Available online at www.sciencedirect.com
SCIENCE DIRECT®
Advances in Water Resources 28 (2005) 15–32
Advances in Water Resources
www.elsevier.com/locate/advwatres

Quantitative comparison of spatial fields for hydrological model assessment—some promising approaches

Stephen R. Wealands [a,b,*], Rodger B. Grayson [a], Jeffrey P. Walker [b]

[a] *Cooperative Research Centre for Catchment Hydrology, The University of Melbourne, Australia*
[b] *Department of Civil and Environmental Engineering, The University of Melbourne, Parkville, Vic. 3010, Australia*

Received 22 January 2004; received in revised form 22 September 2004; accepted 7 October 2004

Abstract

The current practice for assessing spatial predictions from distributed hydrological models is simplistic, with visual inspection and occasional point observations generally used for model assessment. With the increasing availability of spatial observations from remote sensing and intensive field studies, the current methods for assessing the spatial component of model predictions need to advance. This paper emphasises the role that spatial field comparisons can play in model assessment. A review of the current methods used in hydrology, and other disciplines where spatial field comparisons are widely used, reveals some promising methods for quantitatively comparing spatial fields. These promising approaches—segmentation, importance maps, fuzzy comparison and multiscale comparison—are for local comparison of spatial fields. They address some of the weaknesses with the current approaches to spatial field comparison used in hydrological modelling and, in doing so, emulate some aspects of human visual comparison. The potential of these approaches for assessing spatial predictions and understanding model performance is illustrated with a simple example.

Keywords: Spatial; Fields; Patterns; Comparison; Quantitative; Model assessment

1. Introduction

Distributed hydrological models produce spatially explicit predictions that allow more detailed analysis in decision-making than lumped models. Managers in the environmental field can now not only query the magnitude of a hydrological attribute, they can also query the spatial distribution of the attribute and ask 'where' type questions. The presence of spatial predictions has grown out of the increased availability of spatial data sets and cheaper computing power required to process these data [20]. However, there are issues relating to the uncertainty in such predictions due to uncertainty in model inputs and structure. Quantifying the uncertainty in these predictions has been the subject of continued research and debate, due to the large number of degrees of freedom inherent in these models [5,36,51]. Recognition of the limitations with distributed hydrological modelling has resulted in several general methodologies for assessing uncertainty being proposed. Methodologies such as generalised likelihood uncertainty estimation (GLUE) [4] and the 'alternative blueprint' [3], which can address the limitations while still utilising the strengths of distributed hydrological models, focus on trying to quantify the uncertainty in the predictions made [37,59]. These methods use many models and parameter sets that could represent 'reality' to make predictions. Model and parameter combinations that do

* Corresponding author. Address: Cooperative Research Centre for Catchment Hydrology, Department of Civil and Environmental Engineering, The University of Melbourne, Melbourne, Vic. 3010, Australia.
E-mail address: srweal@civenv.unimelb.edu.au (S.R. Wealands).

0309-1708/$ - see front matter
doi:10.1016/j.advwatres.2004.10.001

not 'fit' the observations are termed 'non-behavioural' and are rejected. The 'more likely' parameter sets (and/or models) remain and are used to provide the measure of uncertainty.

Grayson et al. [22] point out that in response to these methodologies for assessing uncertainty and numerous calls for data collection, spatial observations for assessing distributed hydrological models are becoming increasingly available. Furthermore, advances in remote sensing are providing improved spatial and temporal measurements of hydrological attributes that are of increasing value [54]. Spatial fields of hydrological attributes—for example, soil moisture [40,69], snow cover [46,58], saturated area [18,24], runoff [67], erosion [57], precipitation [17,50] and ocean suspended sediment [62]—have been observed and predicted for various study sites. These studies have provided insights about the hydrological processes involved and their function under different conditions, but the tools required to utilise such data have not developed accordingly. As such, spatially-distributed models are still being assessed using the more readily available point measurements (which often represent an integrated response of a larger area). These point measurements can be replicated using many different spatial fields, which makes them poor for constraining the distributed predictions [23].

At present, the value of observed spatial fields for distributed hydrological modelling has been recognised and the use of data from remote sensing and improved field measurements continues to grow. To fully realise the potential of spatial fields for model assessment, the absence of appropriate comparison methods must be addressed [20,22,35]. This paper defines spatial fields as used in hydrology and then reviews the common ways that they have been used in assessing model predictions. Where comparisons of observed and predicted spatial fields are undertaken, we focus on the methods used for comparison and the information thus garnered. The dominant characteristics of human visual comparisons are identified, with a view to emulating these with quantitative comparison methods. Approaches to comparison from the broader image- and pattern-related literature shows how other disciplines approach the problem of comparison. Drawing from these disciplines, some promising methods for quantifying the comparison of spatial fields are detailed. The potential of these methods for providing quantitative measures useful for hydrologic interpretation are illustrated with a simple example and discussed in reference to their use in hydrological model assessment.

2. Observed spatial fields in hydrology

Spatial fields are being increasingly generated in hydrological studies, via both observation and model simulation. Spatial fields are primarily used for model input, but with increasing data availability, they are also being used for model assessment. Spatial observations are usually made at variably spaced points and then interpolated onto a regular grid to produce a complete spatial field. Both the density of the observations and the interpolation method used contribute to how representative the observed spatial field is of reality. Where sufficient point samples are made to represent the spatial field of interest, then the interpolation step can be avoided. For example, if a spatial observation is made for every model element, then this spatial data may be sufficient for assessing the model. When spatial observations are obtained via remote sensing, the spatial field is represented with a regular grid, having a resolution (or pixel size) that defines the density of observation points. Spatial models in hydrology can be based on both regular grids and unstructured networks. In all cases, the model domain is discretised into model elements that have a spatial link to neighbouring elements. When comparing observed and predicted spatial fields, it is desirable for them to be commonly discretised (i.e. have the same structure and resolution). This allows any processing to be applied similarly to both data sets and ensures that spatially coincident values are compared. Throughout this paper, the spatial fields used in the discussion and demonstrations are regular grids. This is due to them being both computationally simple and common, thus making them ideal for presenting the methods.

Spatial observations are usually based on measurements of categorical data (e.g. presence/absence of snow cover [58], low/medium/high level of rill erosion [35]) or continuous data (e.g. soil moisture [69]). The data type is controlled by both the measurement method and logistical factors (e.g. time, personnel). In all spatial analysis tasks (including comparison), the data type determines the methods that can subsequently be applied for analysis [12], although a higher level data type can always be converted into a lower level data type (i.e. a continuous field can be categorised). In this paper, the methods discussed vary in their applicability, although we have attempted to focus on methods for continuous spatial fields (i.e. the higher level data type).

Hydrological spatial observations are obtained in different ways and encompass varied levels of processing (to produce the spatial fields from the raw measurements). In general, observed spatial fields are produced from exhaustive field measurements, remote sensing (such as satellites or ground-based radar) and/or surrogate data (that have correlation with the attribute of interest). Strictly speaking, all measurements are surrogates of some kind, yet those specifically referred to here have low correlations with the hydrological attribute being represented [22]. One of the most common surrogates used in hydrology is terrain, which can be used as a surrogate for soil wetness, solar radiation exposure, precipitation gradients and soil properties [70]. These varied sources all require some level of processing to produce a spatial field in a form that is useful for model input or assessment. This can involve interpolation from points to spatial fields, the reduction of sensor noise, application of retrieval algorithms requiring multiple input data, or deriving the relationship with a surrogate attribute. The processing alters the effective spacing and support of the raw information, acting as a kind of filter. This is discussed more fully by Blöschl and Grayson [6], but it is usual for there to be a tradeoff between the accuracy of a single point observation, the total number of points (i.e. spatial detail) observed and the number (or the extent) of the observations made.

Several case studies from hydrology that involve spatial fields exist and they illustrate the state-of-the-art for comparing spatial fields in hydrological applications [21,22]. Visual comparison is recognised as the most common and useful method, even though it cannot provide a quantitative measure of similarity and is labour-intensive. For quantitative methods, there are two general approaches used—characterisation followed by comparison (i.e. global comparison); and comparison followed by characterisation (i.e. local comparison). In the first approach, the information in the spatial field is characterised into a number or graph (e.g. summary statistics, geostatistical measures, landscape indices) and then the summary measures are compared numerically (to compute the global similarity). The second approach uses numerical comparison at every location (e.g. the residual between spatially coincident pixels) to produce an intermediate spatial field (that reflects the local similarity), from which the overall measure of similarity is then computed (e.g. the mean similarity).

There are many common methods used for global comparison, such as comparing the mean values, the correlation lengths or other indices that characterise the spatial field (for examples see [24,25,52]). In contrast, there are only limited methods used for local comparison. Most methods calculate residuals between spatially coincident pixels and then summarise them (as in a root mean squared error (RMSE) calculation). The other common local comparisons measure the correlation between all spatially coincident pixel values. More advanced local comparison methods are yet to be used in hydrology and are the methods of interest in this paper. In addition to these approaches to comparison (i.e. global and local), the type of similarity measure produced can be either absolute or relative. Absolute measures compare the actual values and are often used to provide a measure of the error (e.g. RMSE). Relative measures ignore the magnitude of the values, instead describing the 'fit' between the data sets (thus ignoring any bias) (e.g. correlation). Both of these types of similarity measures are useful, but need to be used together to fully interpret the comparison.

In most hydrological studies where observed spatial fields are being used, both global and local comparisons are computed. Global comparison is the first step, usually encompassing comparison of the means and variances of the spatial fields. After this, local comparison is commonly done by calculating the root mean squared error (i.e. a summary of the residuals for each data point in the spatial field). These methods are applied to the spatial fields just as they are applied with temporal data (i.e. hydrological time series), with a comparison occurring between each pair of data points. As has been recognised with the comparison of temporal data in hydrology, there are many different aspects to compare that may provide more useful information for assessing similarity (e.g. looking at neighbouring data points to recognise shifts, comparing regions of the time series rather than each point individually, assessing both absolute and relative similarity). Boyle et al. [9] have presented some specialised methods for comparing hydrological time series. In the same way, specialised methods to address aspects of spatial comparison could be developed, although they are not currently used in hydrology.

There have been some attempts to make more specialised comparisons with spatial fields in hydrology. Transects have been extracted through important parts of a spatial field. These can be useful for visual interpretation of lateral shifts in one-dimension, although they are hard to quantitatively compare. Another method called optimal local alignment [22], allows the investigation of lateral shifts between spatial fields. This is done by computing the correlation between patches within the observed and predicted fields for two-dimensional shifts. The vectors for the "optimal shift" (i.e. that resulting in the highest correlation) are then displayed, allowing the user to visually assess areas with consistent shifts and make inferences about the model (or data) deficiencies leading to consistent errors in particular locations. Again, this method lacks a metric to enable quantification of the comparison and use in automatic optimisation schemes. The promising methods discussed throughout the rest of this paper are all for local comparison and produce different similarity measures (including both absolute and relative).

3. Understanding and emulating human visual comparison

Human vision is widely regarded as the most powerful and comprehensive method for comparing spatial fields [22,26]. All modelling studies in which observed and predicted spatial fields are available use this approach. It is also used where only predicted fields are available, as a qualitative check on the plausibility of the predictions. In this situation, the comparison is made against the background knowledge and expectations of the observer. A visual comparison has a number of strengths and weaknesses. The most obvious strength is the simplicity with which a comparison is completed. Humans can observe, recognise and interpret spatial fields automatically, integrating their background knowledge and understanding of the spatial field being viewed. They can then compare two spatial fields and make a qualitative assessment of their similarity, exploiting the outstanding ability of the human brain to synthesise disparate information. The comparison will involve looking at overall similarity, the similarity of specific features and even the possible similarity of features if they were shifted or altered slightly. Yet, amongst all these strengths emerge the weaknesses with this approach. While the spatial field can be interpreted and observed, the observer can personally bias the interpretation and there are limits to the capacity of the brain to assess multiple images or large spatial extents. For example, a modeller who believes that there should be a connected, linear flow path in a particular study catchment would tend to weight such a feature highly in comparisons (i.e. by focussing on it more), while other components of the field such as distribution of moisture on a hillslope may be just as important in an assessment of model performance, yet not as obviously apparent. Visual comparison also takes a more general view of the spatial fields, disregarding individual model elements and focussing more on the relative differences within regions. These variable interpretations make visual comparisons difficult to replicate, impossible to quantify and relatively slow. As a result, there is only limited application of visual approaches for comparing multiple spatial fields (as would be required for rejecting 'non-behavioural' models from a large ensemble of model runs, or in an optimisation procedure). If we can obtain a better understanding of the processes being undertaken by the human visual system, it may be possible to emulate these computationally and address some of the weaknesses. While it is unrealistic to expect to emulate the human approach exactly, we can benefit from exploring the general methods used.

We take the ability to view and interpret for granted yet it involves remarkable computation and processing. Research into eye movement and visual attention has investigated the process occurring when humans view images, which are analogous to spatial fields in this case [33,41]. Humans are able to interpret complex scenes and then select a subset of the available sensory information for further processing. This process of simplifying the scene is useful to understand the way humans analyse images. Saccadic eye movements, where the eye jumps between different points of interest in the image, have been studied to identify characteristics of images that draw visual attention [15]. Using the characteristics of visual attention, a measure of the areas of an image that are of most interest to the human vision system is then produced [33,41]. This is commonly called an importance map, which can then be used to prioritise further processing.

Osberger and Maeder [41] synthesised research on visual attention to come up with a number of low-level (or simplistic) factors that determine the visual importance of regions in an image. After initially segmenting an image into homogeneous regions, the following factors were found to be instrumental in determining the perceptual importance of the individual regions—contrast of a region and its neighbours; region size; region shape; and the location of a region within the image. Other researchers have identified that the 'uniqueness of pixel values' is a powerful way to determine importance. Tompa et al. [60] used this approach with Shannon's self information measure which is detailed later in this paper. Itti et al. [33] have also used the approach with low-level (or simple) image features that were calculated for multiple spatial scales to help decide on the features of highest visual importance. It is evident that the human visual system works predominantly with features that command attention due to their intensity, size, shape, location or value. These aspects of visual importance are also expected to be dominant when visually comparing spatial fields in hydrology. As such, they will be useful approaches to emulate in automated comparisons methods.

Some methods for emulating visual comparisons with spatial fields use image features, while others work with pixel values. Depending on the process being emulated, the fundamental entity varies. Hay et al. [29] identifies that the concept of objects (or features) within the analysis is innate to humans and is something often lacking in approaches to emulate the way humans interpret images. From this we obtain the notion of 'image-objects', which are individual entities in an image that are generated from pixel groups [29]. Most existing methods in hydrology work only with pixels and perform limited analysis with objects, despite their obvious importance to visual processing. These objects may be thought of as "homogeneous regions" in hydrological spatial fields.

Once the features of interest in a spatial field are identified, whether these are pixels or objects, the task of comparison is then undertaken. Power et al. [48] acknowledge that humans intuitively identify a hierarchy of similarities between two images, by firstly noticing global similarity, then focussing on the finer details or local similarities. For example, an overall similarity in intensity (i.e. the mean value) is an initial global comparison, while the similarity of the high and low regions in both spatial fields (i.e. correlation) comprises local comparison. When comparing spatial fields, human imagination permits the objects identified to shift or change slightly to obtain a better match, focussing attention on the basic spatial structure and patchiness, rather than precise co-location of feature boundaries. This is captured by the idea of 'fuzziness' in the comparison and it is particularly powerful for making qualitative statements about how the spatial fields are different. Methods for emulating 'fuzziness' during comparison exist [26], with one approach being investigated later in this paper.

The ability to interpret and compare a spatial field at varying scales is innate to human vision. It involves intuitive recognition of the 'correct' scale at which to interpret an object, a task that requires extensive background information. For example, visual assessment of a spring snow cover field may involve looking at the broad scale for snow cover having some relationship to elevation, while at smaller scale expecting some effects on patchiness due to topographic aspect, while at still smaller scale expecting variations due to vegetation effects on wind/drift fields or albedo effects on snow melt. A multiple scale approach for image processing is used in [30,33] to perform analysis at a range of scales. Features (or measures) are derived from a set of lower-resolution versions of an original image. By conducting spatial field comparisons at multiple scales, the approach performed by humans can be emulated to some extent. Such a multiple scale approach is discussed later for use with the comparison methods presented in this paper.

4. Approaches to comparison in other disciplines

The task of producing a quantitative measure of similarity between two images (or spatial fields) is also encountered in other disciplines, ranging from image processing and pattern recognition to landscape ecology and multimedia. These other disciplines rely on a number of fundamental methods for processing images and conducting comparisons, some of which show promise for use in hydrological applications. However, the nature of the spatial fields present in hydrology must be considered when looking at other disciplines, as this is often the limiting factor in the application of their methods. Spatial fields in hydrology do not always have obvious features, they often have high levels of noise, they are representations of things that humans cannot immediately recognise and often contain only one band of data (making them synonymous with greyscale rather than colour images).

In other disciplines, the major comparison task is to find the images within a large database that are most similar to a query image. This task, which has applications in multimedia, criminology, art, engineering and science, relies on rapid comparison of the query image to the database images. There are numerous ways to achieve this task, which fall into the active area of research known as content-based image retrieval. This area demands fast processing and different types of comparisons, which has led to ongoing refinement of many common methods. These common methods have developed through the application of fundamental image processing operations to new areas.

4.1. Image processing for comparisons

The field of image processing encompasses many fundamental methods that are used in disciplines like content-based image retrieval. The major steps undertaken for image recognition and comparison problems are identified by Haralick and Shapiro [28]. Processing usually begins by conditioning the image, which includes normalisation, histogram operations and various types of filtering (for noise reduction or smoothing). The image histogram describes the frequency of pixel values occurring in the image. It is sometimes modified with conditioning methods, although more commonly, neighbourhood operators (or 'moving windows') are used to condition the pixel values (based on their neighbouring pixels). Noise reduction is a common task, in which individual pixels that are dramatically different from their neighbours are replaced by a value calculated from the neighbourhood (such as the average). Mastin [39] provides a review of various noise reduction methods using neighbourhood operators. For some comparisons, conditioning may be the only processing needed before computing the similarity measure (such as with a RMSE calculation). In other instances, the comparison may require structural features to be identified in the image.

Structural features can sometimes be defined by edges. Edge detectors identify the change in pixel values in certain directions within a neighbourhood. The edges can be calculated using many different approaches [8,28], although they are very sensitive to image noise and are rarely used with natural scenes (due to lack of clearly defined edges). One example in which edge detectors are used is for delineating flood inundation areas from synthetic aperture radar (SAR) imagery [31]. A more common approach for detecting structural features or regions (which ideally represent objects) is image segmentation. This is the process of partitioning an entire image into a set of non-overlapping regions. Segmentation can be achieved using knowledge-driven or data-driven approaches. In knowledge-driven approaches, the image is segmented to extract a predetermined type of target object, like finding individual tree crowns in a satellite image of a forest. In data-driven approaches, the image is segmented into regions (or image-objects) based on some additional information. Some of the common methods used for data-driven segmentation are thresholding, clustering, region growing and region merging. These, and other, methods are covered in reviews about segmentation and its various uses in different applications (see [10,34,42,63]). One of the important points from the broader literature on comparisons and segmentation is that there is no single "best method" [44,55], although some are more applicable to hydrological applications than others and these are described more fully below.

One of the most common segmentation methods is thresholding [28]. Thresholding involves defining a threshold value that is then used to classify each individual pixel as high or low. Fig. 1A shows a spatial field of soil moisture that has been grouped using a threshold of 38% volumetric soil moisture. These data are from the Tarrawarra data set presented in Western et al. [69]. 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 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. Pal and Pal [42] provide a comprehensive review of thresholding. Thresholding can also be applied during the data observation phase. In this case, the observations made are binary (e.g. snow or no snow, wet or dry) and thus produce a spatial field of binary values. Thresholds have been used to produce binary fields in a number of hydrological studies (e.g. [1,18]). Simple categorisation, of which thresholding is a subset, is an approach that will continue to be useful in hydrology due to its simplicity. Categorisation segments by grouping pixel values into pre-determined ranges, assigning pixels to categories using information from measurement (or feature) space. Subsequently, the pixels are grouped into connected regions using some form of connected components labelling [28]. For automated categorisation, the histogram of the spatial field is usually analysed and the category ranges are calculated based on the characteristics of the histogram. Fig. 1B shows a spatial field 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 categorisation. As this method works entirely on the pixel values, the resultant regions are not always spatially contiguous or without holes.

Another popular approach to segmentation, that also relies more on the pixel values (i.e. measurement space) and less on their spatial location (i.e. image space), is clustering. Clustering maps the pixel values from multiple bands of data into a feature space. From the feature space, clustering algorithms are used to identify groups of pixels with features that are compactly grouped and isolated from other features [34,43]. The most common clustering method is K-means clustering, in which the feature space is broken up into K clusters of data points (i.e. pixels). Each data point can then be moved around into other clusters if needed, until it remains in the cluster with the most similar mean value. Eventually, every data point in each cluster is closest to its cluster than to any other cluster, thus producing the optimal segmentation into K regions. Generally, the number of clusters to be detected has to be specified prior to the clustering. Depending on the method used, the resulting segmentation may be strong (with logical regions being grouped in the image) or weak (with many small, disjointed regions). Multi-spectral images provide a large number

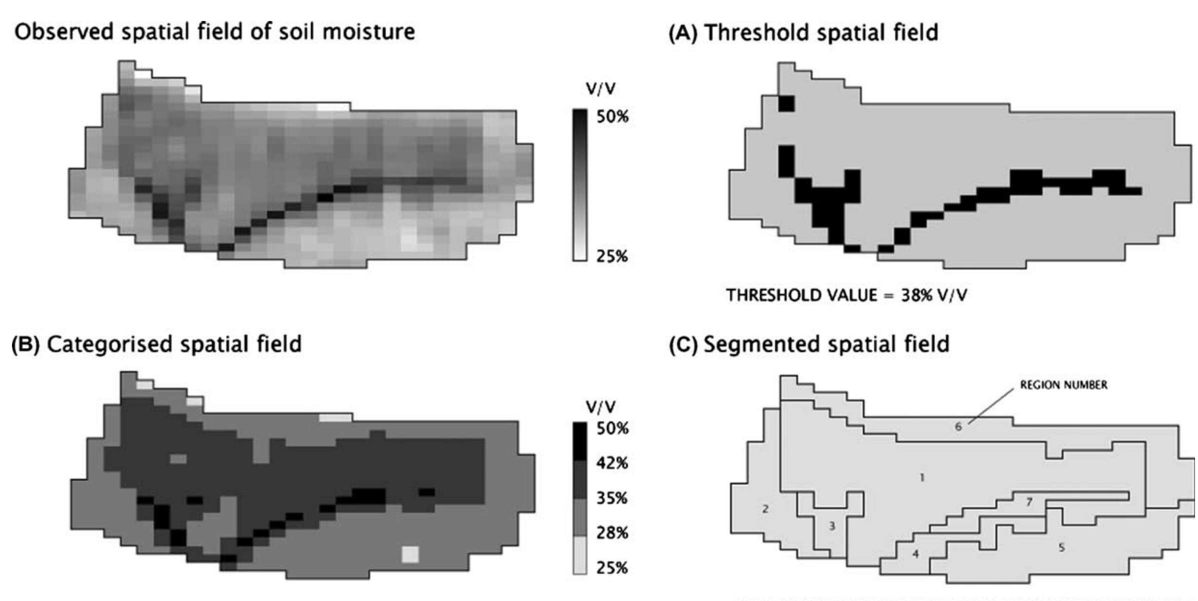

Fig. 1. Identification of regions within a spatial field of soil moisture using three alternate methods: (A) histogram thresholding; (B) histogram slicing; and (C) region merging.

of features (i.e. different image bands) and so are well suited to clustering approaches. With many spatial fields in hydrology, there are often only a small number of attributes predicted or observed and so fewer features can be identified for clustering and other approaches are preferred.

The most promising segmentation methods for spatial fields are those that work predominantly in image space and are data-driven. These methods then 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 (say based on joining pixels that differ in value less than a specified amount) until the entire image is segmented [10]. This method is sensitive to the choice of seeds and is not widely used. However, by considering every pixel as a seed (or splitting the entire image into small regions which become seed regions [10]), the method of region merging has evolved and become popular. Region merging tries to find the best possible merge between neighbouring regions (i.e. closest in value) at each iteration through the image [2,71]. Eventually, there are no more merges that pass the merge criteria and the segmentation is complete. Region merging is dependant on the criteria for homogeneity (which determines which merges are acceptable) and the order of processing (which determines which acceptable merges are done). Other criteria can also be used, which control the size or shape of the resultant segmented regions [2]. Region merging appears suited to a wider range of images than other segmentation methods, having proved successful with simple grey level images (which have limited measurement space data on which to base segmentations) as well as multispectral images. Fig. 1C shows a segmented spatial field of soil moisture, using the region merging method in Baatz and Schäpe [2]. This result is visually the most logical grouping of all examples shown in Fig. 1. It is also the most complex and requires more processing than the other methods, which is disadvantageous when analysing many spatial fields. The thresholding and categorisation methods shown in Fig. 1 are simple to process, although they are only useful when the processing is supervised (as automatic usage can produce highly unsuitable results).

4.2. Content-based image retrieval

Fast and accurate image comparisons are essential in content-based image retrieval (CBIR). One typical application is facial recognition, in which a single query image is provided (i.e. the face being photographed) and then a large database is searched for images with similar features. Facial recognition has to deal with 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. [49] 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 intensity values). 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 [72]. This refined set of pixels (often called the 'interesting' or 'salient' features in pattern literature) is computed using some knowledge of the image context, which is not generally so clearly defined in hydrological spatial fields. In analysing terrain models (which is closely related to many hydrological processes), Peucker and Douglas [45] developed methods for detecting salient points in elevation models. This work enables the detection of characteristic features (i.e. ridges, streams, peaks and pits) in a terrain surface and could possibly be applied to other hydrological spatial fields, provided that they also have obvious characteristic features.

The comparison stage requires the 'interesting features' from the query image to be compared against the sets of interesting features stored in the database, using a similarity metric. A common metric used in pattern matching research is the Hausdorff distance [32,66], which is basically a minimum distance fit between the two sets. There are numerous metrics for comparing two sets of features extracted from images, designed to maximise robustness to noise, shifts, scale changes and rotation [65]. These comparison methods are specifically for structural features defined in the images, so they are only useful for spatial fields with identifiable structural features (e.g. predicted river networks).

With images where the context is not known (such as large photo or art databases on the Internet), meaningful structural features are difficult to define. This lack of useful image structures means that segmentation is used for identifying 'homogeneous regions' or 'image objects'. Using these regions to represent the content of the image, the similarity of regions between a query and the catalogued images can be computed. Chen and Wang [11] used a 'clustering in feature space' approach to segment a set of general images. A number of fuzzy features (reflecting the colour, texture and shape properties) were then calculated to characterise each region. The fuzzy features provide 'blurry boundaries' for the regions, which help to handle some of the uncertainties introduced with segmentation. To do the comparison, the fuzzy features for coincident regions are summarised into a similarity measure. This approach shows how segmentation can be used to divide an image up into regions that can then be characterised from the pixel values in that region. Subsequently, the comparison uses the regions to compute a measure of similarity (as opposed to using the individual pixels). This idea is expanded upon in the discussion in application to hydrology.

4.3. Approaches from landscape ecology

Landscape ecology is a field allied to hydrology, in that it is concerned with the relationship between processes and patterns in the landscape. Both landscape ecology and hydrology use technologies such as geographic information systems (GIS) and remote sensing. However, work in landscape ecology deals with modelling and understanding landscape change and planning [61]. Through this work, there have been ongoing and recent developments that address some issues faced in hydrology when comparing spatial fields. Foody [16] provides a review of accuracy assessment for classified land cover maps. In this application area, the task is to provide a measure of how well a categorised map matches ground observations. While early attempts used visual approaches and were unable to quantify the accuracy, in recent times the confusion matrix is commonly used. This is a cross-tabulation of the mapped class label against the class label measured at particular locations (the ground truth). For each class X, the matrix contains the number of pixels correctly classified as X, as well as the number of pixels incorrectly classified into each other class. The confusion matrix provides a starting point from which classification accuracy can be calculated, as well as providing some evidence of which classes have the most error. From this, the percentage of pixels correctly classified can be measured. While there is no standard approach for accuracy assessment in landscape ecology, the kappa coefficient of agreement [13] is increasingly recognised as the type of measure needed [56]. This measure is derived from the confusion matrix, but accounts for the possibility of chance agreement between classes and produces a similarity measure adjusted for such agreements. Until recently, these measures had been used only for accuracy assessment using discrete samples. The kappa statistic is used to compare two categorical maps, effectively using every pixel as a sample point for the assessment. Some variations on the standard kappa statistic have been developed to quantify how much of the error is due to categorical differences and locational errors [47].

Power et al. [48] further developed the methods from accuracy assessment for the comparison of land use maps. The standard approach for assessing land use change is to conduct a pixel-by-pixel comparison and produce a confusion matrix. Power et al. [48] recognise that this is prone to showing registration errors as land use change, when in reality the spatial fields are the same (but they are displaced). To adjust for this, they use a fuzzy areal intersection approach, in which the similarity between overlapping areas is calculated based on the overlap and fuzzy relationships between different categories. Hagen [26] provides an extension to this approach that allows for fuzziness of both location and category. A variant of this method has been used by Güntner et al. [24] in hydrology and is discussed in more detail later. The ideas from landscape ecology (and related disciplines) have many similarities to those faced in hydrology, although the map comparisons all deal with categorical spatial fields. Many spatial fields in hydrology contain continuous values, and while these can be categorised (e.g. high, medium, low values), adapting these methods for continuous values may prove useful.

5. Promising comparison approaches for hydrological spatial fields

The approaches used in other disciplines for image processing and comparison provide useful ideas for application with hydrological spatial fields. Some of the approaches focus specifically on pre-processing the fields to be in a 'more comparable' form (e.g. by identifying regions or important features), while other approaches focus on new ways of conducting a local comparison between spatial fields (by doing fuzzy or multiscale comparisons). Individually these methods do not address the objective of this research, which is to quantitatively compare spatial fields. However, when used together and with existing methods, they offer approaches that can compare new and different aspects of spatial fields, thus providing additional information when assessing overall similarity. The combination of methods is expanded on later, while in the next section we focus on the individual components.

5.1. Identifying homogeneous regions

Identifying regions (or image-objects) within spatial fields provides a way of representing the fields as a collection of regions rather than pixels. Each pixel usually has only a single value, whereas a region has a range of values and the associated distribution and statistics. This additional information is useful for characterising a large area and is more closely related to the way humans visually interpret spatial fields. Haralick and Shapiro [27] introduce many segmentation methods that have seen further development in recent years. The approach that seems most suited for hydrological fields, due to their similarity with single-band remotely sensed images, is region merging. The following description of segmentation via region merging is a combination of two similar methods [2,71]. The major differences between these two methods are the way they control the merging order and the homogeneity criteria used.

Region merging is generally done by initially considering every pixel within a spatial field as an individual one-pixel region. Larger starting regions can also be used, but this often results in 'artificial looking' segmentation results. A distance metric is required as the measure of difference between the measurement spaces of two regions (e.g. the difference between the mean value of each region and the pixel under consideration for merging). This metric, which is used to assess possible merges, is calculated using the intensity of the values in the region. Woodcock and Harward [71] also include a size criterion to produce merged regions that are not too small or large. This ensures that multiple regions are formed in areas with low variance, as well as eliminating undesirable small regions in areas with high local variance. Baatz and Schäpe [2] use additional criteria relating to the shape of the merged regions. The shape criteria are composed of compactness and smoothness, each of which is calculated as some ratio of the region border length against the region size (a circle being the most compact shape). Baatz and Schäpe [2] also enforce a 'scale parameter' on their segmentation, which determines when merging will stop, therefore controlling the size of the segmented regions. This overall scale parameter is similar to the size criteria used by Woodcock and Harward [71] and is a control on region size (and shape) rather than the number of regions. The purpose of size and shape criteria is to make the resulting segmentation more visually appealing and more suited to the needs of the user.

Once the criteria are defined, region merging works through the spatial field, selecting one region and comparing it to its adjacent regions to decide on which merges should take place. Adjacency has been defined using a 4- or 8-neighbourhood approach. Woodcock and Harward [71] found that the 4-neighbourhood approach produced better segmentation results. Both the order in which regions are treated and the decision heuristics (i.e. the rules determining which merges take place) influence the resulting segmentation. To handle this, they work through the entire image and produce a list of the differences between all adjacent regions. The pair of regions with the lowest difference is then selected and merged, but only if the difference is below a global threshold. Once no adjacent regions pass the global threshold, then the segmentation is complete. This approach ensures that only the most similar regions are merged on each pass through the field. For large spatial fields (e.g. remotely sensed imagery), this type of segmentation is quite slow as there is only one merge for each pass through the image. Baatz and Schäpe [2] call this approach 'global mutual best fitting', but point out that it tends to produce very large regions in areas of low local variance and subsequently, small regions in areas of high local variance. Woodcock and Harward [71] eliminate this effect with their minimum and maximum region size thresholds.

Baatz and Schäpe [2] use a 'local mutual best fitting' for deciding on the most suitable merge. For a region A, this finds the most similar neighbouring region B. Before doing the merge, the most similar neighbour to region B is also found. If both tests decide that region A and region B are most similar then they are merged. If not, then the method searches again, this time with region B as the starting point. This was found to produce negligible quantitative differences to the global mutual best fitting, yet allows regions to grow equally in areas of high or low variance [2]. To control the order in which regions are assessed, a treatment order is created that first treats the region that is located furthest from all other treated regions. Both the treatment order and the merge criteria allow regions to grow in all areas of the spatial field at equal rates.

The approaches presented here describe the general methods and considerations when using region merging for segmentation. The data-driven approach can provide strong segmentations without the need for large number of parameters (and therefore large levels of user input). For example, the approach of Baatz and Schäpe [2] was used to segment the soil moisture spatial field in Fig. 1. The only segmentation parameters were the scale parameter (arbitrarily set to a unitless value of 25 to limit the maximum size of the regions) and the weighting for different merge criteria (set to 1 for intensity criteria and 0 for shape criteria to ensure segmentation based on intensity only). This segmentation shows logical regions without holes, unlike the spatial field categorised using only measurement space.

5.2. Characterising important features

When comparing spatial fields, we need to be careful that the entities being compared are those of most hydrological importance (as opposed to those that are the easiest to compute). The standard mean squared error (MSE) calculated in a pixel-by-pixel comparison is often used in hydrology and weights every pixel in the comparison equally. This can result in differences in pixels of lesser hydrological interest (e.g. ridge cells in a map of runoff producing areas) influencing the error measure in a negative way, while the areas of importance (e.g. the valleys in this same example) were actually quite good. This can be overcome by weighting pixels in the comparison using some criteria for importance. Two methods that deal with pixel and region weightings are detailed here. The first is based on information theory and calculates a measure of importance (or interest) for each pixel. The second models visual attention in images and looks at the characteristics of regions that make them visually important.

Tompa et al. [60] present a simple modification to the standard MSE statistic that takes into account the importance of pixels. This work was developed for comparing original and distorted images, with an aim to assess whether there was a perceptual loss of image information. First, an 'event' is chosen from the image that is the 'event of importance'. This is usually the pixel value (which must be categorised), but can also be a measure like local variance (which is calculated using a neighbourhood operation). For each event, the information content is then computed using a formula from information theory for information content [53]—equal to the logarithm of the inverse of the probability of the event occurring in the image. As such, infrequent events have high information content, while common events contain less information content. Using the logarithm ensures that the situations where the probability is equal to either one or zero are handled correctly. An information content map is then produced, which is used to weight the original pixel values during the standard MSE calculation. Fig. 2 shows examples of information content maps produced from a spatial field of soil moisture. In Fig. 2A, the pixel value is used as the event of interest, with the resultant weighted spatial field (obtained by multiplying the observed field by the importance map) highlighting the less common values and reducing the background detail. In Fig. 2B, the local variance of pixel values (within a 3×3 cell window) is used as the event. This approach favours the more variable areas in the spatial field, which is quite evident in the weighted result. Tompa et al. [60] calls the pixel-by-pixel MSE calculation based on the weighted maps the 'information' mean squared error (IMSE). The IMSE statistic provides a quantitative measure of similarity between compared fields. The relative nature of this measure makes it suited to ranking similarity between pairs of images, rather than providing an absolute measure of error (as the MSE does). Tompa et al. [60] found this quantitative method of comparison to agree better with subjective comparisons of distorted images than the standard MSE approach. For hydrological spatial fields, this approach may improve pixel-by-pixel comparisons that have previously produced poor results. For application to hydrology, it is important to consider the event of importance carefully, as this changes the context of the comparison. Further experience using this approach will identify the most suitable events for different hydrological spatial fields.

The approach used in Osberger and Maeder [41] uses the characteristics of regions in an image to characterise importance. The previous section highlights a way to obtain regions using segmentation. The regions produced, while representing the spatial field in an alternate way, do not have a specific category label. They do have characteristics such as mean pixel value, variance, centroid location, region size and distributional parameters. Osberger and Maeder [41] use some of these low level characteristics to define perceptual importance (i.e. the characteristics that make a region visually important). They found the major characteristics of perceptual importance to be:

- the contrast of the region against its neighbours (high contrast regions attract attention);
- the size of the region (larger regions, up to some saturation point beyond which they are ignored, attract more attention than smaller ones);

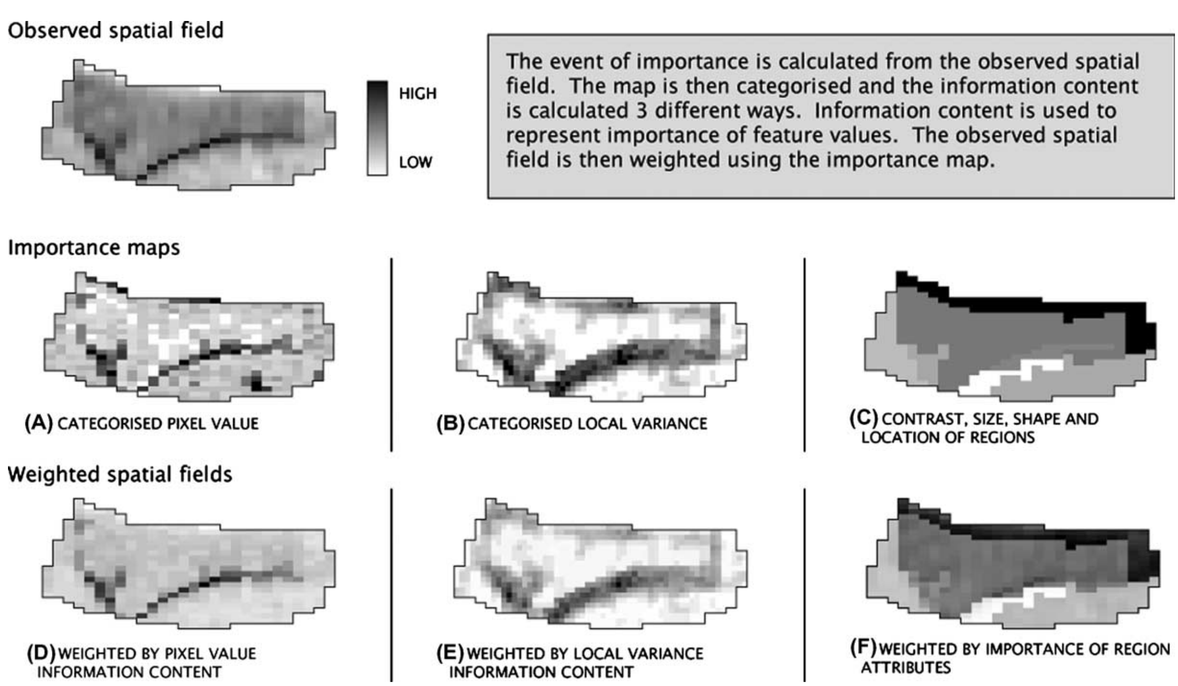

Fig. 2. Creating importance maps and weighted spatial fields using three distinct events: (A) pixel value (i.e. the observation); (B) local variance (within a 3×3 pixel neighbourhood); and (C) region attributes.

- the location of the region in the overall image (focus is more on regions in the centre 25% of the image); and
- the shape of the region (regions that are long and thin are visual attractors).

For each of these characteristics, a measure between zero and one was produced. These measures were then squared (to enhance highly important characteristics) and summed to produce a final importance value, which was then scaled so the maximum importance was one. The importance values were calculated for each region, producing an importance map. Fig. 2C presents an importance map made for the segmented spatial field shown previously (Fig. 1). The importance map produced for these regions gives minimal weighting to the valley in the spatial field, while it weights the region along the top more heavily (due to the long, thin shape). Osberger and Maeder [41] produced these importance criteria as a model of perceptual importance of regions. By combining this with the weighting approach of [60], the importance map can be used for conducting a MSE calculation to produce a comparison measure of perceptual importance. Fig. 2F weights the regions very differently to Fig. 2D and Fig. 2E as it is based on a model of visual perceptual importance rather than information content. For characterising hydrological importance (as opposed to perceptual importance), other measures such as slope, aspect or elevation derived from a digital elevation model (DEM) may be more useful. While this idea is not taken further in this paper, ongoing research is addressing the use of surrogate hydrological data to help define importance.

5.3. Fuzzy map comparison

When comparing spatial fields containing categorical data, some tolerance for locational or categorical errors is desirable. Locational tolerance allows slight shifts of pixel locations to occur without denoting them as a total disagreement. Similarly, categorical tolerance allows similar categories to be related and any mismatches between those categories to be judged more similar than total mismatches. When continuous maps have been categorised using histogram classes (thus producing ordered categories), tolerance for values in the slightly wrong category is particularly beneficial (a common occurrence in hydrological spatial fields). Research into fuzzy comparison methods, incorporating these ideas, have been suggested and developed by Power et al. [48] and more recently by Hagen [26]. Hagen [26] investigated two sources of fuzziness for categorical maps—fuzziness of location and fuzziness of category. This addresses both the spatial and thematic aspects of a comparison, facilitating fuzzy boundaries to be included in both cases. This approach implements fuzzy set theory, by representing each pixel in the spatial field as a fuzzy vector (containing the membership values for all possible categories). From the fuzzy representations of two spatial fields, the fuzzy vectors for coincident pixels are used to calculate the similarity metric (which has a value between zero and one, indicating the level of agreement between the fuzzy vectors). Calculating this metric in a pixel-by-pixel manner produces a fuzzy comparison map. Hagen [26] reduces the fuzzy comparison map down to two different summary statistics, which provide the quantitative measure of the comparison. The following explains how Hagen [26] represents fuzziness of location and category for each pixel in a spatial field (i.e. how the fuzzy vector is produced). It then looks at how the vectors are compared and introduces the summary statistics.

For representing fuzziness of category and location, each pixel is represented by a vector. The fuzziness vector for a pixel represents the membership that pixel has with all other categories in the map. Membership refers to how one category relates to the other categories. For example, values categorised as 'medium' could have some degree of membership with categories labelled 'low' or 'high'. These memberships are assigned subjectively, based on an understanding of the relationship between categories. Membership values are specified between zero and one, with one denoting complete membership (i.e. a pixel categorised as 'medium' has membership value of one for the 'medium' category, while having partial or no membership of other categories). Incorporating a measure of locational fuzziness is then achieved using a distance decay function to define the penalty that is given to matching categories located some distance from the pixel of interest (the central pixel in the neighbourhood). This is also chosen subjectively, depending on the level of tolerance allowed or expected. The result of adding fuzziness is that each pixel that previously had a single category value (e.g. 'medium'), now has a vector stating the likely membership of all the categories (e.g. $[0.2, 1.0, 0.3]$ for high, medium and low respectively) accounting for categorical and locational fuzziness. Once the spatial fields are represented using fuzzy vectors (as opposed to individual pixel values), two fields can be compared. On a pixel-by-pixel basis, the fuzzy vectors for common locations (which hold information about the neighbourhood of the pixel) are compared. The best match between the vectors represents the similarity of the pixels and is put into the fuzzy output map. The complete formulation of the fuzzy vectors and the similarity measure between two vectors is explained fully in the paper by Hagen [26].

The result of this local fuzzy comparison is an intermediate spatial field showing the degree of similarity between the categories assigned to each pixel, with one representing complete agreement and zero representing no agreement. The examples in Fig. 3 show fuzzy

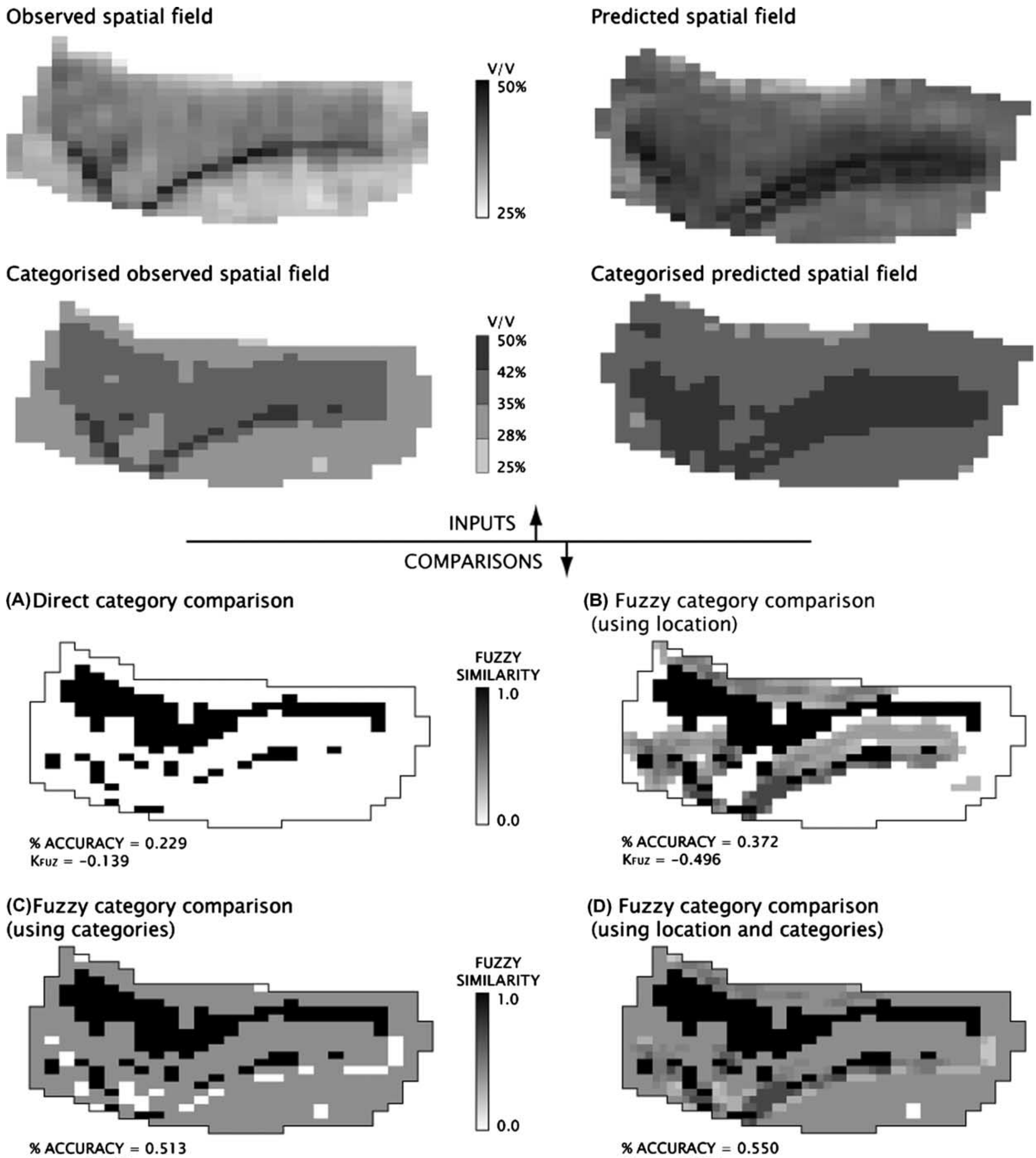

Fig. 3. Fuzzy comparison of observed and predicted spatial fields using four approaches: (A) category comparison; (B) fuzzy comparison with locational tolerance; (C) fuzzy comparison with categorical tolerance; (D) fuzzy comparison with both locational and categorical tolerance.

comparison maps that compare an observed and predicted spatial field of soil moisture. These fields, which originally had continuous values, have been categorised into four categories for comparison using this method. The four examples given in Fig. 3 show comparisons that represent (A) no fuzziness, (B) fuzziness of location only, (C) fuzziness of categories (in which each category has slight membership of the categories either side) and (D) fuzziness of both location and categories. From these fuzzy comparison maps, a summary statistic (in this case the average similarity value for all pixels in the map) is used to express the level of similarity between all the pixels and provide an overall quantitative measure of similarity. From the example given, we can see that as we build in more tolerance (or fuzziness), the similarity measure increases from 0.229 up to 0.550 (i.e. the spatial fields are being judged to be more similar). Using this approach to make many comparisons allows ranking of the most similar results. Hagen [26] has developed a measure that is similar to the kappa coefficient of agreement that was described earlier. This measure, called K_{FUZ} in Fig. 3, represents the observed similarity against an expected similarity, thus showing how much better the match is than with a randomly generated map (of categories with the same histograms). This measure is greater than zero where the match is better than expected, and less than zero when below what is expected. In Fig. 3, all K_{FUZ} values are negative, suggesting that the match is worse than would be expected when compared to a randomly generated map with the same categorical distribution. Hagen [26] recognises that evaluating the expected similarity is difficult with a fuzzy representation of categories and location. Several problems with the K_{FUZ} statistic are identified and improvements must be made for its use as a measure of comparison.

5.4. Multiscale comparisons

The implementation of methods for comparing observed and predicted spatial fields is usually undertaken at the finest resolution available (i.e. the resolution of the spatial field). However, as was identified earlier, humans usually observe global and local aspects of spatial fields when comparing them visually. Conducting comparisons at multiple scales of observation will allow a measure of similarity that varies with scale to be computed. In hydrological modelling, scaling of measurements to different time and space scales is the subject of ongoing research [7,68]. Approaches to changing the spatial scale of a spatial field vary, with certain methods more suitable for certain hydrological attributes. In image processing the multiple scales are just different resolution representations on the original image. Hay et al. [30] explains an approach to object detection in images where there is no prior knowledge of the scale of the objects. This approach, called 'linear scale-space', uses a set of smoothed versions of the original image to help identify image-objects, where the smoothing is simulating a change of scale (support) of the original image. By analysing the set of multiscale images and how they change with scale, the image-objects that are persistent across many scales can be extracted. In hydrology, some recent work by Gallant and Dowling [19] has used multiple scale versions of an elevation model to detect valley bottoms at multiple scales, while in flood hydrology, Horritt and Bates [31] have investigated the effect of scale on calibrating flood flow models. Despite these examples, there is generally limited use of multiscale analysis, and no use of multiscale comparison, in hydrology.

Successful analysis of images in other application areas using multiscale representations suggests that this should also be useful for spatial fields from hydrological models. Using scaling rules (e.g. aggregation into average values), resampling or image processing like convolution, a set of coarser resolution spatial fields can be created. Convolution involves passing a kernel (or 'moving window') over an image to create a new image, where each pixel is a function of the original pixels and the kernel. After producing coarser scale representations of the observed and predicted spatial fields, the comparison methods discussed earlier can be applied for each scale. With some methods, like mean squared error or direct correlation, this is simply a matter of computing the similarity measure. With others, it may involve computing importance weights or segmentations for each spatial field. The result of such a multiscale approach is a measure of similarity that changes with scale. Fig. 4 shows a multiscale comparison of the continuous value spatial fields used in Fig. 3. The fields have been resampled to coarser cell sizes using a bilinear interpolation to assign the resampled cell values. At each resolution, the standard MSE and the IMSE were calculated. This basic example shows that the similarity measure changes with the scale of observation. In this case, there is greater agreement between the spatial fields at the coarser cell sizes of 20 m and 60 m (using two different similarity measures), which may relate to some characteristic of the attribute being modelled or observed (or may be an artefact of the resampling process). An alternative to using smoothed versions of the original field was presented by Costanza [14]. In this case, the size of the areas being compared between two spatial fields was increased until the area covered the entire field. At each size, a summary value of the area being compared was used to produce the similarity measure. As with Fig. 4, the similarity measure also changes over scale (with the largest scale being a global comparison), allowing an overall measure of similarity to be derived. Multiscale comparisons go some way to emulating the global and local comparisons done visually by humans.

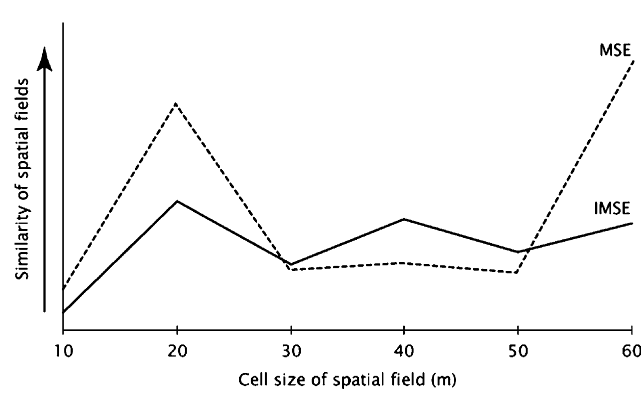

Fig. 4. Multiscale comparisons of the spatial field in Fig. 3 using two different measures: mean squared error (MSE); and information mean squared error (IMSE) using pixel values as the event.

6. Discussion

The comparison methods used for spatial fields from hydrological models have been very limited to date. While many papers present simulated or predicted spatial fields, and more recently some have shown observed spatial fields, only in the last few years have the weaknesses in comparison methods been stated [22,35]. Until a suite of quantitative comparison methods can be developed, calibration, optimisation and uncertainty analysis of distributed models will remain limited to point time series or weak spatial comparisons that provide only limited tests of spatial field similarity. The methods that have been presented in this paper address some issues of comparison that are necessary to improve model testing and interpretation. The inspiration for these methods draws from approaches used by human visual comparison, applications using image processing and the broad area of computer vision. While the background and current uses of these approaches suggests their flexibility and suitability for hydrological spatial fields, ongoing research is required to fully assess their application in this area. The following discussion speculates on how the promising and existing methods might work together to overcome the limitations of current comparisons (i.e. repeatability, quantification, recognition of features, tolerance for minor shifts, multiscale comparisons). It identifies some of the benefits of multiple approaches to comparison and suggests what these methods may infer about the spatial field similarity.

Fig. 5 is a conceptual diagram of comparison methods to indicate how these can work together to achieve the overall task of comparing spatial fields. Processing begins with a range of methods for preparing the raw observations and predictions into a suitable form for making comparisons (i.e. having equivalent resolutions, extents, type of values). This involves methods such as conditioning, segmentation, resampling (for multiscale comparisons) and the calculation of important features (such as homogeneous regions). The pre-processing produces a set of comparable spatial fields, which can also be generated at multiple scales. These processed inputs can be used with the comparison methods. The standard feature-by-feature, weighted feature-by-feature and fuzzy comparisons can be used to produce intermediate spatial fields. These are the measures often used in manual analyses, but for automated comparison or processing large number of spatial fields, the reduction of these measures down to quantitative comparison measures is necessary. Where a multiscale representation of the input spatial fields has been created, the comparison measures can be computed for the multiple scales. Fig. 5 also lists the type of fields suited to each method. In many cases, the methods work with both continuous and categorical fields, although some methods work with only one type. Continuous value spatial fields can readily be placed into categories, making the fuzzy comparison method (which requires categorical inputs) versatile to both data types.

The comparison measures in Fig. 5 all produce a quantitative measure of similarity. These measures sometimes have a direct meaning, such as MSE, which represents the mean of the squared differences between pixel values (in the same units as the pixel values). In other comparisons, such as IMSE, the original values are weighted by a unitless measure, thus making the exact meaning of the measure somewhat unclear and requiring further interpretation (the measure is relative). For example, the correlation coefficient is a unitless measure computed from the available data points and thus biased when there is little spread in the data. Similarly, the weightings assigned in an IMSE calculation are controlled by the homogeneity of the spatial field. As such, care must be taken when using these methods for inter-comparison. For example, comparing two sets of observed and predicted fields of soil moisture (one set from spring and one from winter) could produce comparison measures with vastly different ranges using IMSE. The relatively homogeneous winter spatial fields (nearly all wet) would receive lesser weightings than the heterogeneous spring spatial fields (showing dry hillslopes and wet gullies), making the similarity between the observed and predicted fields in each set difficult to compare.

The methods that have emerged from image and computer vision research are not concerned with intercomparison. They are designed to produce a comparison measure that is powerful for comparing a single template (i.e. an observation) with a database of possible matches (i.e. a set of predictions). This is similar to some uncertainty frameworks discussed in hydrology [3], in which the simulated spatial fields from many models and parameter sets (e.g. an ensemble of outputs at a point in time) are compared to a single observation, with poor matching predictions being rejected. There is a key role for spatial observations in the rejection of nonbehavioural models and parameter sets, thus limiting the possible model realisations and acceptable model parameter sets. As Beven [3] points out, "the real challenge is to find creative ways of using observations to limit those possibilities." The methods presented here offer the potential to increase the utility of observations at this model assessment stage.

All of the pre-processing methods listed in Fig. 5 require some subjective judgement in the choice of parameter values. For example:

- Categorising a continuous spatial field requires the number of output categories to be chosen.
- Segmentation, while often being data driven, requires parameters to be set to control the form of the segmented regions—such as thresholds, the number of

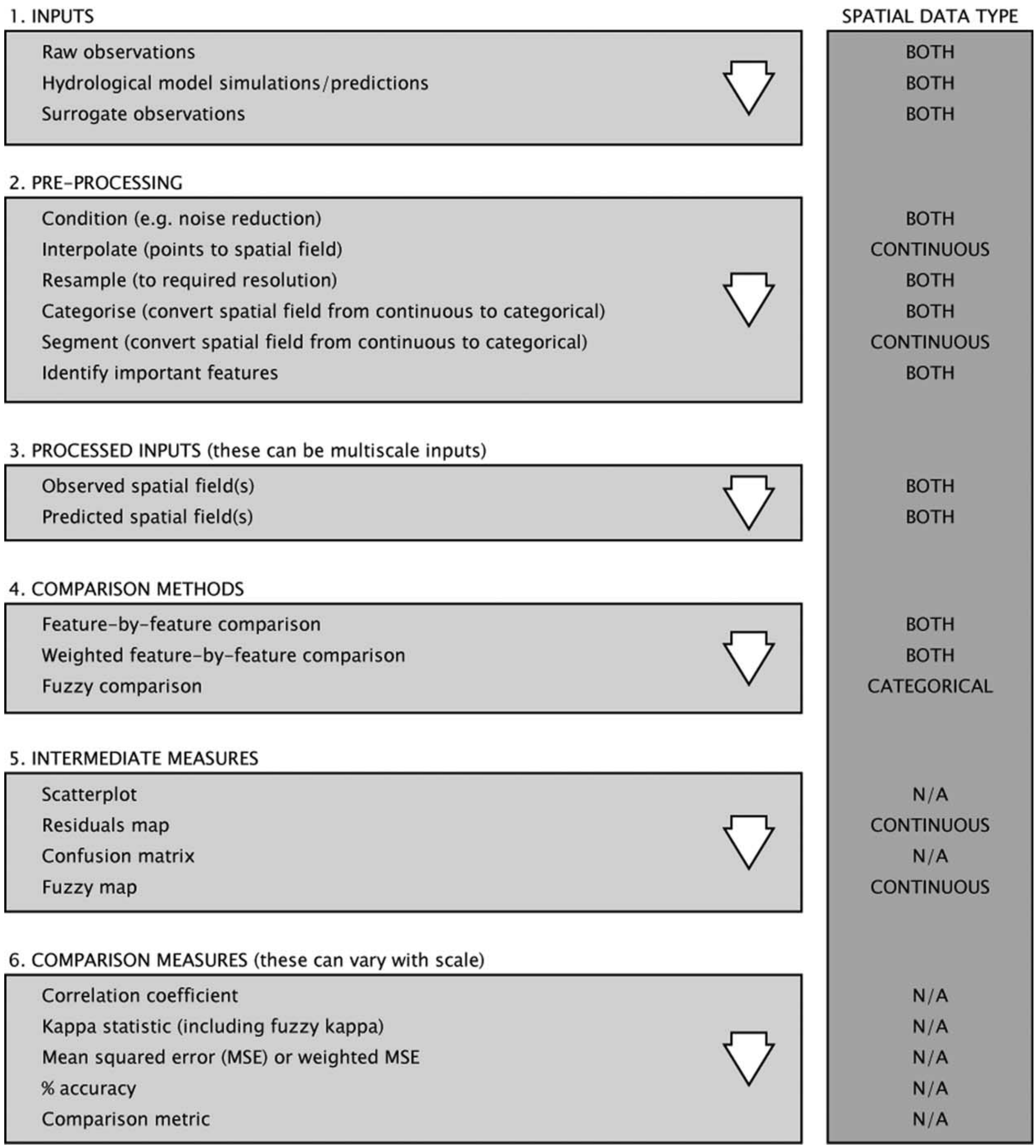

Fig. 5. Conceptual diagram of the methods introduced in this paper and their role in the comparison process. The type of spatial field suitable for each method is listed (i.e. continuous or categorical values).

clusters, region merging tolerances (relating to the size and shape of regions).

- Identification of important features requires definition of the 'event of interest' and how the importance of the event is defined (e.g. with Shannon's self-information measure).
- Importance maps (which are used to weight a standard comparison) must be chosen such that they are strong tests of model performance and do not 'undervalue' other features of the spatial fields that could also be important.
- Fuzzy comparisons (which can tolerate locational and categorical errors between spatial fields) require subjective definition of the memberships for each category and the amount of locational tolerance.

In each case, the choices made are an explicit statement of what the user believes is important in the spatial fields and the system being studied. These choices need to exploit the user's knowledge about the characteristics of the model output and observations. For example, interpolation involves a change in support that may remove some small-scale variability critical to testing model performance. Resampling involves a change in spacing. These interpolation or resampling steps should be aimed at matching the scale triplet of the observations and simulations so comparisons are not biased [6]. Similarly, the 'event of interest' (or important features) can end up being the primary part of a spatial field that most affects the final comparison metrics, as other important features may be under weighted and therefore ignored in the metric. For example, an 'event of interest' in a snow model comparison might be the presence of snow on particular topographic aspects, thus testing the components of the model related to wind drift or radiation exposure, but testing little about elevation or lapse-rate components. Comprehensive comparison approaches will require application of multiple methods to capture multiple features of the spatial fields. These choices are no different in concept to the choice of the objective functions commonly used with time series comparisons (such as catchment runoff)—some emphasise high flows, some emphasise volume matching, some emphasise timing shifts (e.g. [38,64]). However, in the case of spatial fields, there is much less experience to draw on as to the best parameters to use.

As the availability of observed spatial fields increases, the existence of observed spatial fields that vary with time (dynamic spatial fields) will arise. Spatial models already produce fields that change at each time step and there will become a need for comparison methods when we have dynamic spatial observations. The methods applied here can be readily used with multiple time steps of comparable spatial fields. They would produce time-varying measures of similarity (just like the scale-varying measure in Fig. 4) and could be extended to incorporate fuzziness in time (by comparing against spatial fields from previous or subsequent time steps). Further work with data of this type would be possible in meteorology, where simulated rainfall fields and observed radar rainfall fields are available. At present we are not dealing with dynamic spatial fields, although the methods described here could potentially be used.

At present, the state-of-the-art for assessing spatial predictions from models is limited. This paper has focussed on describing and demonstrating some methods used in other disciplines that show promise for application to comparing spatial fields of hydrological phenomena. The methods discussed in this paper are largely steps for pre-processing the spatial fields. Through the pre-processing, different aspects of the spatial fields are recognised and are then used in the subsequent comparison step. Depending on the approach used, the actual comparison step could be a standard pixel-by-pixel process, or it could be an alternative procedure like fuzzy comparison. All of the approaches presented here provide alternatives that can address different aspects of comparison. They are applied to different types of features and different scales, with a view to providing richer quantitative measures of comparison. Future work will focus on application of these methods to a range of spatial fields from hydrology, to reveal their true value for distributed hydrological model calibration, assessment, development and uncertainty estimation.

Acknowledgment

We would like to acknowledge support from the Cooperative Research Centre for Catchment Hydrology and an Australian Postgraduate Award (APA) scholarship.

References

[1] Aronica G, Bates PD, Horritt MS. Assessing the uncertainty in distributed model predictions using observed binary pattern information within GLUE. Hydrol Process 2002;16:2001–16.
[2] Baatz M, Schäpe A. Multiresolution segmentation: an optimization approach for high quality multiscale image segmentation. In: Strobl J, Blaschke T, editors. Angewandte Geogr Informationsverabeitung. Heidelberg, Germany: Herbert Wichmann Verlag; 2000. p. 12–23.
[3] Beven K. Towards an alternative blueprint for a physically based digitally simulated hydrologic response modelling system. Hydrol Process 2002;16:189–206.
[4] Beven K, Binley A. The future of distributed models—model calibration and uncertainty prediction. Hydrol Process 1992;6:279–98.
[5] Beven KJ. Prophecy, reality and uncertainty in distributed hydrological modelling. Adv Water Resour 1993;16:41–51.
[6] Blöschl G, Grayson R. Spatial observations and interpolation. In: Grayson R, Blöschl G, editors. Spatial patterns in catchment hydrology: observations and modelling. Cambridge: Cambridge University Press; 2000. p. 17–50.
[7] Blöschl G, Sivapalan M. Scale issues in hydrological modeling—a review. Hydrol Process 1995;9:251–90.
[8] Bow S. Pattern recognition and image preprocessing. New York, USA: Marcel Dekker; 2002. 698 p.
[9] Boyle DP, Gupta HV, Sorooshian S. Toward improved calibration of hydrologic models: combining the strengths of manual and automatic methods. Water Resour Res 2000;36:3663–74.
[10] Chang YL, Li X. Adaptive image region-growing. IEEE Image Process 1994;3:868–72.
[11] Chen Y, Wang JZ. A region-based fuzzy feature matching approach to content-based image retrieval. IEEE Pattern Anal Mach Intell 2002;24:1252–67.
[12] Clarke KC. Getting started with geographic information systems. New Jersey: Prentice Hall; 1999. 338 p.
[13] Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Measur 1960;20:37–46.
[14] Costanza R. Model goodness of fit—a multiple resolution procedure. Ecol Modell 1989;47:199–215.
[15] Findlay JM, Walker R, Kentridge RW. Eye movement research: mechanisms, processes and applications. Amsterdam, The Netherlands: Elsevier; 1995. 564 p.

[16] Foody GM. Status of land cover classification accuracy assessment. Remote Sens Environ 2002;80:185–201.
[17] Foufoula-Georgiou E, Vuruputur V. Patterns and organisation in precipitation. In: Grayson R, Blöschl G, editors. Spatial patterns in catchment hydrology: observations and modelling. Cambridge: Cambridge University Press; 2000. p. 82–104.
[18] Franks SW, Gineste P, Beven KJ, Merot P. On constraining the predictions of a distributed moder: the incorporation of fuzzy estimates of saturated areas into the calibration process. Water Resour Res 1998;34:787–97.
[19] Gallant JC, Dowling TD. A multiresolution index of valley bottom flatness for mapping depositional areas. Water Resour Res 2003;39:1347.
[20] Grayson R, Blöschl G. Summary of pattern comparison and concluding remarks. In: Grayson R, Blöschl G, editors. Spatial patterns in catchment hydrology: observations and modelling. Cambridge: Cambridge University Press; 2000. p. 355–67.
[21] Grayson RB, Blöschl G, editors. Spatial patterns in catchment hydrology: observations and modelling. Cambridge University Press: Cambridge; 2000. 404 pp.
[22] Grayson RB, Blöschl G, Western AW, McMahon TA. Advances in the use of observed spatial patterns of catchment hydrological response. Adv Water Resour 2002;25:1313–34.
[23] Grayson RB, Moore ID, McMahon TA. Physically based hydrologic modeling. 2. Is the concept realistic?. Water Resour Res 1992;28:2659–66.
[24] Güntner A, Seibert J, Uhlenbrook S. Modeling spatial patterns of saturated areas: an evaluation of different terrain indices. Water Resour Res 2004;40.
[25] Gustafson EJ. Quantifying landscape spatial pattern: what is the state of the art. Ecosystems 1998;1:143–56.
[26] Hagen A. Fuzzy set approach to assessing similarity of categorical maps. Int J Geog Inf Sci 2003;17:235–49.
[27] Haralick RM, Shapiro LG. Image segmentation techniques. Comput Vision Graphics Image Process 1985;29:100–32.
[28] Haralick RM, Shapiro LG. Computer and robot vision. Reading, MA: Addison-Wesley; 1992. 1302 p.
[29] Hay GJ, Blaschke T, Marceau DJ, Bouchard A. A comparison of three image-object methods for the multiscale analysis of landscape structure. J Photog Remote Sens 2003;57:327–45.
[30] Hay GJ, Dube P, Bouchard A, Marceau DJ. A scale-space primer for exploring and quantifying complex landscapes. Ecol Modell 2002;153:27–49.
[31] Horritt MS, Mason DC, Luckman AJ. Flood boundary delineation from Synthetic Aperture Radar imagery using a statistical active contour model. Int J Remote Sens 2001;22:2489–507.
[32] Huttenlocher DP, Klanderman GA, Rucklidge WJ. Comparing images using the hausdorff distance. IEEE Pattern Anal Mach Intell 1993;15:850–63.
[33] Itti L, Koch C, Niebur E. A model of saliency-based visual attention for rapid scene analysis. IEEE Pattern Anal Mach Intell 1998;20:1254–9.
[34] Jain AK, Murty MN, Flynn PJ. Data clustering: a review. ACM Comput Surveys 1999;31:264–323.
[35] Jetten V, Govers G, Hessel R. Erosion models: quality of spatial predictions. Hydrol Process 2003;17:887–900.
[36] Klemeš V. Dilettantism in hydrology—transition or destiny. Water Resour Res 1986;22:S177–88.
[37] Kuczera G, Mroczkowski M. Assessment of hydrologic parameter uncertainty and the worth of multiresponse data. Water Resour Res 1998;34:1481–9.
[38] Legates DR, McCabe GJ. Evaluating the use of goodness-of-fit measures in hydrologic and hydroclimatic model validation. Water Resour Res 1999;35:233–42.
[39] Mastin GA. Adaptive filters for digital image noise smoothing: an evaluation. Comput Vision Graphics Image Process (CVGIP) 1985;31:103–21.
[40] Merz B, Plate EJ. An analysis of the effects of spatial variability of soil and soil moisture on runoff. Water Resour Res 1997;33:2909–22.
[41] Osberger W, Maeder AJ. Automatic identification of perceptually important regions in an image. In: Proc 14th International Conference on Pattern Recognition, Brisbane, Australia, 1998.
[42] Pal NR, Pal SK. A review on image segmentation techniques. Pattern Recog 1993;26:1277–94.
[43] Pauwels EJ, Frederix G. Finding salient regions in images—nonparametric clustering for image segmentation and grouping. Comput Vision Image Underst 1999;75:73–85.
[44] Pavlidis T. 36 years on the pattern recognition front. Pattern Recog Lett 2003;24:1–7.
[45] Peucker TK, Douglas DH. Detection of surface specific points by local parallel processing of discrete terrain elevation data. Comput Graphics Image Process 1975;4:375–87.
[46] Pomeroy JW, Li L. Prairie and arctic areal snow cover mass balance using a blowing snow model. J Geophys Res 2000;105:26619–34.
[47] Pontius RG. Quantification error versus location error in comparison of categorical maps. Photogramm Eng Remote Sens 2000;66:1011–6.
[48] Power C, Simms A, White R. Hierarchical fuzzy pattern matching for the regional comparison of land use maps. Int J Geog Inf Sci 2001;15:77–100.
[49] Pujol A, Villanueva JJ, Alba JL. A supervised modification of the Hausdorff distance for visual shape classification. Int J Pattern Recog Artif Intell 2002;16:349–59.
[50] Purdy JC, Harris D, Austin GL, Seed AW, Gray W. A case study of orographic rainfall processes incorporating multiscaling characterization techniques. J Geophys Res 2001;106:7837–46.
[51] Refsgaard JC. Parameterisation, calibration and validation of distributed hydrological models. J Hydrol 1997;198:69–97.
[52] Scheibe TD. Characterization of the spatial structuring of natural porous media and its impacts on subsurface flow and transport. PhD thesis, Stanford University, 1993.
[53] Schmid C, Mohr R, Bauckhage C. Evaluation of interest point detectors. Int J Comput Vision 2000;37:151–72.
[54] Schmugge TJ, Kustas WP, Ritchie JC, Jackson TJ, Rango A. Remote sensing in hydrology. Adv Water Resour 2002;25:1367–85.
[55] Smeulders AWM, Worring M, Santini S, Gupta A, Jain R. Content-based image retrieval at the end of the early years. IEEE Pattern Anal Mach Intell 2000;22:1349–80.
[56] Smits PC, Dellepiane SG, Schowengerdt RA. Quality assessment of image classification algorithms for land-cover mapping: a review and a proposal for a cost-based approach. Int J Geog Inf Sci 1999;20:1461–86.
[57] Takken I, Beuselinck L, Nachtergaele J, Govers G, Poesen J, Degraer G. Spatial evaluation of a physically-based distributed erosion model (LISEM). CATENA 1999;37:431–47.
[58] Tarboton D, Blöschl G, Cooley K, Kirnbauer R, Luce C. Spatial snow cover processes at kuhtai and reynolds creek. In: Grayson R, Blöschl G, editors. Spatial patterns in catchment hydrology: observations and modelling. Cambridge: Cambridge University Press; 2000. p. 158–86.
[59] Thiemann M, Trosset M, Gupta H, Sorooshian S. Bayesian recursive parameter estimation for hydrologic models. Water Resour Res 2001;37:2521–35.
[60] Tompa D, Morton J, Jernigan E. Perceptually based image comparison. In: Proc Proceedings of the IEEE International Conference on Image Processing, Vancouver, BC, Canada, 2000. p. 489–92.
[61] Turner MG. Landscape ecology: the effect of pattern on process. Annu Rev Ecol Syst 1989;20:171–97.
[62] Turner MRJ, Mynett AE, van den Boogaard HFP. Feature extraction from satellite images for validating suspended sediment

models. In: Proc Fifth International Conference on Hydroinformatics, Cardiff, Wales, UK, 2002.
[63] Udupa JK, Saha PK. Fuzzy connectedness and image segmentation. Proc IEEE 2003;91:1649–69.
[64] Vázquez RF, Feyen L, Feyen J, Refsgaard JC. Effect of grid size on effective parameters and model performance of the MIKE-SHE code. Hydrol Process 2002;16:355–72.
[65] Veltkamp RC. Shape matching: Similarity measures and algorithms. Utrecht University, UU-CS-2001-03, Utrecht, The Netherlands, 2001. 17 pp.
[66] Veltkamp RC, Hagedoorn M. State-of-the-art in shape matching. Utrecht University, UU-CS-1999-27, Utrecht, The Netherlands, 1999. 26 pp.
[67] Vertessy R, Elsenbeer H, Bessard Y, Lack A. Storm runoff generation at la cuenca. In: Grayson R, Blöschl G, editors. Spatial patterns in catchment hydrology: observations and modelling. Cambridge: Cambridge University Press; 2000. p. 247–71.
[68] Western AW, Grayson RB, Blöschl G. Scaling of soil moisture: a hydrologic perspective. Ann Rev Earth Planet Sci 2002;30:149–80.
[69] Western AW, Grayson RB, Green TR. The Tarrawarra project: high resolution spatial measurement, modelling and analysis of soil moisture and hydrological response. Hydrol Process 1999;13:633–52.
[70] Wilson JP, Gallant JC. Terrain analysis: principles and applications. New York: Wiley; 2000. 512 p.
[71] Woodcock C, Harward VJ. Nested-hierarchical scene models and image segmentation. Int J Remote Sens 1992;13:3167–87.
[72] You J, Li WX, Zhang D. Hierarchical palmprint identification via multiple feature extraction. Pattern Recog 2002;35:847–59.