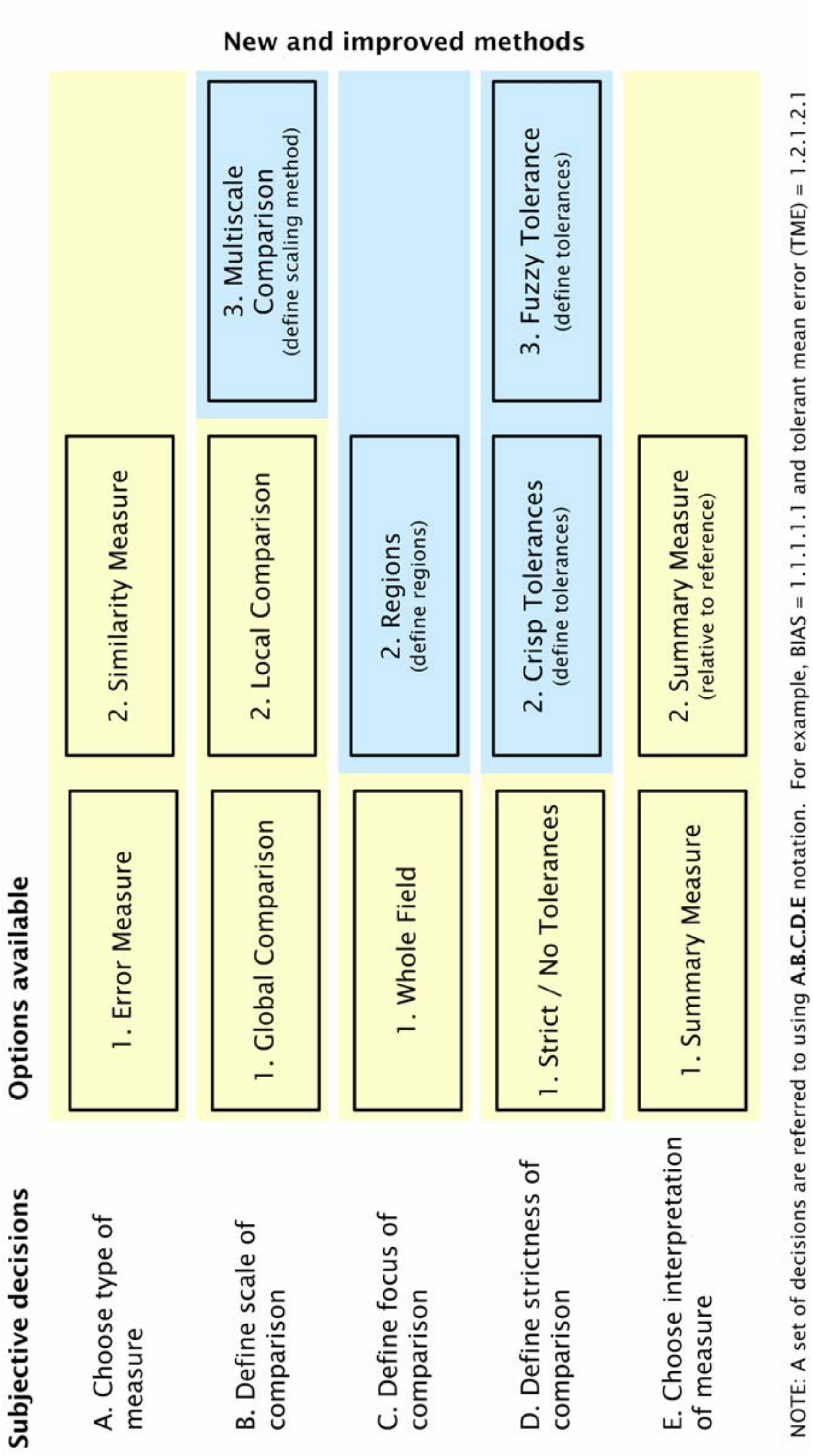
Tolerances allow further customisation to enable highly specialised comparisons. For example, value tolerances that limit the direction of the difference can be used to only tolerate under-prediction. Location tolerances may be specified to only consider connected elements that are upslope (i.e. determine from a terrain model) and within a certain distance. These specialised tolerances are additional decisions available to a user (if required) during the comparison process.

6.3.2 Combining the decisions

The subjective decisions combine to produce specialised comparison measures. The interpretation of these measures is a direct result of the decisions made. Figure 6.1 shows the four groups of subjective decisions described here, as well as a final group that determines whether the measure represents relative performance. One option is chosen (and parameterised if needed) from each group to produce a specific comparison measure. The shaded options identify the decisions that must be explicitly made for the new measures to be used. Users are expected to use pre-processing tools to translate their knowledge into either the regions or tolerances that parameterise their decisions. The type of user knowledge included determines whether the resulting measure has hydrological meaning or not.

Each combination of decisions will produce a measure that belongs to one of the groups of measures identified in the comparison flowchart (Figure 4.20). The informative combinations for hydrological comparisons are listed in Table 6.1 along with the following details for each combination: 1) the associated measures; 2) a description of how the measure is used; and 3) any limiting assumptions.

Any of the measures listed in Table 6.1 could be used individually, but combining multiple measures into an analysis provides a more comprehensive comparison and is the recommended approach for ensuring completeness. The different measures can each quantify different aspects of similarity (e.g. overall model error under uncertainty, similarity in gullies, similarity on hilltops, etc.) and can potentially be combined into automated optimisations to reveal the optimal models (assuming the model space is broad enough). This is a major change of approach to what is currently taken in hydrology, but it is necessary to advance the science of spatial prediction.



Subjective decisions	Options available		New and improved methods
A. Choose type of measure	1. Error Measure	2. Similarity Measure	
B. Define scale of comparison	1. Global Comparison	2. Local Comparison	3. Multiscale Comparison (define scaling method)
C. Define focus of comparison	1. Whole Field	2. Regions (define regions)	
D. Define strictness of comparison	1. Strict / No Tolerances	2. Crisp Tolerances (define tolerances)	3. Fuzzy Tolerance (define tolerances)
E. Choose interpretation of measure	1. Summary Measure	2. Summary Measure (relative to reference)	

NOTE: A set of decisions are referred to using **A.B.C.D.E** notation. For example, BIAS = 1.1.1.1.1 and tolerant mean error (TME) = 1.2.1.2.1

Figure 6.1 The subjective decisions (made either implicitly or explicitly) combine together to define the meaning of the resulting comparison measure. By incorporating user information (via the blue shaded options), specialised hydrological comparison measures can be created.

Table 6.1 The combinations of subjective decisions that are informative for hydrological comparisons. Each combination includes a list of the associated measures; the broad group they belong to in the comparison flowchart (Figure 4.20); a description of how the measure is used; and details of any limiting assumptions. The shaded measures are only possible by making subjective decisions for the options introduced in this thesis.

List of decisions	Group - Measure(s)	Use within hydrological modelling and limitations
1.1.1.1.1	G E - BIAS	Used for assessing over- or under-prediction of a hydrological model; treats entire field as a single element
1.1.2.1.1	R E - BIAS per region	Changes field structure from elements to regions and then compares; used for assessing how much particular regions of hydrological significance differ (e.g. error for each low slope region); assumes no variability within the region; should only be applied within homogeneous regions (e.g. those from data-driven segmentation)
1.2.1.1.1	L E - MAE - RMSE	Strict comparison, so similar nearby elements are not considered; if local errors are squared it produces RMSE; used for assessing the average error per hydrological model element (i.e. model error); requires spatial and temporal coincidence of elements before comparison can be meaningful
1.2.1.2.1 1.2.1.2.2	TL E - TMAE - TCOE	Tolerant comparison allows for uncertainty in fields to be tolerated; error measure is dependent on tolerances (i.e. larger tolerances, lower errors); explicitly manages uncertainty at element scale; used for assessing significant hydrological model errors; can be relative to known reference (e.g. mean field) to facilitate inter-comparison of measures;
1.2.2.1.1 1.2.2.2.1	SL E - MAE per region STL E - TMAE per region	Unlike region-based comparisons, this does not change the field structure into regions before comparison; makes measures more focussed; used for assessing model error in specific regions of hydrologically significant elements (e.g. error on northern facing hillslopes) with or without accounting for uncertainty
1.3.1.1.1	M E - MSME_K	Combines comparisons from global, intermediate and local scales together; comparisons at intermediate scales introduce some degree of tolerance (due to smoothing); can reveal scale differences if used diagnostically; used for assessing error for all important scales in a single holistic measure
2.1.1.1.1	G S - KS	Measures difference between cumulative distributions to describe similarity of elements in fields; used to assess general fit of hydrological model (treated as single element);
2.1.2.1.1	R S - KS per region	Changes field structure from elements to regions and then compares; compares distribution of elements within each region; used to assess how similar particular regions of hydrological significance are (e.g. similarity for each gully region); for small regions, high similarity is difficult to achieve due to limited overlap between distributions
2.2.1.1.2	L S - RSQ - COE	Transforms local errors into similarity by using a reference (e.g. linear fit, observed variance); widely used but not specialised for hydrological comparison; currently used to assess hydrological model performance relative to a known reference; implicit assumptions are made that need to be understood for correct interpretation, but these measures are well-understood in hydrology

List of decisions	Group - Measure(s)	Use within hydrological modelling and limitations
2.2.1.3.1 2.2.1.3.2	TL S - TCOE - TMS - TSE	Fuzzy tolerances used to define the differences considered as similar; explicitly manages expectations of model in terms of value, location and time accuracy; used for assessing level of similarity between hydrological model and observation
2.2.2.3.1 2.2.2.3.2	STL S - TCOE per region - TMS per region - TSE per region	Uses region definition to focus comparison and provide unambiguous measure for specific elements; used for assessing similarity between elements in hydrologically significant regions (e.g. similarity within alluvial soil regions); in small regions, high similarity values can be difficult to obtain due to limited sample
2.3.1.1.1	M S - MSMS_K	Gives holistic measure of similarity (using KS) across local, intermediate and global scales (as defined by user); used for assessing similarity in a single holistic measure

6.3.3 An example comparison strategy

It is not necessary to calculate every comparison measure listed in Table 6.1 to comprehensively compare spatial fields. Instead, there are a number of methods that may be more suitable for some applications than others. The most suitable strategies for different applications are expected to emerge via user experience. However, until such experience is available, the general objectives for a comparison strategy (from Chapter 4) should be considered when faced with a situation requiring comparison.

This section describes a comparison strategy that is devised for a situation in which a numerical weather prediction model is compared with observed RADAR rainfall fields. This example is given to illustrate how the methods presented in this thesis can be easily adapted to different hydrological situations (as opposed to the soil moisture situation in Chapter 5). The actual comparisons and model assessments are not fully undertaken, but some details of the results are given for illustrative purposes.

6.3.3.1 The rainfall fields to be compared

Rainfall RADAR is one of the few methods used in hydrology that produces an observed spatiotemporal series with short temporal spacing. Using this data for model assessment can permit more specialised questions to be asked about how the model and observations compare. The example given here is for 6 hours of observed rainfall RADAR aggregated over a 15 minute period (i.e. 24 fields) for a 256km x 256km area (with regular 2km x 2km elements) covering the south-east of the United Kingdom (A.

Seed, personal communication, 04/05/2006). A numerical weather prediction (NWP) model was used to simulate this same time period and it is being compared against the RADAR observations to measure model performance. The modelled fields have been produced at a comparable scale to the RADAR observations.

6.3.3.2 The suggested comparison strategy

As presented in Chapter 5, the comparison strategy devised for this example is detailed in a table (Table 6.2) and then the rationale for the various decisions is given. As the observed data in this situation is a spatiotemporal series, the comparison can also make use of time tolerances. The strategy listed here meets the objectives set out in Chapter 4 for achieving completeness during comparison.

This strategy has been applied to assess 6 different time steps from the modelled series (shown in Figure 6.2). The results from making the comparisons are listed in Table 6.3 and are referred to in the following section.

Table 6.2 The list of comparison measures used in the comparison strategy for model assessment of a numerical weather prediction model using RADAR rainfall fields. The pre-processing, parameters and a description of what the measure is expected to assess about the models is given for each measure.

Group	Measure(s)	Parameters/Pre-processing	Description
G E	BIAS	-	Over- or under-estimation of rainfall averages but is representative of difference in total rainfall (i.e. it is just divided by number of elements)
G S	KS	-	Similarity of element value distributions between model and observation
L S	RSQ	-	Amount of observed variability that is represented by the model
R E	BIAS	Define two regions in RADAR fields – elements with rain; and elements where no rain is observed	Over- or –underestimation for ‘true rain’ and ‘false rain’ parts of the model
TL E	TMAE	$\|\Delta V\|_{ALLOW} = 0.1$ mm $\|\Delta L\|_{ALLOW} = 1$ element (2km)	Significant error per model element
TL E	TMAE	$\|\Delta V\|_{ALLOW} = 0.1$ mm $\|\Delta L\|_{ALLOW} = 1$ element (2km) $\|\Delta T\|_{ALLOW} = 15$ minutes	Significant error per model element, with inclusion of time tolerance for mismatches between overall fields

Table 6.3 Results from applying the comparison strategy for assessing the numerical weather prediction model against observed RADAR rainfall. Refer to Table 6.2 for details of the region definitions and different tolerance (VL and VLT) definitions. The best matching time step (i.e. lowest BIAS value) has been noted to show whether there was a regular timing difference between the models.

FIELD	BIAS mm	KS	FALSE RAIN BIAS mm	TRUE RAIN BIAS mm	TMAE VL	TMAE VLT	TIME MATCH
T = +1 hr	0.44	0.73	0.24	0.78	0.47	0.42	T = +0.75 hr
T = +2 hr	0.40	0.71	0.35	0.48	0.51	0.51	T = +1.75 hr
T = +3 hr	0.16	0.72	0.33	-0.17	0.63	0.59	T = +2.75 hr
T = +4 hr	-0.07	0.78	0.25	-0.79	0.49	0.49	T = +4 hr
T = +5 hr	-0.17	0.86	0.25	-1.05	0.48	0.48	T = +5.25 hr
T = +6 hr	-0.19	0.93	0.15	-0.91	0.36	0.33	T = +6.25 hr

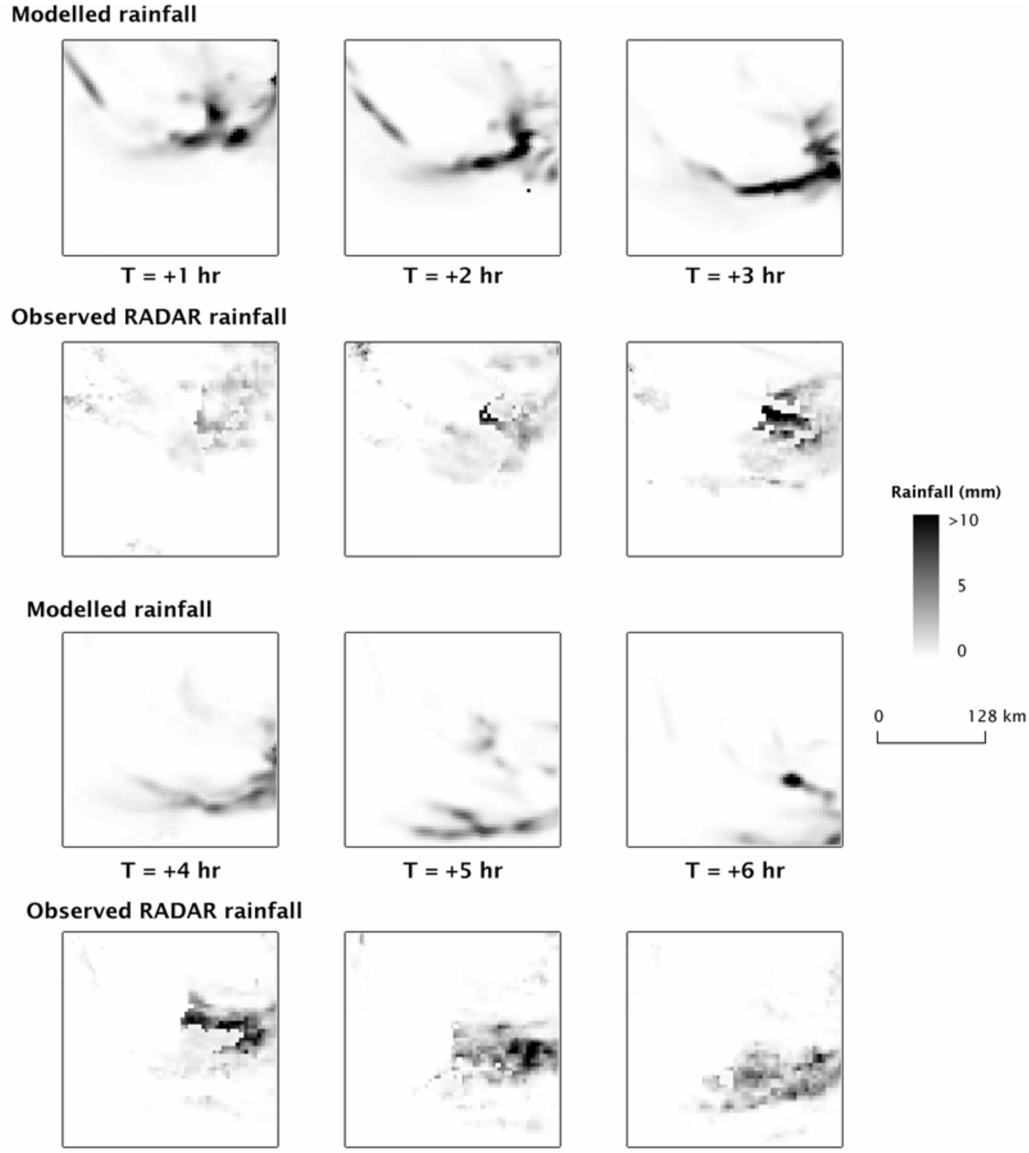


Figure 6.2 The spatial fields from the numerical weather prediction model and RADAR rainfall observations. Six temporally-coincident pairs of spatial fields are shown.

6.3.3.3 Rationale for strategy and parameters chosen

Initial comparison of rainfall fields would generally use global measures of BIAS and KS similarity to describe whether the overall amount of rain in the field is correct and whether the distributions agree. The distributional measure is particularly interesting for rainfall fields, as they are usually heavily skewed by the large number of 'zero rainfall' elements. This can unduly influence the error measures if not considered.

To obtain a measure that considers the 'no rainfall' elements, importance regions are recommended for use. The rainfall fields could be divided into regions using any of the methods introduced in this thesis, although the simplest approach in this situation is to threshold the observed rainfall fields at 0mm. This separates them into areas where rain actually occurred (i.e. the model would be estimating 'true rain') and where it didn't (i.e. the model would be estimating 'false rain'). The resulting BIAS measures calculated for each region describe how well/poorly the modelled fields estimate rain and whether this is due to false positives or other causes. For example, Table 6.3 shows that the model has a similar amount of 'false rain' at each time, but changes from over-estimating to under-estimating 'true rain' at around T = +4 hr.

The above comparisons provide global and region-scale comparisons, but local comparison of the fine details is also necessary for a complete comparison. Tolerant local measures are favoured here, as they allow any uncertainty in the RADAR fields to be considered and investigated. A subjectively-assigned value and location tolerance is used first, to summarise only the significant differences between the fields. RADAR fields are typically quite uncertain, so the small $|\Delta V|_{ALLOW}$ of 0.1 mm is likely to be much larger in a more realistic comparison. There is also a difference in the temporal support of the numerical weather model versus the RADAR data which would lead to errors. These could be included into a value tolerance or else as a time tolerance on the observed RADAR field. An additional set of tolerances, in which the model is permitted to have a 1 time step difference with the RADAR fields, is used to see if a better local match is found nearby (i.e. as evidence of timing problems in the model). Table 6.3 shows that early in the series, the modelled rainfall may be moving faster than the observations, while in the latter times the model has fallen behind.

Further analysis of these fields is not undertaken here, but the strategy shows how readily adaptable the comparison measures from this thesis are for an alternate hydrological situation. They are expected to have widespread applicability to continuous fields.

6.4 Analysis issues

Analysis of the results from comparing spatial fields can be managed in a variety of ways. The measures are generally analysed to evaluate the optimally performing model (or parameter set) from a group of potential models. The method used to determine the optimum can include approaches such as multi-criteria optimisation or rejection of non-behavioural models. Both of these methods have already been discussed in this thesis. However, regardless of the optimisation methods used, there are some issues with how to interpret and analyse the measures that must be considered.

6.4.1 Interpreting comparison measures

Comparison measures have limited value if there is no user knowledge applied when interpreting them. They are simply summaries of the accumulated errors or similarities between two (or more) sets of 'meaningless' numbers. However, this situation is uncommon because the user usually understands the spatial fields they are comparing. The primary knowledge that facilitates interpretation is the 'type' of the spatial fields being compared. It identifies the actual attribute (e.g. rainfall), the units of the resulting error measures (e.g. rainfall in millimetres) and also the characteristic scales of the field elements (e.g. daily rainfall for each 100m x 100m square element). This knowledge applies to all comparisons conducted between the fields, unless they are pre-processed to change their meaning. In some studies, there may be limited additional knowledge about the fields apart from their basic context. In other studies there can be much more ancillary information available (e.g. spatial distribution of topography, climatic information).

In addition to the type of fields being compared, knowledge of the processing involved in the comparison method can be used to aid interpretation. This is where the common methods used in hydrology often fail. The comparison methods used to produce measures such as RMSE, BIAS, COE and RSQ are not interactive. Instead, they are a ‘black box’, into which two data sets are put and a quantitative measure is produced. When the user is not encouraged to make decisions, they tend to blindly apply the methods without thinking about the implicit decisions/assumptions being made. Legates and McCabe (1999) showed that the methods to produce RMSE and COE could be easily modified to require users to control how the comparison is done (i.e. whether absolute or squared errors are used). This type of interaction makes the user consider what is actually being compared and enhances their ability to correctly interpret the results.

The RMSE measure is often misinterpreted as measuring the performance of a model for simulating high magnitude events (e.g. Sun et al. 2000). This method does not specifically compare high magnitude elements, nor does it treat them any differently to others. While RMSE does cause large errors to be more influential (i.e. by squaring them) in the resulting measure, it does not do so unambiguously and can be dangerous to interpret in this way. A more appropriate comparison would require the user to specify a threshold for defining high magnitude events, which are then used as a subset (i.e. a non-contiguous region) for comparison, as done in time-series comparison by Boyle et al. (2000). Explicitly making this decision makes the measure better reflect what the user wants to test, while still being quantitative and repeatable. All that is required is the addition of user knowledge. With spatial fields, the common methods do not produce easily interpretable measures, due to the confounding effects of spatial variability (as in the analysis throughout Chapter 5). Therefore, useful and meaningful spatial comparison measures are particularly reliant on user interaction.

6.4.2 Significance of differences between measures

When a comparison measure is produced between a range of potential models and an observed field (i.e. reality), the resulting measures are generally similar in magnitude. This occurs because many model parameterisations cause only minor changes and this feeds through to the comparison results. However, evaluating whether the performance of a model with similarity of 0.80 versus one with similarity of 0.82 is significantly different demands greater consideration.

Spatial field comparisons are influenced by the actual element values and their spatial arrangement within the fields being compared. They can be considered synonymous with landscape indices, which are also dependent on the composition and configuration of the field being analysed. A recent discussion about significant differences between landscape indices by Remmel and Csillag (2003) recognised that, unlike spatial statistical models in which the distributions are known, the distributions of landscape indices (and spatial field comparisons) are not known. Therefore, the expected values and variance cannot be used to evaluate statistical significance.

Bootstrap methods (Efron 1981) have been applied to try and estimate the distribution of certain landscape indices (e.g. Hess and Bay 1997), but these methods have only considered the element values, thereby ignoring the spatial autocorrelation present in most spatial fields. Remmel and Csillag (2003) have presented a method to estimate confidence intervals for landscape indices, using multiple realisations of different binary landscapes (created with a range of value distributions and arrangements). An empirical distribution is created from these realisations, from which confidence intervals can be determined for specific combinations of composition and configuration. This allows significance to be determined, but relies on the realisations being representative of all the possible landscapes, which is difficult to ensure.

For spatial field comparisons, the summary measure is generally determined from an intermediate error or similarity field (or multiple fields as with multiscale comparisons). The measure is therefore dependent on the initial arrangement of both fields and also on the comparison method used. This introduces additional complexity into estimating the distribution of a measure and makes significance testing problematic with continuous fields. This is an area that deserves further research, but it has not been attempted here because of the complexity in generating feasible realisations of continuous fields, as opposed to the binary landscapes attempted by Remmel and Csillag (2003).

Any analysis and interpretation of a comparison measure should be made relative to other measures that use a similar method and data (so that some ‘feeling’ for the distribution is available). It is not valid to analyse and contrast measures obtained from different data sets or times without having a reference to evaluate the measures against.

6.4.3 Relative comparison measures

Comparison measures should be analysed relative to some reference value to aid with their understanding. This can be incorporated into the comparison method, as is the case with the efficiency measures used throughout this thesis (e.g. COE, TCE, TSE). These measures all use the performance of the observed mean (or mean spatial fields) to standardise the values. When using measures that are not adjusted to a known reference, the reference can be considered during the analysis in a less explicit manner. By contrasting each measure against the results from other fields being compared, the user can understand the distribution of the measure for the particular data sets. This approach has been used throughout Chapters 4 and 5, in which the 'stand out' values from the result tables are highlighted and discussed. When the complete distribution of the measure cannot be known and it is not assumed, these are the most suitable approaches to analysis.

Relative comparison measures, based on evaluating the error or similarity relative to a known reference, are usually restricted to a certain range of possible values (e.g. zero to one). A value of zero may represent no similarity, or it may represent the similarity of a featureless field with the correct mean (as with TCE and TSE). Regular and/or widespread use of these measures with different data sets enables a user to understand their distribution via experience. The COE measure is an existing example from hydrology, whereby most hydrologists would have an expectation about what a 'good' COE value is for a certain situation. It is expected that similar experience will develop for more specialised spatial field comparison method also.

6.5 Implementation issues

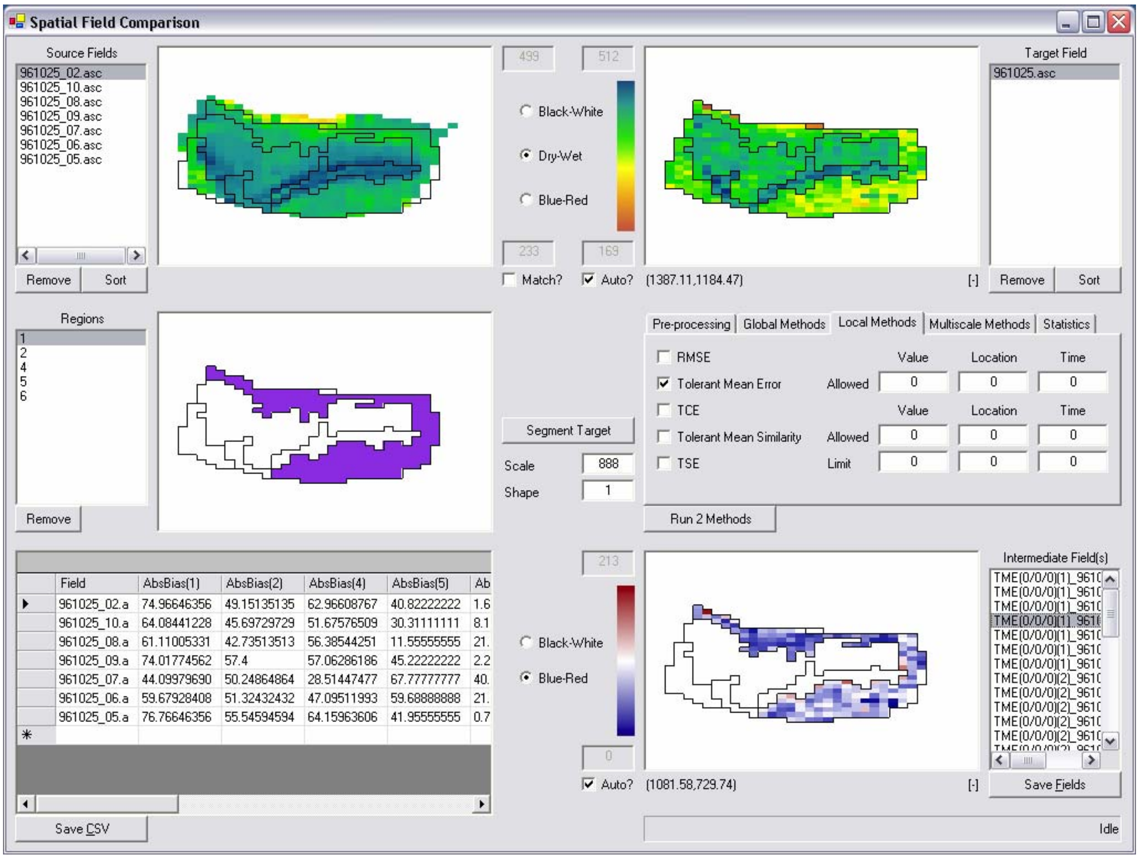
Implementation of specialised spatial field comparison methods requires a computational system to manage element values/locations/times. The common methods used in hydrology only compare spatially and temporally coincident elements, avoiding the need to manage location and time attributes. This makes comparison as simple as summarising the differences between pairs of data values (e.g. in a spreadsheet). When location and/or time are considered, a more advanced method of handling the data is required. GIS software can be used for conducting simple spatial comparisons, but the analysis tools within GIS are only suited to general analyses. They manage time information using either an attribute or the name of the spatial field (or layer). To undertake custom or complex analyses (i.e. implement a prescribed algorithm) with GIS software, a certain amount of programming or scripting is usually required. GIS software generally provides a set of programmable objects that can be manipulated and combined to undertake these more advanced comparisons.

To encourage users to apply new comparison methods, custom software tools that enable easy implementation are desirable. This is particularly relevant for the more complex methods that introduce tolerances and multiscale representations. These methods are not just simple combinations of standard GIS operations. Instead, they are custom algorithms that require a custom solution. With the simpler, strict comparison methods (e.g. global, local, region-based), standard GIS analysis tools can be automated to conduct the comparisons, although considerable experience with GIS is needed to properly use or modify the functions. For example, the ZONAL commands in ArcGIS (ESRI 2005) allow statistical operations to be conducted on a per region basis and thus can perform region-based comparisons. Without specialised software for comparison, the simpler methods are expected to be the most commonly used, primarily because of their accessibility. In this thesis, a spatial field comparison tool has been developed and used to undertake all analyses. The development and use of this tool is detailed in the remainder of this chapter.

6.5.1 The Invisible Modelling Environment (TIME)

Research amongst the hydrological community in Australia has benefited over recent years from the Cooperative Research Centre (CRC) for Catchment Hydrology and more recently the eWater CRC. A major output from these cooperative research programs is the Catchment Modelling Toolkit (Argent et al. 2003; eWater CRC 2006), which is a repository of models and modelling tools (e.g. optimisers, stochastic data generators) aimed at improving catchment modelling practice. Within this toolkit is a development environment named The Invisible Modelling Environment (TIME) (Rahman et al. 2003; Rahman et al. 2004). TIME provides users with access to the source code for the whole toolkit, enabling users to develop custom models/tools. This ensures that any capabilities needed within the Catchment Modelling Toolkit can be added. For example, during this thesis tools were developed for calculating variograms, performing connected components analysis and also for undertaking all spatial field comparisons discussed. Any person seeking access to these models and tools can gain access to the Catchment Modelling Toolkit and there is a growing user base for support, further development, etc.

The TIME source code includes all of the underlying classes (i.e. programmable objects) for the included models and tools, as well as a range of general input and output classes for different data types (including visual interface classes). TIME supports spatial, temporal and spatiotemporal data types, with classes that give access to element values, location and time. TIME is an ideal environment in which to implement the new comparison methods in this thesis. It provides suitable classes for implementing the algorithms described and for developing comparison tools that are readily accessible by the hydrological modelling community.



Source Fields: 961025_02.asc, 961025_10.asc, 961025_08.asc, 961025_09.asc, 961025_07.asc, 961025_06.asc, 961025_05.asc

Target Field: 961025.asc

Regions: 1, 2, 4, 5, 6

499 512 233 169

○ Black-White
● Dry-Wet
○ Blue-Red

□ Match? ☑ Auto? (1387.11,1184.47)

Scale 888
Shape 1

	Value	Location	Time
□ RMSE			
☑ Tolerant Mean Error Allowed	0	0	0
□ TCE			
□ Tolerant Mean Similarity Allowed	0	0	0
□ TSE Limit	0	0	0

	Field	AbsBias(1)	AbsBias(2)	AbsBias(4)	AbsBias(5)	Ab
▶	961025_02.a	74.96646356	49.15135135	62.96608767	40.82222222	1.6
	961025_10.a	64.08441228	45.69729729	51.67576509	30.31111111	8.1
	961025_08.a	61.11005331	42.73513513	56.38544251	11.55555555	21.
	961025_09.a	74.01774562	57.4	57.06286186	45.22222222	2.2
	961025_07.a	44.09979690	50.24864864	28.51447477	67.77777777	40.
	961025_06.a	59.67928408	51.32432432	47.09511993	59.68888888	21.
	961025_05.a	76.76646356	55.54594594	64.15963606	41.95555555	0.7
*						

213 0

○ Black-White
● Blue-Red

☑ Auto? (1081.58,729.74)

Intermediate Field(s): TME(0/0/0)(1)_961(… TME(0/0/0)(2)_961(

Figure 6.3 A typical screen capture from the prototype spatial field comparison tool developed for implementing the comparison methods in this thesis. Details are given in Appendix A.

All of the comparison methods developed in this thesis have been developed using TIME for application to regular element fields (i.e. rasters). The comparison measures for synthetic and hydrological fields (from Chapters 4 and 5) were all undertaken using a prototype spatial field comparison tool (Figure 6.3). This prototype contains a simple interface on which users can explore how spatial fields compare, without the need to implement the methods themselves. It allows all of the methods listed in Table 6.1 to be calculated. There is potential for future development of such a tool in the Catchment Modelling Toolkit that could integrate within other models.

For the interested reader, more complete details about using this tool are given in Appendix A. A detailed explanation is given about the interface, the options available and a tutorial showing how to conduct each type of comparison used in this thesis. The installation files and some sample spatial fields are provided on the CD with this thesis, or alternatively by contacting the author.

6.5.2 Managing computational demands

Working with spatial fields and comparison methods are computationally demanding tasks. Spatial fields are always much larger than temporal hydrological data sets because they represent a value for every location. Large fields, such as high-resolution remotely sensed imagery, can have millions of elements (e.g. a 1000 x 1000 pixel image). When a spatiotemporal series exists, the data size and computational demands increase even more. In addition to the data being quite cumbersome, most comparison algorithms require every element (at every time) to be inspected at least once during processing (e.g. algorithms detailed in Chapter 4). As the comparison methods become more inclusive of spatial relationships between neighbours, these computational demands increase again (with multiple visits to every element). Some of the methods, such as segmentation or interpolation, are iterative processes that require multiple loops through the field before they converge upon a solution. These iterative methods are computationally demanding, but are generally only required for parameterising a comparison method (i.e. they are ‘one-off’ computations). In some situations it is impossible to reduce the computational demands, while in others it is possible to limit the spatial or temporal extent of the data being processed. For example, comparing elements with a limited, but meaningful region may be just as informative as comparing the entire field, yet requires far less computing power.

Regardless of whether the amount of processing required can be reduced, the current cost (i.e. time and money) of computing power makes it feasible to run the comparison methods on standard desktop computers with typical hydrological fields. The fields used throughout this thesis mostly have less than 10,000 elements, being for small research catchments. The strict comparison measures are calculated almost instantly, while tolerant and multiscale measures take longer and are dependent on the size of location tolerances, number of intermediate scales, etc.

To give some example times, a field with 100 x 100 elements has been analysed on a Pentium 4 PC with 1 gigabyte of RAM. Strict comparison measures were produced in less than 1 second, but introducing a location tolerance of 10 elements (in all directions) increased processing time to approximately 3 seconds. The time required for a multiscale measure (with 4 scales from local to global) was approximately 80 seconds. The region merging segmentation algorithm ran in approximately 15 seconds, but this is dependent on the values and arrangement of the field (i.e. an organised field is simpler to segment). The greatest control on processing time is the spatial field size, with memory and processing demands (i.e. both in time and cost) increasing accordingly. Most of these issues can be managed via thoughtful software development or otherwise patience for longer runtimes.

6.6 Chapter summary

This chapter has provided discussion about the major issues relating to the application of new spatial comparison methods in hydrology. These methods require a different approach to model assessment, in which the user directs the comparison process rather than attempting to make qualitative comparisons that are difficult to repeat and report.

The subjective decisions required during comparison are identified and the manner in which they combine to control comparison is conceptualised. From these decisions, a comparison strategy can be devised for different hydrological situations. This is illustrated by using an example from a large-scale weather prediction model and observed RADAR rainfall.

More specific issues regarding application of these methods, including the analysis issues (e.g. interpretation, significance) and the implementation issues (e.g. software, resources), are discussed. The software tool developed throughout this thesis is detailed and it identifies an area where this research may progress in the future. The following chapter concludes this thesis by revisiting the initial research questions, describing how they have been addressed and presenting the important avenues for future research.

Chapter 7

Conclusions and further research

7.1 Conclusions

The ability to quantitatively compare spatial fields used in hydrological modelling has been advanced through the development of new comparison methods. These developments have been guided by an understanding of how qualitative methods (e.g. visual comparison), which are currently depended upon for making specialised comparisons in hydrology, operate during analysis and comparison. Qualitative methods are recognised as being powerful and versatile and they remain useful. However, they cannot be automated and become overwhelmed when applied to large, complex comparison tasks (e.g. large spatial extents, complex spatial arrangements and/or numerous spatial fields requiring comparison).

A review of approaches used in other disciplines facing similar comparison challenges (e.g. image processing, pattern recognition/matching) (Wealands et al. 2005a) has revealed three general themes that are common to most applications:

1. importance;
2. tolerance; and
3. completeness.

The development of these three themes into hydrological comparison methods has led to the significant advances. The themes emulate general processes undertaken during visual comparison. However, to achieve this, they must be guided by user knowledge and experience. The new comparison methods developed in this thesis capture user knowledge in an intuitive manner, through the use of pre-processing, feature definition and recognition, tolerance definitions and parameters that control spatial scale.

The notion of importance is incorporated into comparison by pre-processing and then comparing the processed fields. A range of existing and new methods can be used to change spatial fields so that they are comparable and representative of what is important. A data-driven segmentation method is a new approach for hydrological fields (adapted from image processing) and is capable of automatically recognising homogeneous regions (i.e. features) within the observed fields, which may have specific importance or interpretation. Specific hydrological or process knowledge can also be used to generate knowledge-driven regions, which can then allow region-based comparison measures to represent specialised comparisons. In model assessment, measures that incorporate importance enable model understanding to be achieved without requiring the user to visually inspect the individual spatial fields.

A new and versatile method for defining and incorporating tolerance into local comparison methods has been developed. Tolerances allow insignificant differences (in time, location or value) to be ignored during comparison, making it feasible that a range of fields will have equal similarity (i.e. all within the tolerances) to the target. This is not the same as tolerating errors globally and allows more explicit and specialised tolerances to be applied during comparison. In model assessment, this means that there may be multiple models that are all optimal, based on how strictly defined the target field is (as defined by the user). This is a desirable property and the range of optimal models is used to evaluate model uncertainty. The intermediate field of tolerant residuals also has potential for use in data assimilation, where it can be more effective for determining corrections to model states because the known uncertainties have already been tolerated.

Completeness of comparison is ensured by using a logical comparison strategy. Devising a comparison strategy encourages the user to plan which different aspects and scales of spatial fields are compared, based on those that can be measured using available methods. It is not possible to make unambiguous comparisons with single measures that do not consider the purpose of comparison. Some single measures can combine multiple scales or aspects, although they are difficult to interpret and have limited versatility. It is more suitable to use a range of comparison measures and to analyse performances using multi-criteria optimisation procedures.

The new and current comparison methods, which incorporate the notions of importance and tolerance, combine to produce an extensive suite of measures that can quantify a broad range of aspects and scales of spatial fields. Of the 14 groups of comparison measures shown in the flowchart (Figure 7.1), only 3 groups can be achieved with the methods that are currently used in hydrology. By introducing the KS similarity measure, the new multiscale measure, tolerances and the notion of importance (via regions), the 11 additional aspects of comparison can be measured. These provide the framework for devising a comparison strategy for a specific application. The strategy will generally reflect how the user would undertake the comparison qualitatively. However, using this approach, the user is only required to make subjective decisions once. The strategy can then be applied to numerous spatial fields to produce a complete quantitative comparison, including all aspects and scales the user deemed suitable.

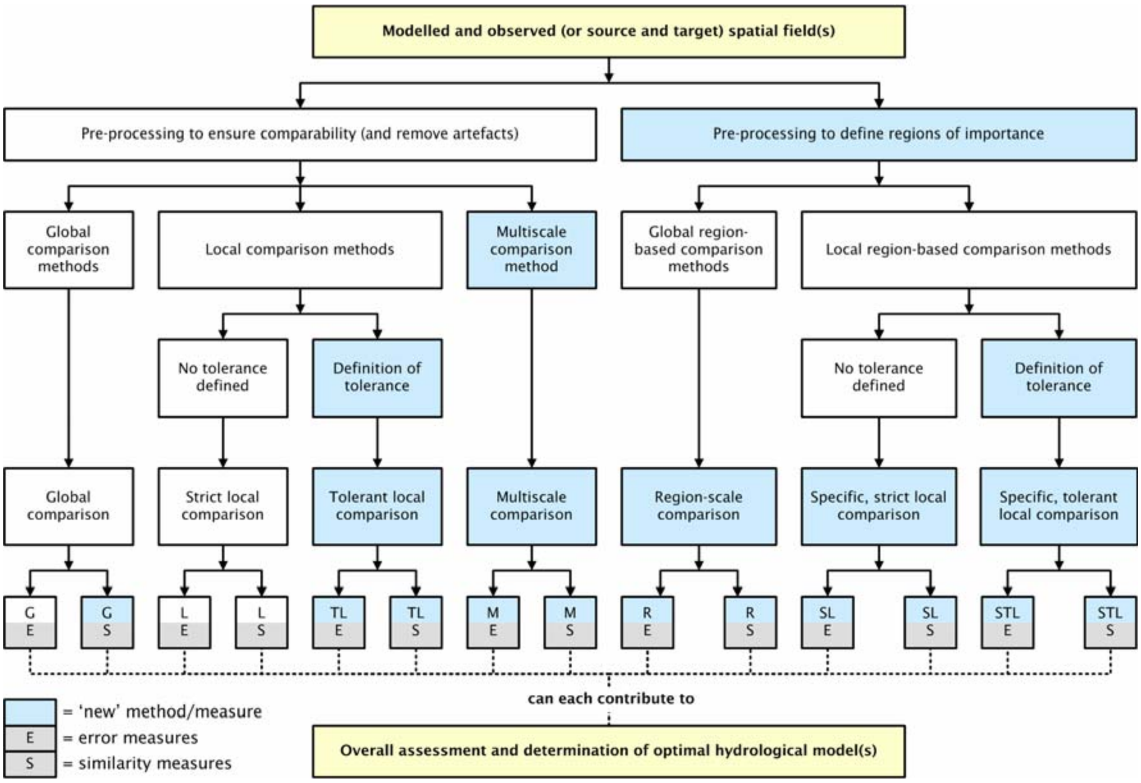

Modelled and observed (or source and target) spatial field(s)
Pre-processing to ensure comparability (and remove artefacts)
Pre-processing to define regions of importance
Global comparison methods
Local comparison methods
Multiscale comparison method
Global region-based comparison methods
Local region-based comparison methods
No tolerance defined
Definition of tolerance
No tolerance defined
Definition of tolerance
Global comparison
Strict local comparison
Tolerant local comparison
Multiscale comparison
Region-scale comparison
Specific, strict local comparison
Specific, tolerant local comparison
G E
G S
L E
L S
TL E
TL S
M E
M S
R E
R S
SL E
SL S
STL E
STL S
can each contribute to
Overall assessment and determination of optimal hydrological model(s)
= 'new' method/measure
E = error measures
S = similarity measures

Figure 7.1 The comparison flowchart (repeated from Figure 4.20) details the expanded range of measures used for spatial field comparison, which should be used for assessing the optimally modelled fields from the input set of potential models and observed fields.

The actual process of making the subjective decisions forces the user to consider in detail what to compare, which is an improved approach from common hydrological modelling practice. The new methods ask the user what parts of the field are important; what differences are significant; what uncertainty exists in the fields; and what scale(s) to compare. Simple software tools for translating user knowledge into comparison parameters (e.g. regions, tolerances) must be available to assist users with converting their knowledge into parameters. The software presented in this thesis is an example of such a tool that could assist with changing comparison practices.

This type of approach is a contrast to most current comparison measures, which are applied to all types of spatial fields in the same manner. Different branches of hydrology have varying demands for comparison, due mainly to the different modelling expectations and uncertainties involved. Therefore, the new methods can be better adapted to different applications. The strict, non-specific methods are still useful for providing a strict measure of difference, but they become more useful when rigorously combined with user knowledge.

The new comparison measures have been evaluated against expert assessment and visual evidence. They achieve results that are visually intuitive and they provide more detailed quantitative information than previously available. The unbiased and rigorous comparison of different aspects and scales produced clear, repeatable and reportable comparison findings that address the major weaknesses with current comparison methods. The examples used in this thesis were relatively simple to facilitate evaluation, but the true power of these methods lies in their application to larger, more complex fields and comparison tasks. Spatial fields that contain numerous important regions or model assessment tasks requiring numerous comparisons (e.g. Monte Carlo simulations) will truly benefit from more advanced and meaningful comparisons.

Comprehensive model assessment can only be achieved by using comprehensive comparisons. For spatial fields, the current methods are limited by their lack of user interaction. By including the user into the comparison process in a rigorous and intuitive manner, better comparison measures have been developed. These measures subsequently produce better model testing and more informative measures for model understanding, which ultimately improves the modelling process.

7.2 Further research

This thesis has addressed an area of hydrology that is increasingly being encountered – the need to quantify the error or similarity between spatial fields. Apart from some comparison methods designed for specific applications (e.g. Grayson et al. 2002; Güntner et al. 2004), there has been limited research on this topic in hydrology prior to this thesis. As a result, there remain many avenues for future research that can contribute to the improved comparison of spatial fields.

As the comparison methods introduced in this thesis become used in more complex hydrological situations, their true value will become better understood. At present, there are only a few studies in which spatial observations have been used for comprehensive model assessment. Apart from the soil moisture modelling at Tarrawarra, these studies have mainly worked with binary spatial fields of flood inundation (Hunter et al. 2005), snow cover (Blöschl et al. 1991; Blöschl and Kirnbauer 1992) or saturated areas (Güntner et al. 2004). The work on flood inundation by Hunter et al. (2005) is the best example so far of applying comparison methods to reduce modelling uncertainty, in this case with binary fields. Future research should focus on using these new methods to make better assessments of complex situations. Ideally, the measures will be used for assessing 'Monte Carlo type' modelling situations, where they determine the likelihood of each potential model.

The significance (in a statistical sense) of differences between comparison measures is unclear and this makes interpretation of results difficult. Further research in this area is expected, following on the recent progresses made with categorical comparisons by Remmel and Csillag (2003). Continuous fields present a more difficult problem when estimating the distributions of comparison measures, but some methods to help with the general interpretation may be sufficient. The many degrees of freedom in comparison problems is expected to limit the development of simple significance testing methods, so understanding via experience and relative performance (e.g. the COE in hydrology) is more likely. There is also potential for geostatistical methods to be used for producing multiple realisations of continuous fields that adhere to specific spatial relationships (e.g. connectivity measures).

These methods address spatial field comparisons, with some limited consideration of temporal information. As data availability and modelling approaches advance, the comparison of spatiotemporal series is expected to become necessary. In this thesis, the comparison of spatiotemporal series is undertaken by first making spatial comparisons and then using the numerical results within a simple temporal series. However, understanding changes at local spatial scales throughout time may also be useful. For example, tracking feature movements or changes throughout time may be possible, although heavily dependent on the definition of a feature, which remains unclear in continuous hydrological fields. Fuzzy region definitions may be a potential solution to this problem, although this is likely to increase the difficulty and versatility of comparison methods. Tracking applications will be reliant on recognising a 'dynamic feature' via analysis of a spatiotemporal series. For categorical fields, the change of features over time will be simpler to recognise (i.e. because boundaries exist between categories) and is likely to be where initial advances are made.

This thesis has focused on using quantitative measures, in which visual inspection and analysis of the result is not used. However, the intermediate measures produced by tolerant comparisons are potentially useful for various applications, but particularly within data assimilation. Existing data assimilation studies usually assimilate either absolute values or residuals to correct states within the model. However, by tolerating insignificant differences between elements, the intermediate field can be modified to better reflect where corrections are needed (i.e. avoiding false findings). Using these methods would alter the magnitude and arrangement of the corrections applied during assimilation and may prove to be useful, although this requires further investigation.

To encourage the use of newer methods by hydrologists, software tools that are readily available and easy to use are required. A prototype software tool has been developed in this thesis to apply and test all of the comparison methods presented (Appendix A). A spatial comparison tool such as this would benefit from greater software development to improve both the user interface and computational efficiency. There are also other considerations for the development and design of such software, including user needs assessment and user-centred design. It is also necessary to publicise and distribute such a tool, although through the internet and the Catchment Modelling Toolkit (eWater CRC 2006) this is quite achievable.

There is great scope for future application and development of comparison methods for spatial fields. At present, the hydrological community continues to develop its ability to observe and model spatial fields, but comprehensive and rigorous model testing is still in its infancy. The methods developed in this thesis provide a new suite of tools to the modeller and, if adopted, should significantly advance the art and science of spatial hydrological modelling.

Bibliography

Abbott, M.B., Bathurst, J.C., Cunge, J.A., O'Connell, P.E. & Rasmussen, J. 1986, 'An introduction to the European Hydrological System -- Systeme Hydrologique Europeen, "SHE", 2: Structure of a physically-based, distributed modelling system', *Journal of Hydrology,* vol.87, no.1-2, pp.61-77.

Abbott, M.B. & Refsgaard, J.C. 1996, *Distributed hydrological modelling,* Water science and technology library; v.22, Kluwer Academic, Dordrecht ; Boston.

Abu El-Nasr, A., Arnold, J.G., Feyen, J. & Berlamont, J. 2005, 'Modelling the hydrology of a catchment using a distributed and a semi-distributed model', *Hydrological Processes,* vol.19, pp.573-587.

Aerts, J.C.J.H., Clarke, K.C. & Keuper, A.D. 2003, 'Testing Popular Visualization Techniques for Representing Model Uncertainty', *Cartography and Geographic Information Science,* vol.30, no.3, pp.249-261.

Argent, R.M., Vertessy, R.A., Rahman, J.M., Cuddy, S.M., Podger, G.D. & Perry, D.R. 2003, 'Building a modelling toolkit for prediction of catchment behaviour', *In: MODSIM 2003 International Congress on Modelling and Simulation*, Post, D. ed., Modelling and Simulation Society of Australia and New Zealand Inc., July 2003, Townsville, pp.1715-1720.

Aronica, G., Bates, P.D. & Horritt, M.S. 2002, 'Assessing the uncertainty in distributed model predictions using observed binary pattern information within GLUE', *Hydrological Processes,* vol.16, pp.2001-2016.

Baatz, M., Binnig, G., Eschenbacher, P., Melchinger, A. & Sogtrop, M. 2004, *Method of iterative segmentation of a digital picture,* US Patent 6832002.

Baatz, M. & Schäpe, A. 2000, 'Multiresolution segmentation: an optimization approach for high quality multiscale image segmentation', in *Angewandte Geogr. Informationsverabeitung*, vol.12, Strobl, J. & Blaschke, T. eds., Herbert Wichmann Verlag, Heidelberg, Germany, pp.12-23.

Barron, J.L., Beauchemin, S.S. & Fleet, D.J. 1994, 'On optical flow', *In: 6th Int Conf on Artificial Intelligence and Information-Control Systems of Robots*, September 12-16, Bratislava, Slovakia, pp.3-14.

Bates, P.D., Horritt, M.S., Aronica, G. & Beven, K. 2004, 'Bayesian updating of flood inundation likelihoods conditions on flood extent data', *Hydrological Processes,* vol.18, pp.3347-3370.

Beven, K. 1995, 'TOPMODEL', in *Computer Models of Watershed hydrology*, Singh, V.P. ed., Water Resources Publications, Highlands Ranch, Colorado, pp.627-668.

Beven, K. 2002, 'Towards an alternative blueprint for a physically based digitally simulated hydrologic response modelling system', *Hydrological Processes,* vol.16, no.2, pp.189-206.

Beven, K. & Binley, A. 1992, 'The Future of Distributed Models - Model Calibration and Uncertainty Prediction', *Hydrological Processes,* vol.6, no.3, pp.279-298.

Beven, K.J. 2001, 'How far can we go with distributed hydrological modelling?' *Hydrology and Earth Science Systems,* vol.5, no.1, pp.1-12.

Beven, K.J. & Kirkby, M.J. 1979, 'A physically-based, variable contributing area model of basin hydrology', *Hydrol. Sci. Bull.,* vol.24, pp.43-69.

Blöschl, G. & Grayson, R. 2000, 'Spatial Observations and Interpolation', in *Spatial Patterns in Catchment Hydrology: Observations and Modelling*, Grayson, R. & Blöschl, G. eds., Cambridge University Press, Cambridge, pp.17-50.

Blöschl, G. & Kirnbauer, R. 1992, 'An Analysis of Snow Cover Patterns in a Small Alpine Catchment', *Hydrological Processes,* vol.6, no.1, pp.99-109.

Blöschl, G., Kirnbauer, R. & Gutknecht, D. 1991, 'Distributed Snowmelt Simulations in an Alpine Catchment, 1. Model Evaluation on the Basis of Snow Cover Patterns', *Water Resources Research,* vol.27, no.12, pp.3171-3179.

Blöschl, G. & Sivapalan, M. 1995, 'Scale issues in hydrological modeling - a review', *Hydrological Processes,* vol.9, no.3-4, pp.251-290.

Boegh, E., Thorsen, M., Butts, M.B., Hansen, S., Christiansen, J.S., Abrahamsen, P., Hasager, C.B., Jensen, N.O., van der Keur, P. & Refsgaard, J.C. 2004, 'Incorporating remote sensing data in physically based distributed agro-hydrological modelling', *Journal of Hydrology,* vol.287, no.1-4, pp.279-299.

Bow, S. 2002, *Pattern recognition and image preprocessing,* Signal processing and communications., 2nd edn. Marcel Dekker, New York, USA.

Boyle, D.P., Gupta, H.V. & Sorooshian, S. 2000, 'Toward improved calibration of hydrologic models: Combining the strengths of manual and automatic methods', *Water Resources Research,* vol.36, no.12, pp.3663-3674.

Brewer, C.A. 2003, 'A Transition in Improving Maps: The ColorBrewer Example', *Cartography and Geographic Information Science,* vol.30, no.2, pp.159-162.

Bruce, N.D.B. 2003, *Evolutionary Design for Computational Visual Attention*, Masters Thesis, University of Waterloo, Waterloo, Ontario, Canada

Burnash, R.J.C. 1995, 'The NWS river forecast system--catchment modelling', in *Computer Models of Watershed hydrology*, Singh, V.P. ed., Water Resources Publications, Highlands Ranch, Colorado, pp.311-366.

Burnett, C. & Blaschke, T. 2003, 'A multi-scale segmentation/object relationship modelling methodology for landscape analysis', *Ecological Modelling,* vol.168, no.3, pp.233-249.

Chang, Y.L. & Li, X. 1994, 'Adaptive image region-growing', *Image Processing, IEEE Transactions on,* vol.3, no.6, pp.868-872.

Chen, Y. & Wang, J.Z. 2002, 'A region-based fuzzy feature matching approach to content-based image retrieval', *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* vol.24, no.0162-8828, pp.1252-1267.

Cheng, T., Molenaar, M. & Lin, H. 2001, 'Formalizing fuzzy objects from uncertain classification results', *International Journal of Geographical Information Science,* vol.15, no.1, pp.27-42.

Chirico, G.B., Grayson, R.B. & Western, A.W. 2003, 'A downward approach to identifying the structure and parameters of a process-based model for a small experimental catchment', *Hydrological Processes,* vol.17, no.11, pp.2239-2258.

Cobby, D.M., Mason, D.C. & Davenport, I.J. 2001, 'Image processing of airborne scanning laser altimetry data for improved river flood modelling', *Isprs Journal of Photogrammetry and Remote Sensing,* vol.56, no.2, pp.121-138.

Cohen, J. 1960, 'A coefficient of agreement for nominal scales', *Educational Psychology Measurement,* vol.20, pp.37-46.

Costanza, R. 1989, 'Model Goodness of Fit - a Multiple Resolution Procedure', *Ecological Modelling,* vol.47, no.3-4, pp.199-215.

Crawford, N.H. & Linsley, R.K. 1966, *Digital simulation in hydrology, STANFORD Watershed Model IV*, Department of Civil Engineering, Stanford University, Technical Report 39.

Definiens Imaging 2003, *eCognition User Guide 3*, Definiens Imaging GmbH, Munich, Germany.

Dingman, S.L. 2002, *Physical hydrology*, 2nd edn. Prentice Hall, Upper Saddle River, NJ.

Dungan, J.L. 2001, 'Scaling up and scaling down: the relevance of the support effect on remote sensing of vegetation', in *Modelling scale in geographical information science*, Tate, N.J. & Atkinson, P.M. eds., Wiley, Chichester ; New York, pp.xiv, 277.

Efron, B. 1981, 'Nonparametric estimates of standard error: The jackknife, the bootstrap, and other methods', *Biometrika,* vol.68, pp.589-599.

Entekhabi, D. & Eagleson, P.S. 1989, 'Land surface hydrology parameterization for atmospheric general circulation models including subgrid scale spatial variability', *Journal of Climate,* vol.2, pp.816-831.

ESRI 2005, *ArcGIS ArcEditor 9.1*, [Software], Environmental Systems Research Institute, Inc., Redlands, CA.

eWater CRC 2006, *Catchment Modelling Toolkit*, [Online], Available: http://www.toolkit.net.au [11/05/2006].

Findlay, J.M., Walker, R. & Kentridge, R.W. 1995, *Eye movement research : mechanisms, processes and applications,* Studies in visual information processing, Elsevier, Amsterdam, The Netherlands.

Flood, M. 2001, 'Laser altimetry: From science to commercial lidar mapping', *Photogrammetric Engineering and Remote Sensing,* vol.67, no.11, pp.1209-1217.

Foody, G.M. 2002, 'Status of land cover classification accuracy assessment', *Remote Sensing of Environment,* vol.80, no.1, pp.185-201.

Foody, G.M. 2006, 'What is the difference between two maps? A remote senser's view', *Journal of Geographical Systems,* vol.8, pp.119-130.

Foufoula-Georgiou, E. & Vuruputur, V. 2000, 'Patterns and Organisation in Precipitation', in *Spatial Patterns in Catchment Hydrology: Observations and Modelling*, Grayson, R. & Blöschl, G. eds., Cambridge University Press, Cambridge, pp.82-104.

Franks, S.W., Gineste, P., Beven, K.J. & Merot, P. 1998, 'On constraining the predictions of a distributed model: The incorporation of fuzzy estimates of saturated areas into the calibration process', *Water Resources Research,* vol.34, no.4, pp.787-797.

Fritz, S. & See, L. 2005, 'Comparison of land cover maps using fuzzy agreement', *International Journal of Geographical Information Science,* vol.19, no.7, pp.787-807.

Gallant, J.C. & Dowling, T.I. 2003, 'A multiresolution index of valley bottom flatness for mapping depositional areas', *Water Resources Research,* vol.39, no.12, p.1347.

Grayson, R. & Blöschl, G. 2000a, 'Spatial Modelling of Catchment Dynamics', in *Spatial Patterns in Catchment Hydrology: Observations and Modelling*, Grayson, R. & Blöschl, G. eds., Cambridge University Press, Cambridge, pp.51-81.

Grayson, R. & Blöschl, G. 2000b, 'Summary of Pattern Comparison and Concluding Remarks', in *Spatial Patterns in Catchment Hydrology: Observations and*

Modelling, Grayson, R. & Blöschl, G. eds., Cambridge University Press, Cambridge, pp.355-367.

Grayson, R.B., Blöschl, G. & Moore, I.D. 1995, 'Distributed parameter hydrologic modelling using vector elevation data: THALES and TAPES-C', in *Computer Models of Watershed hydrology*, Singh, V.P. ed., Water Resources Publications, Highlands Ranch, Colorado, pp.669-96.

Grayson, R.B., Blöschl, G., Western, A.W. & McMahon, T.A. 2002, 'Advances in the use of observed spatial patterns of catchment hydrological response', *Advances in Water Resources,* vol.25, pp.1313-1334.

Grayson, R.B., Moore, I.D. & McMahon, T.A. 1992, 'Physically Based Hydrologic Modeling. 2. Is the Concept Realistic?' *Water Resources Research,* vol.28, no.10, pp.2659-2666.

Grayson, R.B., Western, A.W., Chiew, F.H.S. & Blöschl, G. 1997, 'Preferred states in spatial soil moisture patterns: Local and nonlocal controls', *Water Resour. Res.,* vol.33, no.12, p.97WR02174.

Güntner, A., Seibert, J. & Uhlenbrook, S. 2004, 'Modeling spatial patterns of saturated areas: An evaluation of different terrain indices', *Water Resources Research,* vol.40, no.5, p.W05114.

Gupta, H.V., Sorooshian, S. & Yapo, P.O. 1998, 'Toward improved calibration of hydrologic models: Multiple and noncommensurable measures of information', *Water Resources Research,* vol.34, no.4, pp.751-763.

Gustafson, E.J. 1998, 'Quantifying Landscape Spatial Pattern: What Is the State of the Art?' *Ecosystems,* vol.1, pp.143-156.

Hadjidemetriou, E., Grossberg, M.D. & Nayar, S.K. 2004, 'Multiresolution Histograms and Their Use for Recognition', *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol.26, no.7, pp.831-847.

Hagen, A. 2003, 'Fuzzy set approach to assessing similarity of categorical maps', *International Journal of Geographical Information Science,* vol.17, no.3, pp.235-249.

Hagen-Zanker, A. 2005, 'Further developments of a fuzzy set map comparison approach', *International Journal of Geographical Information Science,* vol.19, no.7, pp.769-785.

Hagen-Zanker, A. 2006, 'Map comparison methods that simultaneously address overlap and structure', *Journal of Geographical Systems,* vol.8, p.165–185.

Han, J., Kamber, M. & Tung, A.K.H. 2001, 'Spatial Clustering Methods in Data Mining: A Survey', in *Geographic Data Mining and Knowledge Discovery*, Miller, H. & Han, J. eds., Taylor and Francis.

Haralick, R.M. & Shapiro, L.G. 1992, *Computer and robot vision*, Addison-Wesley, Reading, Mass.

Hargrove, W.W. 2006, 'Mapcurves: A quantitative method for comparing categorical maps', *Journal of Geographical Systems,* vol.8, pp.187-208.

Hay, G.J., Blaschke, T., Marceau, D.J. & Bouchard, A. 2003, 'A comparison of three image-object methods for the multiscale analysis of landscape structure', *ISPRS Journal of Photogrammetry and Remote Sensing,* vol.57, no.5-6, pp.327-345.

Hess, G.R. & Bay, J.M. 1997, 'Generating confidence intervals for composition-based landscape indexes', *Landscape Ecology,* vol.12, pp.309-320.

Horritt, M.S. & Bates, P.D. 2002, 'Evaluation of 1D and 2D numerical models for predicting river flood inundation', *Journal of Hydrology,* vol.268, no.1-4, pp.87-99.

Horritt, M.S., Mason, D.C. & Luckman, A.J. 2001, 'Flood boundary delineation from Synthetic Aperture Radar imagery using a statistical active contour model', *International Journal of Remote Sensing,* vol.22, no.13, pp.2489-2507.

Hosking, J.R.M. 1990, 'L-moments: analysis and estimation of distributions using linear combinations of order statistics', *Journal of the Royal Statistical Society, Series B,* vol.52, pp.105-124.

Houser, P.R., Goodrich, D. & Syed, K. 2000, 'Runoff, Precipitation, and Soil Moisture at Walnut Gulch', in *Spatial Patterns in Catchment Hydrology: Observations and Modelling*, Grayson, R. & Blöschl, G. eds., Cambridge University Press, Cambridge, pp.125-157.

Hunter, N.M., Bates, P.D., Horritt, M.S., De Roo, P.J. & Werner, M.G.F. 2005, 'Utility of different data types for calibrating flood inundation models within a GLUE framework', *Hydrology and Earth System Sciences,* vol.9, no.4, pp.412-430.

Hutchinson, M.F. & Gessler, P.E. 1994, 'Splines - more than just a smooth interpolator', *Geoderma.,* vol.62, pp.45-67.

Huttenlocher, D.P., Klanderman, G.A. & Rucklidge, W.J. 1993, 'Comparing Images Using the Hausdorff Distance', *Ieee Transactions on Pattern Analysis and Machine Intelligence,* vol.15, no.9, pp.850-863.

Isaaks, E.H. & Srivastava, R.M. 1989, *Applied geostatistics*, Oxford University Press, New York.

Itti, L., Koch, C. & Niebur, E. 1998, 'A model of saliency-based visual attention for rapid scene analysis', *Ieee Transactions on Pattern Analysis and Machine Intelligence,* vol.20, no.11, pp.1254-1259.

Jain, A.K., Murty, M.N. & Flynn, P.J. 1999, 'Data clustering: a review', *ACM Computing Surveys,* vol.31, no.3, pp.264-323.

Jetten, V., Govers, G. & Hessel, R. 2003, 'Erosion models: quality of spatial predictions', *Hydrological Processes,* vol.17, no.5, pp.887-900.

Jolliffe, I.T. & Stephenson, D.B. 2003, *Forecast verification : a practitioner's guide in atmospheric science*, J. Wiley, Chichester, West Sussex, England ; Hoboken, NJ.

Journel, A.G. & Huijbregts, C.J. 1978, *Mining geostatistics*, Academic Press, London ; New York.

Klemeš, V. 1983, 'Conceptualization and scale in hydrology', *Journal of Hydrology,* vol.65, no.1-3, pp.1-23.

Klemeš, V. 1986, 'Operational Testing of Hydrological Simulation-Models', *Hydrological Sciences Journal,* vol.31, no.1, pp.13-24.

Legates, D.R. & McCabe, G.J. 1999, 'Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation', *Water Resources Research,* vol.35, no.1, pp.233-242.

Li, H. & Wu, J. 2004, 'Use and misuse of landscape indices', *Landscape Ecology,* vol.19, no.4, pp.389-399.

Li, X., He, H.S., Bu, R., Wen, Q., Chang, Y., Hu, Y. & Li, Y. 2005, 'The adequacy of different landscape metrics for various landscape patterns', *Pattern Recognition,* vol.38, no.12, pp.2626-2638.

Light, A. & Bartlein, P.J. 2004, 'The End of the Rainbow? Color Schemes for Improved Data Graphics', *Eos Transactions. AGU, 85(40), 385.,* vol.85, no.40, p.385.

Littlewood, I.G., Croke, B.F.W., Jakeman, A.J. & Sivapalan, M. 2003, 'The role of top-down modelling for Prediction in Ungauged Basins (PUB)', *Hydrological Processes,* vol.17, no.8, pp.1673-1679.

Lundquist, J.E., Lindner, L.R. & Popp, J. 2001, 'Using landscape metrics to measure suitability of a forested watershed: A case study for old growth', *Canadian Journal of Forest Research,* vol.31, no.10, p.1786.

MacQueen, J.B. 1967, 'Some Methods for Classification and Analysis of Multivariate Observations', *In: Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press, Berkeley, pp.281-297.

Madsen, H. 2003, 'Parameter estimation in distributed hydrological catchment modelling using automatic calibration with multiple objectives', *Advances in Water Resources,* vol.26, pp.205-216.

Maeder, A.J. 2005, 'The image importance approach to human vision based image quality characterization', *Pattern Recognition Letters,* vol.26, no.3, pp.347-354.

Martin, D.R. 2002, *An Empirical Approach to Grouping and Segmentation*, PhD Thesis, University of California, Berkeley

Mastin, G.A. 1985, 'Adaptive Filters for Digital Image Noise Smoothing: An Evaluation', *Computer Vision, Graphics and Image Processing (CVGIP),* vol.31, pp.103-121.

McCabe, M.F., Franks, S.W. & Kalma, J.D. 2005, 'Calibration of a land surface model using multiple data sets', *Journal of Hydrology,* vol.302, no.1-4, pp.209-222.

McGarigal, K. & Marks, B. 1995, *FRAGSTATS: spatial pattern analysis program for quantifying landscape structure*, USDA Forest Service, Pacific Northwest Research Station, Portland, General Technical Report PNW-GTR-351.

Merz, B. & Plate, E.J. 1997, 'An analysis of the effects of spatial variability of soil and soil moisture on runoff', *Water Resources Research,* vol.33, no.12, pp.2909 - 2922.

Monserud, R.A. & Leeemans, R. 1992, 'Comparing global vegetation maps with the Kappa statistic', *Ecological Modelling,* vol.62, pp.275-293.

Moore, I.D., Grayson, R.B. & Ladson, A.R. 1991, 'Digital Terrain Modelling: A Review of Hydrological, Geomorphological, and Biological Applications', *Hydrological Processes,* vol.5, pp.3-30.

Morin, E., Goodrich, D.C., Maddox, R.A., Gao, X., Gupta, H.V. & Sorooshian, S. 2006, 'Spatial patterns in thunderstorm rainfall events and their coupling with watershed hydrological response', *Advances in Water Resources,* vol.29, no.6, pp.843-860.

Nash, J.E. & Sutcliffe, J.V. 1970, 'River flow forecasting through conceptual models, I, A discussion of principles', *Journal of Hydrology,* vol.10, pp.282-290.

Osberger, W. & Maeder, A.J. 1998, 'Automatic Identification of Perceptually Important Regions in an Image', *In: 14th International Conference on Pattern Recognition*, 16–20 August, Brisbane, Australia.

Pal, N.R. & Pal, S.K. 1993, 'A review on image segmentation techniques', *Pattern Recognition,* vol.26, no.9, pp.1277-1294.

Pappenberger, F. & Beven, K.J. 2006, 'Ignorance is bliss: Or seven reasons not to use uncertainty analysis', *Water Resources Research,* vol.42, p.W05302.

Pauwels, E.J. & Frederix, G. 1999, 'Finding salient regions in images - Nonparametric clustering for image segmentation and grouping', *Computer Vision and Image Understanding,* vol.75, no.1-2, pp.73-85.

Pauwels, V.R.N., Hoeben, R., Verhoest, N.E.C. & De Troch, F.P. 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', *Journal of Hydrology,* vol.251, no.1-2, pp.88-102.

Pavlidis, T. 2003, '36 years on the pattern recognition front', *Pattern Recognition Letters,* vol.24, no.1-3, pp.1-7.

Pontius, R.G. 2002, 'Statistical methods to partition effects of quantity and location during comparison of categorical maps at multiple resolutions', *Photogrammetric Engineering & Remote Sensing,* vol.68, no.10, pp.1041-1049.

Pontius, R.G., Huffaker, D. & Denman, K. 2004a, 'Useful techniques of validation for spatially explicit land-change models', *Ecological Modelling,* vol.179, no.4, pp.445-461.

Pontius, R.G. & Malizia, N.R. 2004, 'Effect of Category Aggregation on Map Comparison', *In: Geographic Information Science, Third International Conference, GIScience 2004*, 3234 edn., Egenhofer, M.J., Freksa, C. & Miller, H.J. eds., Springer, October 20-23, Adelphi, Maryland, USA, pp.251-268.

Pontius, R.G., Shusas, E. & McEachern, M. 2004b, 'Detecting important categorical land changes while accounting for persistence', *Agriculture, Ecosystems & Environment,* vol.101, no.2-3, pp.251-268.

Power, C., Simms, A. & White, R. 2001, 'Hierarchical fuzzy pattern matching for the regional comparison of land use maps', *International Journal of Geographical Information Science,* vol.15, no.1, pp.77-100.

Press, W.H., Teukolsky, S.A., Vetterling, W.T. & Flannery, B.P. 1992, *Numerical Recipes in C: The Art of Scientific Computing, Second Edition*, Cambridge University Press, New York.

Pujol, A., Villanueva, J.J. & Alba, J.L. 2002, 'A supervised modification of the Hausdorff distance for visual shape classification', *International Journal of Pattern Recognition and Artificial Intelligence,* vol.16, no.3, pp.349-359.

Quinn, P.F., Beven, K.J. & Lamb, R. 1995, 'The Ln(a/Tan-Beta) Index - How to Calculate It and How to Use It within the Topmodel Framework', *Hydrological Processes,* vol.9, no.2, pp.161-182.

Rabus, B., Eineder, M., Roth, A. & Bamler, R. 2003, 'The shuttle radar topography mission - a new class of digital elevation models acquired by spaceborne radar', *Isprs Journal of Photogrammetry and Remote Sensing,* vol.57, no.4, pp.241-262.

Radke, R.J., Andra, S., Al-Kofahi, O. & Roysam, B. 2005, 'Image change detection algorithms: a systematic survey', *IEEE Transactions on Image Processing,* vol.14, no.3, pp.294-307.

Rahman, J.M., Seaton, S.P. & Cuddy, S.M. 2004, 'Making frameworks more useable: using model introspection and metadata to develop model processing tools', *Environmental Modelling & Software,* vol.19, no.3, pp.275-284.

Rahman, J.M., Seaton, S.P., Perraud, J.-M., Hotham, H., Verrelli, D.I. & Coleman, J.R. 2003, 'It's TIME for a new environmental modelling framework', *In: MODSIM 2003 International Congress on Modelling and Simulation*, Post, D. ed., Modelling and Simulation Society of Australia and New Zealand Inc., July 2003, Townsville, pp.1727-1732.

Rawls, W.J., Brakensiek, D.L. & Miller, N. 1983, 'Green-Ampt infiltration parameters from soils data', *Journal of Hydrologic Engineering,* vol.109, pp.62-70.

Refsgaard, J.C. 1997, 'Parameterisation, calibration and validation of distributed hydrological models', *Journal of Hydrology,* vol.198, no.1-4, pp.69-97.

Remmel, T.K. & Csillag, F. 2003, 'When are two landscape pattern indices significantly different?' *Journal of Geographic Systems,* vol.5, pp.331-351.

Rock, I. & Palmer, S. 1990, 'The legacy of Gestalt psychology', *Scientific American,* vol.263, pp.84-90.

Rykiel, E.J. 1996, 'Testing ecological models: the meaning of validation', *Ecological Modelling,* vol.90, no.3, pp.229-244.

Sarkar, S. & Boyer, K.L. 1993, 'Perceptual Organization in Computer Vision - a Review and a Proposal for a Classifactory Structure', *Ieee Transactions on Systems Man and Cybernetics,* vol.23, no.2, pp.382-399.

Schmugge, T.J., Kustas, W.P., Ritchie, J.C., Jackson, T.J. & Rango, A. 2002, 'Remote sensing in hydrology', *Advances in Water Resources,* vol.25, no.8-12, pp.1367-1385.

Shamir, E., Imam, B., Gupta, H.V. & Sorooshian, S. 2005, 'Application of temporal streamflow descriptors in hydrologic model parameter estimation', *Water Resources Research,* vol.41, p.W06021.

Singh, V.P. (ed.) 1995, *Computer Models of Watershed Hydrology*, Water Resources Publications, Highlands Ranch, Colorado.

Sivapalan, M., Blöschl, G., Zhang, L. & Vertessy, R. 2003, 'Downward approach to hydrological prediction', *Hydrological Processes,* vol.17, no.11, pp.2101-2111.

Skoien, J.O. & Bloschl, G. 2006, 'Scale effects in estimating the variogram and implications for soil hydrology', *Vadose Zone Journal,* vol.5, no.1, pp.153-167.

Smeulders, A.W.M., Worring, M., Santini, S., Gupta, A. & Jain, R. 2000, 'Content-based image retrieval at the end of the early years', *Ieee Transactions on Pattern Analysis and Machine Intelligence,* vol.22, no.12, pp.1349-1380.

Smith, B.J. 1986, 'Perceptual organization in a random stimulus', in *Human and Machine Vision II*, Rosenfeld, A. ed., Academic, New York, pp.237-242.

Sun, X., Mein, R.G., Keenan, T.D. & Elliott, J.F. 2000, 'Flood estimation using radar and raingauge data', *Journal of Hydrology,* vol.239, no.1-4, pp.4-18.

Takken, I., Beuselinck, L., Nachtergaele, J., Govers, G., Poesen, J. & Degraer, G. 1999, 'Spatial evaluation of a physically-based distributed erosion model (LISEM)', *CATENA,* vol.37, no.3-4, pp.431-447.

Tarboton, D., Blöschl, G., Cooley, K., Kirnbauer, R. & Luce, C. 2000, 'Spatial Snow Cover Processes at Kuhtai and Reynolds Creek', in *Spatial Patterns in Catchment Hydrology: Observations and Modelling*, Grayson, R. & Blöschl, G. eds., Cambridge University Press, Cambridge, pp.158-186.

Tompa, D., Morton, J. & Jernigan, E. 2000, 'Perceptually based image comparison', *In: Proceedings of the IEEE International Conference on Image Processing*, 10-13 September, Vancouver, BC, Canada, pp.489-492.

Troch, P., Verhoest, N., Gineste, P., Paniconi, C. & Merot, P. 2000, 'Variable Source Areas, Soil Moisture and Active Microwave Observations at Zwalmbeek and Coet-Dan', in *Spatial Patterns in Catchment Hydrology: Observations and Modelling*, Grayson, R. & Blöschl, G. eds., Cambridge University Press, Cambridge, pp.187-208.

Tsotsos, J.K., Culhane, S.M., Kei Wai, W.Y., Lai, Y., Davis, N. & Nuflo, F. 1995, 'Modeling visual attention via selective tuning', *Artificial Intelligence,* vol.78, no.1-2, pp.507-545.

Udupa, J.K. & Saha, P.K. 2003, 'Fuzzy connectedness and image segmentation', *Proceedings of the IEEE,* vol.91, no.10, pp.1649-1669.

Ulaby, F.T., Dubois, P.C. & van Zyl, J. 1996, 'Radar mapping of surface soil moisture', in *Soil moisture theories and observations.*, vol.184 (1-2), Georgakakos, K.P. ed., Elsevier, Amsterdam, Netherlands, pp.57-84.

Ulichney, R. 1987, *Digital halftoning*, MIT Press, Cambridge, Mass.

Veltkamp, R.C. 2001, *Shape matching: Similarity measures and algorithms*, Technical Report, Utrecht University, Utrecht, The Netherlands, UU-CS-2001-03.

Veltkamp, R.C. & Hagedoorn, M. 1999, *State-of-the-art in shape matching*, Technical Report, Utrecht University, Utrecht, The Netherlands, UU-CS-1999-27.

Verhoest, N.E.C., Troch, P.A., Paniconi, C. & De Troch, F.P. 1998, 'Mapping basin scale variable source areas from multitemporal remotely sensed observations of soil moisture behavior', *Water Resources Research,* vol.34, no.12, pp.3235-3244.

Wealands, S.R., Grayson, R.B. & Walker, J.P. 2003, 'Hydrologic Model Assessment from Automated Spatial Pattern Comparison Techniques', *In: MODSIM 2003 International Congress on Modelling and Simulation*, Post, D. ed., Modelling and Simulation Society of Australia and New Zealand Inc., July 2003, Townsville.

Wealands, S.R., Grayson, R.B. & Walker, J.P. 2005a, 'Quantitative comparison of spatial fields for hydrological model assessment--some promising approaches', *Advances in Water Resources,* vol.28, no.1, pp.15-32.

Wealands, S.R., Grayson, R.B. & Walker, J.P. 2005b, 'Quantitative measures for the local similarity of hydrological spatial patterns', *In: 29th Hydrology and Water Resources Symposium*, The Institute of Engineers, Australia, 21-23 February, Canberra.

Wealands, S.R., Grayson, R.B., Walker, J.P. & Blöschl, G. 2004, 'Quantitative measures for the local similarity of hydrological spatial patterns', *In: 2nd international CAHMDA workshop on: The Terrestrial Water Cycle: Modelling and Data Assimilation Across Catchment Scales*, Teuling, A.J., Leijnse, H., Troch, P.A., Sheffield, J. & Wood, E.F. eds., 25-27 October, Princeton, pp.40-43.

Western, A.W., Blöschl, G. & Grayson, R.B. 1998, 'Geostatistical characterisation of soil moisture patterns in the Tarrawarra catchment', *Journal of Hydrology,* vol.205, pp.20-37.

Western, A.W. & Grayson, R.B. 2000, 'Soil Moisture and Runoff Processes at Tarrawarra', in *Spatial Patterns in Catchment Hydrology: Observations and Modelling*, Grayson, R. & Blöschl, G. eds., Cambridge University Press, Cambridge, pp.209-246.

Western, A.W., Grayson, R.B. & Blöschl, G. 2002, 'Scaling of soil moisture: A hydrologic perspective', *Annual Review of Earth and Planetary Sciences,* vol.30, pp.149-180.

Western, A.W., Grayson, R.B., Blöschl, G., Willgoose, G.R. & McMahon, T.A. 1999a, 'Observed spatial organization of soil moisture and its relation to terrain indices', *Water Resources Research,* vol.35, no.3, pp.797-810.

Western, A.W., Grayson, R.B. & Green, T.R. 1999b, 'The Tarrawarra project: high resolution spatial measurement, modelling and analysis of soil moisture and hydrological response', *Hydrological Processes,* vol.13, no.5, pp.633-652.

Western, A.W., Zhou, S., Grayson, R.B., McMahon, T.A., Blöschl, G. & Wilson, D.J. 2004, 'Spatial correlation of soil moisture in small catchments and its relationship to dominant spatial hydrological processes', *Journal of Hydrology,* vol.286, no.1-4, pp.113-134.

White, R. 2006, 'Pattern Based Map Comparisons', *Journal of Geographical Systems,* vol.8, pp.145-164.

Wilson, D.J., Western, A.W. & Grayson, R.B. 2005, 'A terrain and data-based method for generating the spatial distribution of soil moisture', *Advances in Water Resources,* vol.28, no.1, pp.43-54.

Woodcock, C. & Harward, V.J. 1992, 'Nested-Hierarchical Scene Models and Image Segmentation', *International Journal of Remote Sensing,* vol.13, no.16, pp.3167-3187.

You, J., Li, W.X. & Zhang, D. 2002, 'Hierarchical palmprint identification via multiple feature extraction', *Pattern Recognition,* vol.35, no.4, pp.847-859.

Zepeda-Arce, J., Foufoula-Georgiou, E. & Droegemeier, K.K. 2000, 'Space-time rainfall organization and its role in validating quantitative precipitation forecasts', *JGR Atmospheres,* vol.105, no.D8, pp.10129-10146.

Appendix A

Tutorial for comparison tool

A.1 Overview

A software comparison tool has been developed to enable the comparison methods described throughout this thesis to be computed and analysed. This tool has been developed using the Catchment Modelling Toolkit (Argent et al. 2003; eWater CRC 2006) and is a prototype that can be developed in future research. This tool is provided here for the reader to experiment with, using the sample fields provided and/or their own spatial field data. The software is only a prototype, but it is representative of the type of interface and functionality expected in a comparison tool.

The interface is designed to facilitate making the subjective decisions and defining any parameters needed for the comparison, rather than actually doing the pre-processing (which can be done externally using other tools). However, pre-processing tools for standardising, equalising histograms and data-driven segmentation are accessible through the interface. Other pre-processing methods (e.g. terrain analysis, noise filtering) are not yet developed into the interface, but some are available in other Catchment Modelling Toolkit products.

A.2 Installation

The software has been written in C#.NET and is supplied as an executable file and a variety of associated dynamic link libraries (DLLs). The executable requires the Microsoft .NET Framework Version 1.1 (or greater) to be installed onto a Microsoft Windows operating system. Many applications already use this framework and it has often been installed as part of another application (therefore not requiring reinstallation).

A.2.1 Installing the comparison tool

The comparison tool is provided on the CD or as a zipped file containing one executable file and eleven DLLs from http://www.civenv.unimelb.edu.au/wealands/tool.zip. This file can be unzipped into any target directory.

To run the software, execute the file named comptool.exe that is in the CD directory (\tool) or the target directory chosen previously. If an error message is shown, the .NET framework will need to be installed.

A.2.2 Installing the .NET Framework

The .NET framework can be installed from the CD directory (X:\dotnet\) or alternatively from http://www.microsoft.com/downloads (search for ".net framework"). To install, execute the file named dotnetfx.exe and follow the prompts. After installation, the comparison tool can be executed again and should run.

A.2.3 Troubleshooting

If there are any problems encountered with the software or installation, the reader is encouraged to contact the author for assistance (mailto:sweal@unimelb.edu.au), although the availability of technical support may be limited.

This software tool is provided primarily for the experimentation of the interested reader and should not be considered a final, packaged product. It has limited error handling and is yet to be optimised for performance with large spatial fields (i.e. more than 100,000 elements). A brief tutorial is given in the following section to demonstrate the use of the interface with a range of sample fields provided.

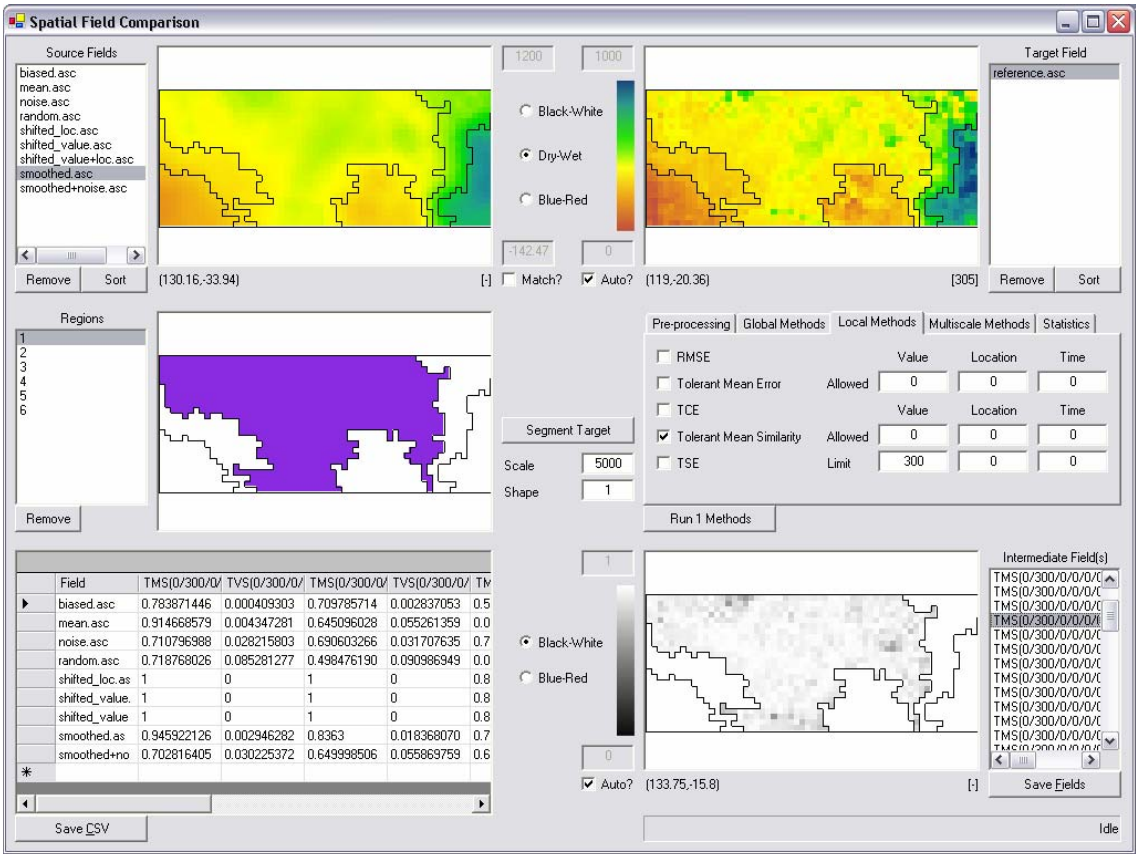
A.3 Tutorial

A.3.1 Sample spatial fields

Two sets of spatial fields have been supplied for the reader to use when experimenting with different comparison methods. They are the synthetic fields from Chapter 4 (\data\synth*\) and the soil moisture fields from Tarrawarra used in Chapter 5 (\data\tarra*\). These fields are provided on the CD (\data) or are downloadable from http://www.civenv.unimelb.edu.au/wealands/data.zip. A readme.txt file in each directory details what each directory and sample spatial field represents and where it has been used in the thesis. The reader may also use their own spatial field data within this tool. The fields must be in Arc ASCII raster format and have comparable headers to be used in the tool.

A.3.2 Interface

The comparison tool interface is comprised of six main sections, each representing a different part of the comparison process. Figure A.1 shows the appearance of the interface when populated with spatial fields and comparison results. Table A.1 identifies the major purpose of each section of the interface, detailing the interactions available to the user. The interface uses a drag-and-drop approach for inputting the source, target and region fields. The other parameters (e.g. tolerances, scales) are entered using the keyboard via simple text boxes arranged on the tab control.


Spatial Field Comparison

Source Fields
biased.asc
mean.asc
noise.asc
random.asc
shifted_loc.asc
shifted_value.asc
shifted_value+loc.asc
smoothed.asc
smoothed+noise.asc

Target Field
reference.asc

1200 1000 -142.47 0
○ Black-White
● Dry-Wet
○ Blue-Red

(130.16,-33.94) [-] □ Match? ☑ Auto? (119,-20.36) [305]

Regions
1
2
3
4
5
6

Segment Target
Scale 5000
Shape 1

Pre-processing | Global Methods | Local Methods | Multiscale Methods | Statistics

		Value	Location	Time
□ RMSE				
□ Tolerant Mean Error	Allowed	0	0	0
□ TCE		Value	Location	Time
☑ Tolerant Mean Similarity	Allowed	0	0	0
□ TSE	Limit	300	0	0

Run 1 Methods

	Field	TMS(0/300/0/	TVS(0/300/0/	TMS(0/300/0/	TVS(0/300/0/	TM
▶	biased.asc	0.783871446	0.000409303	0.709785714	0.002837053	0.5
	mean.asc	0.914668579	0.004347281	0.645096028	0.055261359	0.0
	noise.asc	0.710796988	0.028215803	0.690603266	0.031707635	0.7
	random.asc	0.718768026	0.085281277	0.498476190	0.090986949	0.0
	shifted_loc.as	1	0	1	0	0.8
	shifted_value.	1	0	1	0	0.8
	shifted_value	1	0	1	0	0.8
	smoothed.as	0.945922126	0.002946282	0.8363	0.018368070	0.7
	smoothed+no	0.702816405	0.030225372	0.649998506	0.055869759	0.6
*						

Save CSV

1 0
○ Black-White
○ Blue-Red
☑ Auto? (133.75,-15.8) [-]

Intermediate Field(s)
TMS(0/300/0/0/0/0
TMS(0/300/0/0/0/0
TMS(0/300/0/0/0/0
TMS(0/300/0/0/0/0
TMS(0/300/0/0/0/0
TMS(0/300/0/0/0/0
TMS(0/300/0/0/0/0
TMS(0/300/0/0/0/0
TMS(0/300/0/0/0/0
TMS(0/300/0/0/0/0
TMS(0/300/0/0/0/0
TMS(0/300/0/0/0/0

Save Fields

Idle

Figure A.1 The interface of the prototype comparison tool, illustrating the appearance of the tool after performing analysis.

Table A.1 A listing of the purpose of each section of the interface shown in Figure A.1. The interactions available to the user within each section are listed. In general, the user controls the meaning of the comparison by the fields and parameters defined in sections 1-4, while the numerical and graphical results are presented in sections 5-6.

1. Input and display source fields · User can drop files onto graphical control · Fields are added if extent and element size matches existing fields in list · Symbology can be defined by selecting colour ramp and min/max values · Defined regions are overlain on display	**2. Input and display target field** · Same operation as section 1 · Temporal representation is not implemented, so only one target field can be added · Matching symbology between source and target aids visual comparison · Target field can be segmented to make regions
3. Input and display regions of comparison · Regions must be chosen using section 4 (e.g. source/target extents or custom regions) · If custom regions are chosen, a region field can be dropped onto the interface · Alternatively, the current target field can be segmented using the scale/shape parameters	**4. Define comparison methods and parameters** · Use check boxes to select comparison methods · Use text boxes to define parameters such as tolerance and scale for the associated methods · Comparison methods are applied between each source field and the target field · Some parameters can introduce long run times
5. Tabulation of comparison measures · Comparison measures are produced for every source/target/method combination · Each row is a different source field · Each column represents a particular method, set of parameters and region · Table can be exported to CSV format	**6. Intermediate fields from comparisons** · Any intermediate field (e.g. residual field) produced is shown here and labelled · Symbology can be defined automatically or manually using colour ramps and values · Intermediate fields can be saved to ASC format

A.3.3 Step-by-step analyses

The following step-by-step tutorials show how to perform some different analyses using the comparison tool and the sample fields. The instructions point out which part of the interface to work with and what selections to make.

A.3.3.1 Calculating current measures (BIAS, COE, RSQ, RMSE, MAE)

1. Select the 9 source fields (\data\synth_source\*.asc) in Windows Explorer and drag onto the ‘Source Fields’ display box.

2. Select the target field (\data\synth_targ\observed.asc) in Windows Explorer and drag onto the ‘Target Field’ display box.

3. Check the ‘Match?’ option to make the colour schemes match between the source and target fields. This helps with visual comparison.

4. Choose the region(s) of analysis under the ‘Pre-processing’ tab. For this analysis, select ‘Common Extents’ (i.e. the intersection of extents between both fields).

5. Choose the comparison methods used for different scales under the ‘Global Methods’ and ‘Local Methods’ tabs. For this analysis, check the ‘Bias’, ‘RSQ Correlation’, ‘COE (Nash-Sutcliffe)’ options under the ‘Global Methods’ tab. Check the ‘RMSE’ and ‘Tolerant Mean Error’ options under the ‘Local Methods’ tab. Tolerant Mean Error is the same as MAE when no tolerances are set (i.e. they are all set to zero).

6. Click the ‘Run’ button to perform each checked comparison method for each source field against the target field. The numerical results are shown in the table, while the intermediate fields used in calculating the two local measures can also be viewed.

7. The mouse cursor can be hovered over the displayed fields to reveal the coordinates and value. By clearing the ‘Auto?’ option, custom values for the colour ramps can also be entered.

8. The tabulated and intermediate field results of the analysis can be saved by clicking the ‘Save CSV’ and ‘Save Fields’ buttons.

A.3.3.2 Calculating region-based measures

1. Select and drag-drop the source and target fields. Use the same fields as in the previous analysis (A.3.3.1).

2. Choose the ‘Custom Extents/Regions’ from the ‘Pre-processing’ tab. This allows a custom set of regions to be dropped onto the display, or alternatively enables the segmentation button to segment the current target field (using the defined parameters).

3. Set the scale parameter at ‘5000’ and the shape parameter at ‘0.75’. These values produce a general segmentation of the target field used in this analysis. Click the ‘Segmentation’ button to segment the target field. This should take less than 5

seconds on a Pentium 4 PC with 1 Gb RAM. Any regions that are not important can be removed using the 'Remove' button. Remove the 'Cust' region, which is the combination of all regions.

4. Check the 'Absolute Bias' option under the 'Global Comparison tab. Be sure to clear any other check boxes for the other comparison methods so that they are not computed.

5. Click the 'Run' button to calculate the 'Absolute Bias' for each region of each source field. The numerical results are tabulated (one column for each region). No intermediate fields are produced because the comparisons calculated were only 'Global Methods'.

A.3.3.3 Calculating multiscale measures

1. Select and drag-drop the source and target fields. Use the same fields as in the previous analysis (A.3.3.1).

2. Choose the 'Common Extents' from the 'Pre-processing' tab. This ensures that the multiscale comparison method is applied to the whole field.

3. Check the 'KS Similarity' option on the 'Multiscale Methods' tab. The parameter values define the different radii used when upscaling (specified as the number of field elements). Large scales will take much longer to compute due to large number of elements being summarised at each scale.

4. Set the parameters for minimum scale, scale step and maximum scale to 0, 2 and 6 respectively. This produces measures for scales of 0, 2, 4 and 6 element radii. The upscaling can be limited to only the source or target fields using the check boxes, but apply upscaling to both in this analysis.

5. Clear any other comparison methods so that they are not also calculated. It is possible to do multiple comparison methods on the one run, but this is avoided here for simplicity.

6. Click the 'Run' button to calculate the 'KS Similarity' value for each scale between each source field and the target field. This took less than 15 seconds (on

a Pentium4 PC with 1 Gb RAM) to calculate the 4 different scales for the 9 source fields.

A.3.3.4 Calculating tolerant measures

1. Select and drop the ‘shifted’ and ‘biased’ source fields (\data\synth_source\) onto the ‘Source Fields’ display box. Use the same target field as in A.3.3.1. Alternatively, selecting a field that is no longer needed and click the ‘Remove’ button.

2. Choose the ‘Common Extents’ from the ‘Pre-processing’ tab.

3. Check the ‘Tolerant Mean Similarity’ option on the ‘Local Methods’ tab. A set of different tolerances will be specified in this analysis and the results from the comparisons will be discussed. The tolerances are specified in the same units as the underlying data. In this analysis, the element values range from 0-1000 and the element size is 0.5 degrees. Time tolerances are not yet implemented through this interface.

4. Leave the tolerances set to 0 and ‘Run’ the comparison. Visually inspect the intermediate fields and note that all fields have a small section that is different (while the rest of the field is identical).

5. Set the value tolerance parameters to ‘allow’ = 0 and ‘limit’ = 100. This causes any difference between element values <100 to be considered as similar (using a linear tolerance definition). Click ‘Run’ and visually inspect the intermediate fields. Similarity (i.e. values greater than 0) is now found between some of the elements, except those where the values have been shifted by more than 100.

6. Reset the value tolerance back to 0 and set the location tolerance parameters to ‘allow’ = 1 and ‘limit’ = 2. Click ‘Run’ and inspect the intermediate fields. Similarity is now found between half of the shifted elements in the ‘locally_shifted’ field.

7. Set the value tolerance to ‘allow’ = 0 and ‘limit’ = 100, while leaving the location tolerances set as in the previous step. Click ‘Run’ and inspect the intermediate fields. Different amounts of similarity are now found between a number of

elements and the numerical summary measures reflect this increased similarity (due to increasingly relaxing the strictness of comparison).

A.4 Summary

The prototype comparison tool introduced and described here is a useful example of how hydrologists could better undertake comparison of spatial fields. It encourages the user to make essential decisions about how the comparison is calculated. The short tutorial shows how the different comparison methods can be applied using the tool. These methods can be extended and combined together to make more specialised comparisons (e.g. using tolerant measures within segmented regions). The synthetic fields from Chapter 4 are used throughout the step-by-step analyses, while the fields from Tarrawarra (Chapter 5) are provided for further experimentation by the reader. The reader may also choose to experiment with their own spatial field data.

Appendix B

List of acronyms and symbols

Acronym or symbol	**Description**
TDR	Time domain reflectometry
GLUE	Generalised likelihood uncertainty estimation
GIS	Geographic information system
DEM	Digital elevation model
SAR	Synthetic aperture RADAR
x_i	Location (x) of element i
$f(x_i)$	Value of element i
MEAN	Mean
SDEV	Standard deviation (σ)
SKEW	Skewness
KURT	Kurtosis
BIAS	Bias
$\gamma(h)$	Semivariance
RMSE	Root mean squared error
MAE	Mean absolute error
RSQ	Coefficient of determination (R^2)
COE	Coefficient of efficiency
PCOR	Percentage correct
KAP	Kappa index
KS	Kolmogorov-Smirnov similarity measure
V, L, T	Value, location, time
CTOL, FTOL	Crisp tolerance, fuzzy tolerance
SV, SL, ST	Similarity for each component of an element – value, location and time
EV, EL, ET	Error for each component of an element – value, location and time
ELOC, SLOC	Local error, local similarity
$\Delta V, \Delta L, \Delta T_{ALLOW}$	Allowed difference in value, location and time
$\Delta V, \Delta L, \Delta T_{LIMIT}$	Limit of difference in value, location and time

Acronym or symbol	Description
TMAE	Tolerant mean absolute error
TCOE	Tolerant coefficient of efficiency
TMS	Tolerant mean similarity
TSE	Tolerant similarity efficiency
|BIAS|	Absolute bias
K	Scale weighting parameter
R	Radius of elements included in multiscale comparison
MSME	Multiscale measure of error
MSMS	Multiscale measure of similarity
G E	Group of ‘global error’ measures
G S	Group of ‘global similarity’ measures
L E	Group of ‘local error’ measures
L S	Group of ‘local similarity’ measures
TL E	Group of ‘tolerant local error’ measures
TL S	Group of ‘tolerant local similarity’ measures
M E	Group of ‘multiscale error’ measures
M S	Group of ‘multiscale similarity’ measures
R E	Group of ‘region-based error’ measures
R S	Group of ‘region-based similarity’ measures
SL E	Group of ‘specific local error’ measures
SL S	Group of ‘specific local similarity’ measures
STL E	Group of ‘specific tolerant local error’ measures
STL S	Group of ‘specific tolerant local similarity’ measures
% v/v	Percentage volumetric soil moisture
SVET	Spatially variable evapotranspiration
K_{SAT}	Saturated hydraulic conductivity
K_{DEEP}	Deep seepage
D_{SOIL}	Total soil depth
D_{UPP}	Depth of laterally transmissive upper soil layer
CRC	Cooperative Research Centre
TIME	The Invisible Modelling Environment
DLL	Dynamic link library