Using Tolerances When Assessing Models Using Spatial Fields

S.R. Wealands [1, 2], **R.B. Grayson** [1, 2] **and J.P. Walker** [1]

[1] **Department of Civil and Environmental Engineering, The University of Melbourne, Victoria**
[2] **Cooperative Research Centre for Catchment Hydrology, The University of Melbourne, Victoria**
Email: srweal@civenv.unimelb.edu.au

***Keywords**: Spatial fields; Similarity; Error; Tolerance; Comparison*

EXTENDED ABSTRACT

Model assessment involves expressing the performance of a model for a given purpose (e.g. a particular range of conditions or locations). The assessment of distributed hydrological models (i.e. spatially explicit models for hydrology) is usually done with limited point samples, but this is inadequate for assessing spatial performance. Spatial fields provide a more complete picture of 'reality' against which spatial models should be assessed. This type of data is increasingly available in hydrology, through improved remote sensing techniques and other methods of spatial sampling. There are limited examples of spatial fields being used for model assessment. In cases where they have been, the fields provided sensitive checks on the modelling and were generally also interrogated to reveal issues with model structure. However, the majority of these analyses were done visually. This is because the standard comparison methods (i.e. the objective functions) used do not currently utilise the rich information on spatial organisation that spatial fields contain.

Visual comparison is a valuable method for comparing fields, as it allows background knowledge (e.g. experience, understanding of purpose) to be incorporated into the process. Unfortunately, visual comparison is neither rigorous, repeatable, unbiased nor quantitative. When measures of error or similarity are wanted, visual comparison cannot be used. It can, however, be used to learn what aspects of comparison should be pursued by any new quantitative methods. The general pattern analysis literature has been reviewed previously to identify comparison methods that can potentially emulate these aspects (Wealands *et al.* 2005). The methods that emulate the ability *to tolerate differences in value and location between elements* are pursued in this paper. These methods are for use after standard measures (e.g. bias, RMSE) have been applied. They give an overall measure of error or similarity under specified tolerances. They also produce graphical measures that can be inspected for more localised analyses.

Tolerant comparisons require tolerances to be specified for differences in value (ΔV) and location (ΔL). The tolerances can be specified as crisp or fuzzy. Crisp tolerances control which values and locations between elements in the fields are judged as being equal. In contrast, fuzzy tolerances define a scale (from one to zero) to describe how similar elements are in value and location. Figure 1 shows an example of how the crisp and fuzzy tolerances can be defined. Using these tolerances, each element in a modelled field is compared to the observed field. All elements that are within a distance of ΔL of the modelled element are treated as being similar (to some degree). The tolerances (ΔV and ΔL) are combined to determine the optimum local measure for each element and this field of measures is summarised to produce a final measure.

One observed field is compared against five 'model' fields, which are created by introducing distortions to the observed. The results illustrate how the measures respond to differences. They show that fields with differences within the tolerances produce equivalent results. By contrasting measures with and without tolerances, the presence/absence of shifts, noise or scale differences can be inferred. Such inferences apply to the whole field, although more localised analysis can provide information on local effects.

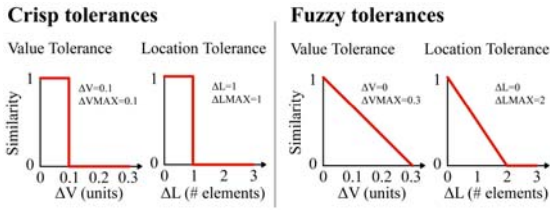

Figure 1 Crisp or fuzzy tolerances are used to translate differences between elements into similarity values (or error values).

1. INTRODUCTION

To advance the use of modelling for decision making, the assessment of model performance must be undertaken and conveyed. Model assessment involves expressing the expected performance of a model for a given purpose (e.g. a range of conditions or locations). In hydrology, distributed hydrological models have developed over the last 20 years and can now provide spatially-explicit simulations of various aspects of hydrology. However, they are usually assessed using only limited point samples, which are inadequate for assessing the performance of spatial simulations. Grayson *et al.* (2002) refer to examples where spatial fields have been used for assessment, but find that in many cases the comparison methods (or objective functions) used with spatial fields do not utilise the rich information they contain. Some efforts are made to address this, although all the methods still require visual interpretation.

Comparison methods that work with spatial fields provide alternative tests of distributed model performance. These are important throughout the modelling process, but particularly during calibration and testing. Calibration involves adjusting model parameters and simulating a period for which observations exist. The similarity (or error) between the simulated and observed data is determined using comparison methods. The parameters are then adjusted until the similarity is maximised (or the errors are minimised). By using multiple objective functions (Gupta *et al.* 1998), the calibration can ensure that multiple desirable characteristics of the observed data are represented in the simulation. If using an uncertainty framework such as GLUE (Beven 1993), the comparison methods are used to assign likelihoods to all possible models and parameter sets. The likelihoods are then used to reject non-behavioural models and subsequently provide an estimate of the uncertainty expected in predictions. Model testing also requires improved comparison methods for quantifying the similarity (or error) between the calibrated model and independent observations.

The primary limitations with the use of spatial fields for assessment are the availability of observed fields and quantitative comparison methods. Remote sensing technologies are addressing the needs for hydrology to some extent, although there is continued work on interpreting remote sensing signals for hydrological studies. In studies where an effort has been made to obtain observed spatial fields, the information garnered from these observations has proven useful for model assessment (e.g. Western *et al.* 1999, Güntner *et al.* 2004, Jetten *et al.* 2003).

The studies that utilise spatial fields have recognised the lack of standard methods for comparing the data. There is a need for quantitative comparison methods that measure different aspects of similarity and error. At present, hydrologists depend largely on visual comparison to assess spatial model outputs (Grayson *et al.* 2002). Visual comparison is neither rigorous, repeatable, unbiased nor quantitative. But it can be used to learn what aspects of comparison should be pursued by any new quantitative methods. Wealands *et al.* (2005) reviewed the pattern analysis literature to reveal comparison methods that emulate aspects of visual comparison. From these methods, those that showed promise for assessing hydrological models were – fuzzy map comparison, weighted analysis, image segmentation and multiscale comparison.

This paper builds on the ideas of fuzzy map comparison (Hagen 2003), in which tolerances can be implemented for differences in both value and location between two or more spatial fields. The information that can be garnered from these methods is presented by comparing fields with known distortions. A discussion about the value of tolerant measures for model assessment is provided. The discussion explains what these measures can reveal when used automatically and also interactively.

2. DEFINITONS

A number of definitions are provided here to ensure readers understand the meaning of terms used when discussing comparison methods.

A *spatial field* is a set of associated elements, where each *element* represents a certain attribute at a given location. A spatial field must have a physical or logical relationship between elements known as *topology*, otherwise it is simply a set of non-related elements. Each spatial field has one attribute at each location and represents a single time. Comparison methods use specific algorithms to produce comparison measures of either error or similarity. The algorithms use numerical relationships (e.g. differences, ratios) between the characteristics of elements (e.g. value, location, shape) in each spatial field to derive their quantity.

Global measures summarise the characteristics of all elements in each spatial field into one summary value. The numerical relationship between the summary values is used to derive the measure. *Local measures* work with each pair of elements

that have a specific spatial relationship (e.g. spatially coincident); one element of the pair is from the modelled spatial field and the other is from the observed spatial field. The numerical relationship between a characteristic of the elements is used to derive a *local measure* between each pair. At this stage a *graphical measure* (e.g. a residual field or scatterplot) can be produced for analysis. The local measures (from each pair of elements) are further processed into the local measure that represents the error or similarity between the spatial fields.

Error measures use the numerical difference (or residual) between spatial field elements in their derivation. They are expressed in the same units as the spatial fields, requiring their magnitude to be interpreted relative to the fields being compared. *Similarity measures* use various numerical relationships between spatial field elements in their derivation. They often use ratios to provide a measure that is relative to a known benchmark (e.g. the observed variance). Similarity measures are expressed on a different scale to the fields being compared. They usually have a maximum value of one to denote perfect similarity under the conditions imposed by the comparison method. A value of zero is used as either the minimum (i.e. no similarity) or to denote the similarity of the benchmark (e.g. coefficient of efficiency).

3. METHOD

The comparison methods presented here build on recent developments in the land-use modelling and geographical information literature (Hagen 2003, Power *et al.* 2001). They have been adapted to work with the interval/ratio type data that is most commonly produced by hydrological models. These methods are relatively easy to implement and provide useful alternatives to the current methods used in hydrology for comparing spatial fields. The measures produced can be interpreted in a familiar way for hydrologists.

3.1. Specifying Tolerances

When comparing fields, tolerances can be specified for differences in location and/or value. These tolerances are either crisp or fuzzy (Power *et al.* 2001). *Crisp tolerances* are used to define the exact differences that are accepted within a comparison method. A tolerance for the allowable difference in value (ΔV) and the allowable difference in location (ΔL) are defined by the user (e.g. Figure 2B). When both the tolerances are set to zero, the standard comparisons like RMSE are the result (e.g. Figure 2A).

Fuzzy tolerances are used to translate differences into similarity values (on a scale from zero to one). If there is no difference, the similarity is one. If the difference is greater than the tolerance, the similarity is zero. Fuzzy tolerances assign a value between zero and one to all possible differences between elements. For differences in value, the value similarity (SV) can be assigned subjectively or with a decay function. For simplicity, a linear decay function is used here, in which the allowable difference in value (ΔV) and the maximum difference in value (ΔVMAX) are required. For differences in location, the location similarity (SL) can also be assigned subjectively or with a decay function. Hagen (2003) suggests the use of exponential decay, requiring the radius and halving distance to be specified. The definition varies with the application, but linear decay is used here for simplicity. The allowable difference in location (ΔL) and the maximum difference in location (ΔLMAX) must be specified (e.g. Figure 2D). Where the allowable difference and maximum difference are equal (ΔL=ΔLMAX), a crisp tolerance is effectively applied to calculate a similarity measure (e.g. Figure 2C).

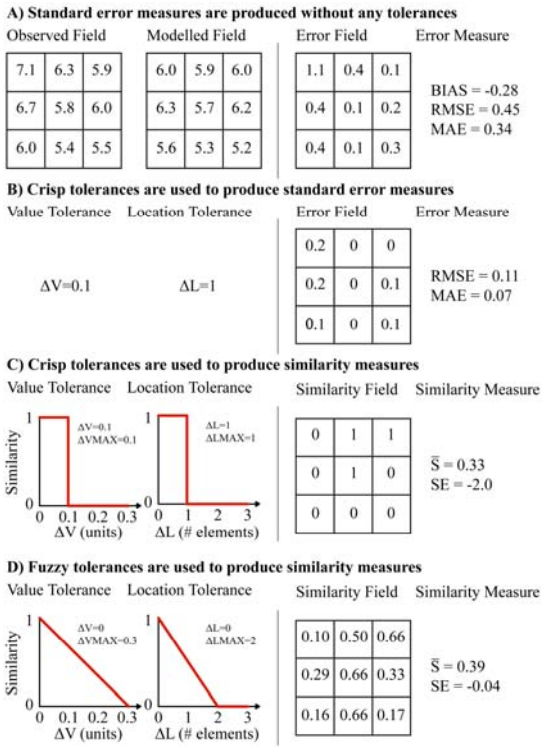

Figure 2 Following the comparison methods described in 3.2, the modelled field is compared using a variety of crisp and fuzzy tolerances. These tolerances are applied during the comparison of each modelled element to the observed field, resulting in both graphical and summary measures (of error or similarity).

Tolerances should be chosen based on known issues in the data or processing. For example, Güntner *et al.* (2004) used a location tolerance of 5 elements, due to support differences between DEM-derived topographic indices (50m) and the observations (10m). Both crisp and fuzzy tolerances were used to understand similarity between binary fields of saturated area. Hagen (2003) used subjective weightings to define the similarity between related category labels (i.e. nominal data) in a land use modelling study (e.g. high- and low-density residential). For interval/ratio data, the tolerances are set using knowledge about the error in the observed field and the differences that are considered to be similar 'to some degree'.

3.2. The Comparison Process

This comparison method treats each element in the modelled field once (i.e. it compares each modelled element against the 'reality' it is supposed to represent). A measure of error (E) or similarity (S) is made between this element and every element from the observed field that is within the shift tolerance (limited by either ΔL or $\Delta LMAX$). The best match (either minimum error or maximum similarity) found for the modelled element amongst all comparisons with the observed elements is retained. When completed for every element in the field, a field of error values or similarity values is created. This is a graphical measure that can be visually interpreted.

For crisp tolerances (e.g. Figure 2B), the error measure between a pair of elements (one modelled element and one observed element) uses the absolute difference between element values. If this is less than ΔV, then E equals zero. If greater, then ΔV is subtracted from the value difference. If producing a similarity measure using crisp tolerances, S equals one when the value difference is less than ΔV, otherwise S equals zero (e.g. Figure 2C).

For fuzzy tolerances, the similarity measure between a pair of elements uses the absolute difference between element values and the distance between elements. The difference between values and locations are converted into SV and SL accordingly. SV and SL are multiplied to give a value to S for each element (e.g. Figure 2D). Error measures cannot be produced using fuzzy tolerances.

This field of errors or similarity is the basis for producing any subsequent measures. It is also a graphical measure that can be visually interpreted to understand the spatial distribution of similarity or error. This is a very useful exercise that should be undertaken when limited comparisons are needed. For larger comparison situations, the numerical summaries are more manageable, although the graphical measures provide a valuable check on the numerical findings. They can be used to direct further comparisons in parts of the field with large differences. This can then lead to understanding why the differences are occurring.

3.3. Error Measures

The error field obtained from using crisp tolerances can be analysed like any error field. Common error measures like root mean squared error (RMSE) can be calculated from it. These measures, if produced for increasing levels of tolerance, will decrease until they reach zero (i.e. all the error is tolerated). As with any absolute measure, the error value is only useful if it can be assessed against another model or background knowledge. Error measures accumulate the error for all elements, thus making them susceptible to very large errors biasing the measurement. This is the where error measures act very differently to similarity measures. If large errors are not biasing the assessment, then the error measures should have high inverse correlation with the similarity measures.

3.4. Similarity Measures

The similarity field obtained by using either crisp or fuzzy tolerances can be summarised into an overall similarity value. If crisp tolerances are used, each similarity value is either zero or one. If fuzzy tolerances are used, each similarity value is in the range zero to one. Taking the average of the similarity values ($\overline{S}$) gives measure of the proportion of the field judged as being similar. As the tolerances increase, this value approaches one.

A similarity measure that is relative to a known reference is described here. The similarity efficiency (SE) measure allows similarity results from one set of fields to be compared against another set (e.g. for a different attribute, location or time). When comparing spatial fields, the observed mean field (i.e. a field containing the observed mean in every element) is a suggested standard reference. This has the most general characteristic (i.e. the mean) of the observation correct, but lacks any variability in spatial arrangement. This is considered a suitable benchmark against which to judge spatial model performance, although the performance of any other spatial field could be used. SE uses the same idea as the coefficient of efficiency (Nash and

Sutcliffe 1970) that is widely used in hydrology, in which the model performance is scaled by the performance of the observed mean, producing a measure that is comparable across different data sets. This measure is given by

$$SE_{O,M} = \frac{\sum_{i}^{n} S(M_i, O) - \sum_{i}^{n} S(\overline{O}_i, O)}{n - \sum_{i}^{n} S(\overline{O}_i, O)}, \quad (1)$$

where $S(M_i,O)$ is the similarity measure between modelled element i and the observed field O, $\overline{O}$ is the mean observed field, and n is the total number of elements. SE is less than zero when the mean field is more similar than the modelled field. It is positive and approaches one as $\overline{S}$ increases relative to the similarity of the mean field. This measure is interpreted in the same way as the coefficient of efficiency. The $\overline{S}$ and SE measures are always perfectly correlated, but SE allows values to be meaningfully analysed against other comparisons (e.g. from a different time or location).

Similarity measures are often favoured over error measures as they reward the elements that meet the criteria for similarity (i.e. those within the tolerances). They do not allow occasional large differences to dominate the result, therefore being insensitive to extremes that are not modelled correctly. Hydrologists are often concerned with the extremes, so error measures should also be produced to ensure that important differences are not overlooked.

3.5. Data

The data used for demonstrating these measures is from part of the Mahurangi catchment in New Zealand (Wilson *et al.* 2005) Soil moisture observations were collected at regularly spaced points using TDR probes. These observations have been interpolated onto a gridded field. All elements have connectivity with their eight neighbouring elements. From this observed field, five 'model' fields have been produced to demonstrate the use of tolerances during comparisons. These 'models' are distortions of the observed field and are shown, along with their global statistics, in Figure 3.

4. RESULTS

One error measure (RMSE) and two similarity measures ($\overline{S}$, SE) have been computed by comparing the five fields shown in Figure 3. Additionally, the observed mean field (MN) has been compared to illustrate the benchmark performance of each comparison. Four different sets of tolerances have been applied. Value tolerances range from 0-10%, while location tolerances vary from 0-90m (the distance between element centroids). The results in Table 1 are used to show how the measures respond to the introduced differences between the fields.

Noise: When $\Delta V=2$, major improvement is seen in all the measures for NS (relative to the other fields). The RMSE under this tolerance does not reduce to zero, as this tolerance is applied at each element (rather than simply being subtracted from the summary measure). The similarity for NS is not as high as SM when some ΔL is introduced because the NS still has some differences that are greater than the tolerance.

Shifts: When ΔV and ΔL are applied, the LS and GS fields have almost zero error and perfect similarity. The improvement from the standard RMSE is more apparent with GS due to the complete shift, while LS produces a less dramatic change in values. The introduced shift of 20m led to a standard RMSE of 2.60 with no tolerance, while tolerances correctly recognise this as being almost identical. Standard measures

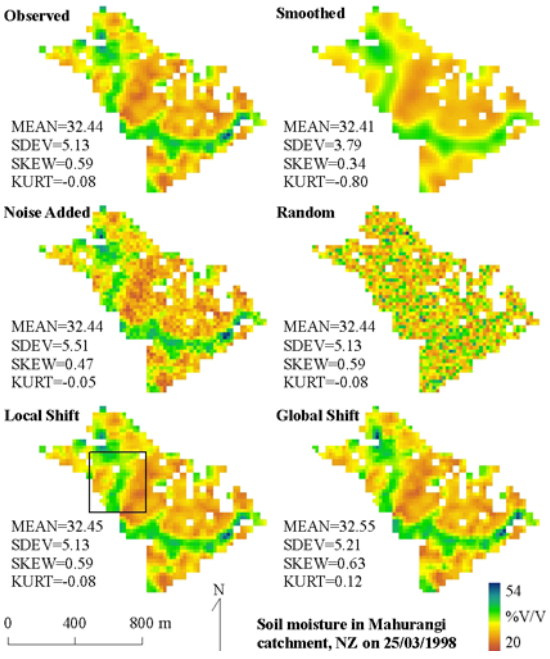

Figure 3 Five synthetic fields have been created from the observed field for comparison. These are: observations smoothed with mean filter (SM); observations plus σ =2% random noise (NS); observations with random arrangement (RN); observations within box shifted 20m NE (LS); and all observations shifted 20m NE (GS). The global summary measures are very similar for each field.

cannot quantify this.

Smoothing: The SM field is judged more similar (and having less error) than the NS field when both tolerances are applied. The smoothing process makes nearby values less variable, making it synonymous with a scale difference between fields. These results illustrate how allowing small tolerances can deal with consistent scale differences.

Random: The random field produces higher error measures and lower similarity under all tolerances. The few elements that were coincidentally similar are rewarded in the similarity measure. The relative measure (SE) reflects that the random field is a poor model of the observed under all tolerances, while the other measures are better models than the mean field.

As expected, the tolerant measures show a strong response (i.e. a major change in value) when the fields compared have globally consistent differences (e.g. NS, GS, SM). When the differences are less consistent (e.g. localised), the response is less apparent (due to being averaged over all elements). Inspecting the graphical measures visually can reveal the causes, but an automated method would involve performing the analysis on limited parts of the field. This is another promising extension to current approaches that was suggested in Wealands *et al.* (2005) but it is not pursued further in this paper.

Individually, the measures provide a statement of similarity or error under specified conditions, which is useful because it quantifies a process that is often done visually. However, looking at the change of a measure under different tolerances can reveal thresholds at which more substantial improvements occur. This can be used to infer that noise, shifting or some kind of scale inconsistency exists between the fields.

Table 1 Comparison results for error (RMSE) and similarity measures ($\overline{S}$ and SE). Column headings denote the tolerance values used for each comparison (e.g. 0,0 means no tolerances for value or location differences). Values that are shaded have been referred to in the results section.

RMSE (ΔV (%), ΔL (m))

	0,0	2,0	0,30	2,30
SM	2.22	1.09	0.61	0.06
NS	2.00	0.54	0.91	0.15
RN	7.29	5.79	4.58	3.41
LS	0.73	0.27	0.17	0.01
GS	2.60	1.42	0.64	0.18
MN	*5.13*	*3.62*	*2.43*	*1.33*

$\overline{S}$ (ΔV to ΔVMAX, ΔL to ΔLMAX)

	2-2,0-0	2-2,30-30	0-5,0-60	0-10,0-90
SM	0.68	0.99	0.70	0.84
NS	0.56	0.95	0.66	0.83
RN	0.22	0.55	0.34	0.54
LS	0.96	1.00	0.96	0.98
GS	0.62	0.98	0.67	0.82
MN	*0.25*	*0.68*	*0.41*	*0.64*

SE (ΔV to ΔVMAX, ΔL to ΔLMAX)

	2-2,0-0	2-2,30-30	0-5,0-60	0-10,0-90
SM	0.57	0.96	0.50	0.56
NS	0.42	0.84	0.42	0.52
RN	-0.04	-0.41	-0.12	-0.25
LS	0.94	1.00	0.92	0.94
GS	0.49	0.93	0.45	0.51
MN	*0.00*	*0.00*	*0.00*	*0.00*

5. DISCUSSION

Tolerances for value differences (ΔV) achieve two different tasks. Firstly, they allow observation errors (e.g. noise) to be managed for each individual element during comparison. This produces an estimate of average error under these conditions, rather than having to evaluate the average error (e.g. RMSE) against the observation error globally. The second role for ΔV is to specify what values will be considered similar. ΔV will usually be small, as value differences will want to be penalised. For fuzzy tolerances the ΔVMAX controls how severe the penalty is for differences. This is quite different from categorical comparisons, where each element clearly belongs to one category or another. The use of tolerances with continuous data effectively specifies category boundaries for each individual element during comparison. Without them, similarity between continuous values cannot be defined.

Building in tolerances at the element level opens up a number of other potential tolerance options. Apart from the locational tolerance described in this paper, most topologies can define which neighbouring elements are up or downhill (by using additional elevation information). The tolerances can be specified to look, say, only at upslope neighbours in the observed field, thus making the measure tolerant for situations where the modelled value is too far downslope. If a time series of spatial fields exists for assessment, a similar tolerance could be implemented for element values that are modelled too early or late. Where the data facilitates such analysis, a

diagnostic field showing where each tolerance has been invoked could also provide useful feedback.

To draw specific hydrological meaning (rather than overall model assessment) from the results would require further localisation of the results. The summary measures presented here identify the average response of the field to differences in value or location. If the measures are calculated for more localised parts of the field (e.g. specific parts of the landscape), more specific reasons for the response can be determined. Using the soil moisture example, if high errors tend to be found on hillslopes, but when tolerating shifts these are reduced more than other areas, then the hydraulic conductivity parameter (which controls the rate of moisture movement) for the hillslopes may have been incorrect. Alternatively, visual inspection of the error or similarity fields produced can reveal this, but not quantify its impact on model performance.

The measures introduced here are for use after initial comparisons (e.g. bias, RMSE) are made. They facilitate further investigation of similarity between fields, which can reveal fields that are similar under tolerances that were not considered similar otherwise.

6. CONCLUSION

The comparison methods presented in this paper are new measures based on the work of Hagen (2003). They apply tolerances for differences in value and location to spatial fields containing interval/ratio data, the type most commonly produced by distributed hydrological models. Observing the response of these measures to changes in tolerance can highlight fields that have regular differences in values (e.g. observation noise) and/or location (e.g. georeferencing or modelling issues). They utilise values from nearby elements to include aspects of spatial arrangement. These measures are readily implemented with any type of spatial field and can be used for model assessment against observations (i.e. a reality) or other models. They allow the user to specify exact or variable tolerances and produce measures that are readily interpreted.

7. REFERENCES

Beven, K.J. (1993), Prophecy, reality and uncertainty in distributed hydrological modelling, *Advances in Water Resources*, 16, 41-51.

Grayson, R.B., G. Blöschl, A.W. Western and T.A. McMahon (2002), Advances in the use of observed spatial patterns of catchment hydrological response, *Advances in Water Resources*, 25, 1313-1334.

Güntner, A., J. Seibert and S. Uhlenbrook (2004), Modeling spatial patterns of saturated areas: An evaluation of different terrain indices, *Water Resources Research*, 40(5), W05114.

Gupta, H.V., S. Sorooshian and P.O. Yapo (1998), Toward improved calibration of hydrologic models: Multiple and noncommensurable measures of information, *Water Resources Research*, 34(4), 751-763.

Hagen, A. (2003), Fuzzy set approach to assessing similarity of categorical maps, *International Journal of Geographical Information Science*, 17(3), 235-249.

Jetten, V., G. Govers and R. Hessel (2003), Erosion models: Quality of spatial predictions, *Hydrological Processes*, 17(5), 887-900.

Nash, J.E. and J.V. Sutcliffe (1970), River flow forecasting through conceptual models, i, a discussion of principles, *Journal of Hydrology*, 10, 282-290.

Power, C., A. Simms and R. White (2001), Hierarchical fuzzy pattern matching for the regional comparison of land use maps, *International Journal of Geographical Information Science*, 15(1), 77-100.

Wealands, S.R., R.B. Grayson and J.P. Walker (2005), Quantitative comparison of spatial fields for hydrological model assessment--some promising approaches, *Advances in Water Resources*, 28(1), 15-32.

Western, A.W., R.B. Grayson and T.R. Green (1999), The Tarrawarra project: High resolution spatial measurement, modelling and analysis of soil moisture and hydrological response, *Hydrological Processes*, 13(5), 633-652.

Wilson, D.J., A.W. Western and R.B. Grayson (2005), A terrain and data-based method for generating the spatial distribution of soil moisture, *Advances in Water Resources*, 28(1), 43-54.