RMetS

ROYAL METEOROLOGICAL SOCIETY

Evaluating forecasts of extreme events for hydrological applications: an approach for screening unfamiliar performance measures

Hannah L. Cloke[a]* and Florian Pappenberger[b]
[a] *Department of Geography, King's College London, London, UK*
[b] *European Centre for Medium-Range Weather Forecasts, Reading, UK*

ABSTRACT: Many different performance measures have been developed to evaluate field predictions in meteorology. However, a researcher or practitioner encountering a new or unfamiliar measure may have difficulty in interpreting its results, which may lead to them avoiding new measures and relying on those that are familiar. In the context of evaluating forecasts of extreme events for hydrological applications, this article aims to promote the use of a range of performance measures. Some of the types of performance measures that are introduced in order to demonstrate a six-step approach to tackle a new measure. Using the example of the European Centre for Medium-Range Weather Forecasts (ECMWF) ensemble precipitation predictions for the Danube floods of July and August 2002, to show how to use new performance measures with this approach and the way to choose between different performance measures based on their suitability for the task at hand is shown. Copyright © 2008 Royal Meteorological Society

1. Introduction

There are many different methods with which a meteorological practitioner or researcher can measure the quality of precipitation forecasts. One of the most familiar methods is the Brier Skill Score, but for more examples see references in Gober *et al*. (2004); Venugopal *et al*. (2005); Weisheimer *et al*. (2005) and Gandin and Murphy (1992). In addition, many more methods have been developed in other related disciplines such as hydrology (Pappenberger and Beven, 2004; Hagen-Zanker, 2006; Hagen-Zanker *et al*., 2006; Pappenberger *et al*., 2006b; Pappenburger *et al*., 2007b). These *performance measures* (also known as *scores*) are usually used to compare a forecasted field with an observed field (see definition by Murphy and Winkler (1987)).

With such a large number of performance measures available, a researcher or practitioner may find that newly developed or unfamiliar measures are difficult to understand and that the meaning of the computed numeric is not adequately conveyed to them. For example, if a measure has results that range from 0 to 1, with 1 being optimal, a value of 0.7 would be meaningless to the unfamiliar user without further explanation and training. Only by using this new measure alongside another familiar measure, or by using it on many different cases,

can a meaning be established. There is often a reliance on craft skill and 'tacit ways of knowing' (Polanyi, 1967) in meteorological verification. A perception that a new or unfamiliar measure is 'too troublesome' may lead to reliance on those that are familiar. This could have important consequences, as the selection of a performance measure can affect inferences about the quality and uncertainty of a forecast (Venugopal *et al*., 2005). This is doubly important for interdisciplinary work, as, for example, hydrologists and meteorologists come from different scientific cultures (Demeritt *et al*., 2007) and are familiar with different measures.

In the context of evaluating forecasts of extreme events for hydrological applications, this article aims to promote the use of a range of performance measures so that the quality of a forecast can be assessed as rigorously as possible. A six-step approach for using new or unfamiliar measures is established, including comparing them with familiar measures and selecting a subset of measures to use based on their suitability for the task at hand (Section 2). Some of the types of performance measure that are available are introduced and a description of those that are used, in particular in this article, is given (Section 3). An example case to illustrate the approach and to show how to choose performance measures according to the evaluation task at hand is then presented (Section 4). The example focuses on the precipitation predictions by the Ensemble Prediction System (EPS) of the European Centre for

---

* Correspondence to: Hannah L. Cloke, Department of Geography, King's College London, London, UK. E-mail: hannah.cloke@kcl.ac.uk

Medium-Range Weather Forecasts (ECMWF) for the July and August 2002 Danube floods. This event has been chosen as it was one of the most extreme events of recent years in Europe, in which extreme rainfall led to extreme flooding. The possibility of eliminating measures based on the properties calculated is illustrated, and a subset of measures for future use is derived.

Although this article focuses on performance measures that compare fields of hydrological extreme events (rainfall) that are likely to lead to flood runoff, the methodology can be extended to other meteorological issues and also time series comparison. In addition, although this article concentrates on deterministic comparison, the concepts could be extended to probabilistic measures.

2. A six-step approach to screen performance measures

A six-step process for screening an unfamiliar measure in order to develop understanding and so that the measure can be used effectively. The approach is described in Table I, and is based on our experience with trying to understand new measures. When following this approach it is suggested that the results of each step are clearly documented, which will increase the possibility that a high value of a performance measure really means a good forecast (Mason, 2007). This remains a research paper and not an instruction manual. Those readers unfamiliar with the details of various performance measures should refer to the references given.

Table I. Six-step approach for becoming familiar with a new performance measure and using it effectively.

Step	Name	Description
1	Classification	Measure is classified according to some generic properties
2	Scatterplots	Measures are directly compared
3	Magnitude analysis	Robustness of the measure is evaluated with respect to scaled forecasts
4	Displacement analysis	Robustness of the measure is evaluated with respect to shifted forecasts
5	Spatial dependences	Performance of fields with known spatial dependence is computed
6	Human visual systems	Measure is compared to the results of a subjective analysis of a forecaster

The basis of our approach lies in the concept of 'meta verification', which determines whether or not verification methods satisfy specific criteria and/or possess particular properties. The outline of the concept was first suggested by Murphy (1991, 1996) and has been used previously in the context of hedging (Jolliffe, 2007). Here, an approach and application to a hydrometeorological problem is described. The interested reader is also referred to the Spatial Forecast Verification Intercomparison Project (http://www.ral.ucar.edu/projects/icp), which promotes a similar approach to the comparison and learning of novel performance metrics.

2.1. STEP 1: classification of measures

Each measure used should first be classified according to the categories described in Table II in order to give the foundation of the approach. This classification concentrates on the characteristics of the measures and not on the aspects of forecast quality that they measure. Thus, measures are classified into multiple categories such as deterministic, probabilistic, continuous, categorical, physical, vulnerability and many others as described in Table II. The basis for the categories used are from Weisheimer et al. (2005) and Murphy (1987). The categories are not clear-cut, and for example, 'categorical' measures based on multiple thresholds could also, at the same time, be 'continuous' measures. Also, any spatially ignorant measure could be easily transformed into a spatially aware measure by applying a moving-window approach.

2.2. STEP 2: scatterplots

The second step is for the measures to be compared with scatterplots. A scatterplot is a graphical representation consisting of ordered pairs possibly showing a relationship between two variable quantities. It allows one to investigate whether performance measures behave similarly, both in general but also in the tails of the distribution. Measures that show exactly the same response as another measure can be excluded from further analyses, as they give no additional information.

It is recommended that a substantial amount of time is spent exploring the relationships between different methods, e.g by studying scatterplots, in order to gain the craft skill needed to use measures effectively.

2.3. STEP 3: magnitude

Third, the sensitivity of measures to changes in magnitude should be evaluated and understood. Magnitude errors control the volume of water routed through a hydrological system, and thus are one controlling factor of the shape of a flood hydrograph (Obled et al., 1994). The influence of magnitude errors can be evaluated by comparing percentiles of, for example, an EPS. For this, the percentiles of the distribution given by precipitation forecasts or observations are computed and the resulting maps are compared.

Table II. Categories of performance measures.

Category	Description	Example
Deterministic/non-probabilistic	Compare one observation with one forecast.	Absolute error
Probabilistic (1 M/O-Way)	Acknowledge the uncertainty in model/observations only by comparing one observation with multiple forecasts. M stands for modelled and O for observed. 1-M Way means that the model results are treated as probabilistic and 1-O Way means that the observations are treated as probabilistic	Brier Score
Probabilistic (2 M/O-Way)	Acknowledge the uncertainty in forecasts and observations by comparing a probability distribution of observations with a probability distribution of forecasts.	Ignorance score (Roulston and Smith, 2002)
Continuous	Use of continuous variables such as precipitation volume	Absolute error
Categorical	Use of categorical variables such as occurrence of precipitation above a certain limit	Brier Skill Score
Object oriented	Measure can be used to evaluate objects such as catchments areas	Fractional Storm Coverage (Paulat *et al.*, 2007)
Grid oriented	Measure can be used to evaluate grids	Comparison of forecasted field with observed field
Hazard	Evaluation based on direct physical model results	Comparison of precipitation predictions
Risk	Evaluation which takes account of additional components such as vulnerability	Predictions are weighted by additional properties such as soil moisture (Pappenberger *et al.*, 2006a, 2007a)
Spatially ignorant	Ignore spatial correlation of fields	Absolute error
Spatially aware	Incorporate spatial correlation of fields	Fuzzy measure (Hagen-Zanker, 2006)
Temporally ignorant	Ignore temporal correlation of fields	Absolute error
Temporally aware	Incorporate temporal correlation of fields	Fuzzy measures
Scale ignorant	Ignore scaling effects	Absolute error
Scale aware	Incorporate scaling effects	Wavelets (Briggs and Levine, 1997); see also references quoted by Casati (2007)
Formal	Based on formal Bayesian Likelihoods	Woodhead measure (Woodhead, 2007) for binary fields and Bayesian Correlation Score (Krzysztofowicz, 2002) for a continuous variable. Jewson (2006) provides an introduction for meteorology
Strictly proper	A forecaster maximizes (or minimizes, depending on the nature of the score) by forecasting exactly his or her true beliefs about the situation.	Ignorance Score, Probability Score, Brier Score
Improper	Not strictly proper (see discussion in Wilson and Gneiting, 2007)	The Linear Score (Broecker and Smith, 2007)
Sufficient	If a performance measure is sufficient then it provides an unequivocal ordering on the quality of forecast (for discussion see Murphy, 1996 and Jolliffe and Stephenson, 2003)	See formal Bayesian
Equitable	Has the same no-skill value for random forecasts and for univarying forecasts of a constant category	Kuiper's Performance Index (Murphy and Daan, 1985)

2.4. STEP 4: displacement

Fourth, the sensitivity of measures to displacement should be evaluated and understood. Although it may seem obvious that a good quantitative precipitation forecast system will predict the correct location of a precipitation system (Ebert and McBride, 2000), displacement errors occur regularly. Moreover, spatial patterns of precipitation can play an important role in the prediction of flow hydrographs (Arnaud *et al.*, 2002). Several techniques such as the continuous rain area (Ebert and McBride, 2000) or domain-based methods (Hoffman *et al.*, 1995) have been developed to quantify the location (displacement) errors. However, it is argued that the displacement error cannot be directly quantified for most forecasts because of the covariance of the errors. Displacement can be evaluated by shifting or rotating a field and comparing it to its original.

2.5. STEP 5: spatial dependency

Fifth, it is important to understand the relationship between the spatial correlation of the field and the performance measures. This is because performance measures, especially spatially aware measures, can be very sensitive to such correlations. Evaluation based on spatial dependence is very similar to the displacement analysis.

2.6. STEP 6: human visual experiment (eyeball verification)

Sixth, the use of eyeball verification, which is often used as a standard verification tool (e.g. Mariani *et al.*, 2005) and tests for consistency between numerical value and forecasters' experience and opinions is suggested. Although time consuming, it is a valuable tool for evaluating location, size, shape, magnitude and patterns, and many measures have tried to mimic it (see, for example, references quoted in Shnayderman *et al.*, 2006). However, realistically only a limited amount of information can be reliably compared in this way, as the way in which the brain detects and discriminates such information is not yet fully understood (Olzak and Wickens, 1999).

The values of the performance measures should be compared to the criteria used in the eyeball verification directly. If there is no correlation between the performance measures and the individual criteria, it is possible to test for higher-dimensional relationships with the help of a regression tree. The method uses the computed performance measures as input variables and predicts the criteria of the eyeball verification. A regression tree is a sequence of questions that can be answered as 'yes' or 'no', plus a set of fitted response values. Each question asks whether a predictor satisfies a given condition. Predictors can be continuous or discrete. Depending on the answers to one question, one either proceeds to answer another question or arrives at a fitted response value. There are many different types of regression trees (see

references in Pappenberger *et al.*, 2006a), but the methodology of Breiman and Cutler (2004), which generates multiple regression trees, a 'random forest', by taking account of the uncertainty in the model structure, is recommended. A more detailed explanation of this methodology is beyond the scope of this article and the reader is, therefore, referred to related studies (Breiman *et al.*, 1984; Grieb *et al.*, 1999; Dietterich, 2000; Bobbin and Recknagel, 2001; Freund, 2001; Ho, 2002; Breiman and Cutler, 2004; Pappenberger *et al.*, 2006a). The random forest is optimized on 80% of the training dataset, and 20% is used for verification.

3. Review of some of the performance measure available

In any particular investigation, the initial use of as many performance measures as time and resources allow is advocated. Table II gives the main categories into which any measure can be classified. Here, 15 different deterministic, continuous, hazard performance measures, which range from fuzzy-based to more formal approaches, have been selected. Deterministic-based performance measures play an important role in all applications in which models can be run multiple times for calibration purposes (calibration is defined here as adjusting effective parameters or factors of the model to fit the observations).

As well as traditional measures such as correlation and root mean square error, seven measures that the reader may be more unfamiliar with have been selected. Each measure is only briefly introduced, as further information can easily be found in the quoted references. A summary of the measures is given in Table III. The definitions for the equations are only given the first time the variables are used, and are summarized in the Appendix. Concentration is on the overall performance measures of the entire fields, and thus, the aggregated information content. Most spatially aware (and spatially unaware) measures are originally designed to attribute errors towards particular grid cells and have aggregated information only as a secondary information content, and so, they only present an average of the underlying fields.

3.1. Fuzzy numerical space (FNS)

The fuzzy numerical measure was first introduced by Hagen (2003) and Hagen–Zanker *et al.* (2006). It compares maps with a moving-window-based approach using fuzzy logic to compute a measure of similarity.

$$s_i(A, B) = \max_j^N(f(A_i, B_j) \times w(d_{i,j}))$$
$$S_i(A, B) = \min(s_i(A, B), s_i(B, A))$$
$$S(A, B) = \frac{1}{n} \sum_{i=1}^{n} S_i(A, B) \qquad (1)$$

$$f(a, b) = 1 - \frac{|a - b|}{\max(|a|, |b|)} \qquad (2)$$

Table III. Summary of performance measures used in this study.

Abbreviation	Performance measure	References
BSA	Bivariate Spatial Association	Lee (2001)
BSSS	Bivariate Spatial Smoothing Scalar	Tan *et al.* (2006)
CoD	Cosine distance	Tan *et al.* (2006)
Corr	Correlation	Tan *et al.* (2006)
FNS	Fuzzy numerical Space	Hagen (2003); Hagen-Zanker *et al.* (2006)
HaD	Hamming distance	Tan *et al.* (2006)
IQA	Image quality assessment	Wang *et al.* (2004)
IWC	Information weighted comparison	Tompa *et al.* (2000)
JaD	Jaccard distance	Tan *et al.* (2006)
LM	Local Moran	Zang and Gove 2005; Hagen-Zanker *et al.* (2006)
MAE	Mean absolute error	Tan *et al.* (2006)
MSE	Mean squared error	Tan *et al.* (2006)
PSNR	Peak signal-to-noise ratio	Tan *et al.* (2006)
RMSE	Root mean squared error	Tan *et al.* (2006)
SVD	Singular vector decomposition	Shnayderman *et al.* (2006)
WAV	Wavelets	Briggs and Levine (1997); Hagen-Zanker (2006)

where: A and B are map A (observed) and B (modelled); n is the number of cells; s is the one-way similarity; S is the overall similarity; i, j is the cell index; $f(a, b)$ is the similarity measure; N is the number of cells in proximity; $w(d)$ is the distance weight.

The similarity measure can be replaced by different formulations. The distance weight is here given by the Euclidian distance. The number of neighbouring cells in proximity is set to 4.

3.2. Image quality assessment (IQA)

In this article the IQA follows the method of Wang *et al.* (2004) who decomposed the images into luminance (l), contrast (ct) and structure (st) within a moving-window approach. The method is conceptually similar to the wavelet decomposition (see Section 3.7) and the FNS (see Section 3.1).

$$\mu_{A_x} = \sum_{i=1}^{N} w_{G_i} A_{x_i}$$

$$\sigma_{A_x B_x} = \left(\sum_{i=1}^{N} w_{G_i} (A_{x_i} - \mu_{A_x})(B_{x_i} - \mu_{B_x}) \right)^{0.5}$$

$$\sigma_{A_x} = \left(\sum_{i=1}^{N} w_{G_i} (A_{x_i} - \mu_{A_x})^2 \right)^{0.5} \quad (3)$$

where w_G is the distance weight based on a normalized Gaussian function with a standard deviation of 1.5, A_x is the window used in map A, B_x is the window used in map B, σ is the standard deviation and μ is the mean.

Equation (4) is then used to compute the image properties:

$$l(A_x, B_x) = \frac{2\mu_{A_x}\mu_{B_x} + C_1}{\mu_{A_x}^2 + \mu_{B_x}^2 + C_1}$$

$$ct(A_x, B_x) = \frac{2\sigma_{A_x}\sigma_{A_x} + C_2}{\sigma_{A_x}^2 + \sigma_{A_x}^2 + C_2}$$

$$st(A_x, B_x) = \frac{\sigma_{A_x B_x} + C_3}{\sigma_{A_x}\sigma_{A_x} + C_3} \quad (4)$$

where l is the range of pixel values (set to the absolute maximum) and C is a constant with $C_1 = (0.01 \times l)^2$, $C_2 = (0.03 \times l)^2$, $C_2 = 0.5$.

The properties can be summarized in a similarity index:

$$\text{SSIM}(A_x, B_x) = c(A_x, B_x) s(A_x, B_x) l(A_x, B_x)$$

$$\text{IQA} = \frac{1}{N} \sum_{i=1}^{N} \text{SSIM}(A_{x_i}, B_{x_i}) \quad (5)$$

where SSIM is the similarity index of luminance, contrast and structure.

3.3. Information weighted comparison (IWC)

The IWC by Tompa *et al.* (2000) is a spatial derivate of the Ignorance Score. Values which are common in the map are weighted less than values which lie in uncommon ranges. It can be used either in a deterministic framework or in a 1-/2-way probabilistic one.

$$I_A(z) = \log\left(\frac{1}{P_A(z)}\right)$$

$$\text{IMSE}_i(A, B) = (A_i I_A(A_i) - B_i I_B(B_i))^2$$

$$\text{IWC}(A, B) = \frac{1}{N} \sum_{i=1}^{n} \text{IMSE}_i(A, B) \quad (6)$$

where $P(z)$ is the frequency of value z; I is the ignorance and IMSE is the mean weighted ignorance.

3.4. Local Moran (LM)

The LM approach has been primarily designed to cluster sources of errors in a local indicator (Zang and Gove, 2005). It computes a weighted average of mean errors based on an evaluation window.

$$ME = \frac{1}{N} \sum_{i=1}^{N} A_i - B_i$$

$$MC_i = ((A_i - B_i) - ME) \sum_{j=1}^{N} w(d_{i,j})$$

$$\times ((A_j - B_j) - ME)$$

$$LM = \frac{1}{N} \sum MC_i \tag{7}$$

where ME is the local mean error and MC is the weighted ME.

The weighting function has the form of a pyramidal frustum. Hagen-Zanker $et\ al.$ (2006) have pointed out that interpretation of the Local Moran maps is very complex, as for example, even small location errors will lead to large negative Moran values.

3.5. Bivariate spatial association (BSA)

Lee (2001) combines the LM measure described above (Section 3.4) with Pearson correlation to compute the BSA measure. The method compares two windows with a bivariate spatial smoothing scalar and multiplies the result by the Pearson correlation.

$$SSS_A = \sqrt{\frac{\sum_{i=1}^{N} (\hat{A}_{x_i} - ME)^2}{\sum_{i=1}^{N} (A_i - ME)^2}}$$

$$BSSS_{A,B} = SSS_A - SSS_B$$

$$BSA_{A,B} = \sqrt{BSSS_{A,B}} \times r_{A,B} \tag{8}$$

where r is the Pearson correlation, SSS is the spatial smoothing scalar and BSSS is the bivariate spatial smoothing scalar.

3.6. Singular vector decomposition (SVD)

Shnayderman $et\ al.$ (2006) developed a multidimensional image quality measure using singular value decomposition. Every real matrix can be factorized and decomposed into a product of three matrices of which one is a scalar by which each corresponding input is multiplied to give a corresponding output. This can be used in a graphical distance measure

$$SVD_i = \sqrt{\sum_{i=1}^{N} (\sigma_{A_X} - \sigma_{B_X})^2} \tag{9}$$

From this a global measure can be derived through averaging.

3.7. Wavelets (WAV)

The wavelet method has been pioneered by Briggs and Levine (1997) who decomposed the observations and forecasted fields into maps at different scales by a discrete wavelet transformation. The maps at the different scales are then compared by any similarity measure such as root mean squared error (RMSE) or anomaly correlation coefficient (ACC). Hagen–Zanker (2006) describes the three steps necessary for this analysis. First, the observed map is transformed into a series of maps representing different characteristic scales (wavelets) of the observed magnitude. The transformation with the lowest Shannon Entropy is chosen. Second, noise is removed with a soft threshold function (this is omitted in this research article due to the ambiguity of the definition of noise). Third, the maps are compared with RMSE. Hagen–Zanker (2006) points out that the decomposition is purely based on the information theory and thus, has limited physical meaning. Moreover, this method is extremely sensitive to offsetting of maps.

3.8. Traditional measures

Measures that are commonly used in meteorological verification include: mean squared error (MSE), mean absolute error (MAE), RMSE, peak signal-to-noise ratio (PSNR), Hamming distance (HaD), Jaccard distance (JaD), Cosine distance (CoD) and correlation (Corr). For the last, the Pearson correlation has been used. These are not described in detail as it is assumed that the reader is familiar with the more common measures. However, a description can be found in Tan $et\ al.$ (2006).

4. A methodology for testing the approach

4.1. Description of case study: July and August 2002 flooding in the Danube catchment.

The approach described in Section 2 was tested with the measures described in Section 3 on precipitation data from a flood on the River Danube. The Danube has been chosen as it is part of the PREVIEW research program (www.preview-risk.com) which analyses the predictability of medium-range flood forecasts. PREVIEW is a project funded by the European Commission as part of the sixth framework program. This study concentrates on the flooding of July and August 2002, which affected over 600 000 people and caused damages in excess of 15×10^6 USD (EM-DAT, 2007). The major cause of this flooding was heavy rain. A high-altitude low-pressure system caused heavy precipitation in Germany and in lower and upper Austria between the 6th and 8th of August. Additionally, heavy precipitation occurred over Romania, South Bohemia and the eastern coastal regions of the Black Sea. On the 10th and 11th of August, a

second depression led to strong precipitation in northern and central Italy and generated torrential rainfall in upper Bavaria and Lower Saxony (SE and NW Germany). Many small and medium-sized rivers in Austria and Germany were in flood (SwissRe, 2003).

The precipitation prediction set used here consists of 50 ensemble members, 43 days in length, each with a forecast of 6 days ahead (258 forecasts in total). The ECMWF variable resolution ensemble prediction system (VAREPS) was used, which has a resolution of $T_L399L62$, with T42L62 singular vectors (Buizza *et al.*, 2006). The observations have been taken from a high-resolution network described by Ghelli and Lalaurette (2000).

4.2. Methodology for the six-step approach

First, the measures described in Section 3 will be classified according to Table II (step 1). Second, scatterplots of the results of the performance measures will be compared (step 2). For step 3, the analysis of magnitude, all forecasts are lumped together. From each of the 258 forecast distributions of 50 ensembles, the 10, 20, 30, 40, 60, 70, 80 and 90% percentiles are computed (2322 experiments). These percentiles are then compared to the 50% percentile of each forecast. Some displacement error is necessarily included in this analysis as, e.g. one forecast can predict the storm centre in one area of the domain and another elsewhere. However, this will be partially overcome by using the median of each cell.

For step 4, displacement errors are simulated by shifting the forecasted precipitation fields. The fields are shifted on the regular Gaussian grid (one field north, one field west, one field north/west, four fields north, four fields west, four fields north/west), in a total of 1548 experimental set ups. Edges of the fields are excluded from further analysis. It should be noted that this displacement error experiment will, indirectly, be dependent on the spatial correlation of the field.

For step 5, the spatial dependence is based on geostatistical simulations. A sequential Gaussian simulation is used to generate multiple maps based on variograms (Deutsch and Journel, 1998). The full variogram used in this study is based on the variograms fitted to observations, and can be described as:

$$\gamma(h) = c_1 \left[1.5\frac{h}{a} - 0.5\left(\frac{h}{a}\right)^3 \right] + c_0, \text{ for } h \leq a$$

$$\gamma(h) = c, \text{ for } h > a \qquad (10)$$

where: h is the lag (distance) between two locations; c_1 is the spatially correlated variance; c_0 is the spatially uncorrelated variance; a is the range of variogram; and c is the sum of c_1 and c_0 with a maximum of 1. This article does not explain the mathematical details of variogram analysis and refers the reader to Michaelides and Wilson (2007) for a relevant summary. Fifty alternative realizations (maps) have been created for ranges

of $a = 20$ and 40 and a fixed c_0 of 0.3 (which represents ranges estimated from observations). The larger the range, a, the more connected fields appear (in principle, moving from small-scale to large-scale storms). The effect of choosing different values for a and c is described in more detail by Michaelides and Wilson (2007). The distribution of the precipitation values of each map has been transformed from a normal distribution to an exponential distribution estimated based on the observations with inverse normal quantile transformation *via* probability matching (van der Waerdens, 1953; Kelly and Krzysztofowicz, 2000). This ensures that the artificial stochastic fields have similar properties to the observed fields. A displacement analysis is performed within these alternative maps and between the different maps. The derivates of the displacements with the two different ranges is computed and the fraction determines how sensitive a measure is towards spatial correlation.

For step 6, an interface has been developed, which allows a subjective performance classification (Figure 1). Forecasted images are selected at random and presented to the human interpreter who, as part of the interface, had unlimited time in which to respond. This methodology type is frequently used in other research fields (see, for example, Fairhurst and Lettington, 2000). In the top right corner an image of the observed data is displayed. In the bottom right corner is the forecasted image. The bottom left corner shows a difference map. The display of this difference map may condition the analysis somewhat, however, it was deemed necessary for better orientation after a trial run without it. A human interpreter conducted an evaluation based on four categories, each with measures ranging from 1 (excellent) to 5 (bad). It was also possible to give N/O (no opinion) in case of ambiguous forecasts. The four categories are overall impression, magnitude, pattern and displacement. Every 20 images or 10 min (whichever is reached first) the program encourages the evaluator to take a break. Many secondary factors such as tiredness, experience or light conditions can influence the analysis. The analysis can be based on raster data (as shown in Figure 1) or catchment outlines.

5. Case study results

5.1. STEP 1: results of classification of measures

In Table IV the performance measures used in this study have been classified according to their properties. In the Table a closed circle indicates what this particular measure was primarily designed for, an open circle suggests that this measure could be used in this context, and an empty space indicates that this measure is not suitable for this application. The distinction between categories is not always clear-cut and measures can fall into several categories at the same time. For example, a measure which is integrated over several thresholds can be seen as an evaluation of a continuous variable. It is apparent that in this case study only a certain subset of performance measures has been chosen, e.g. there
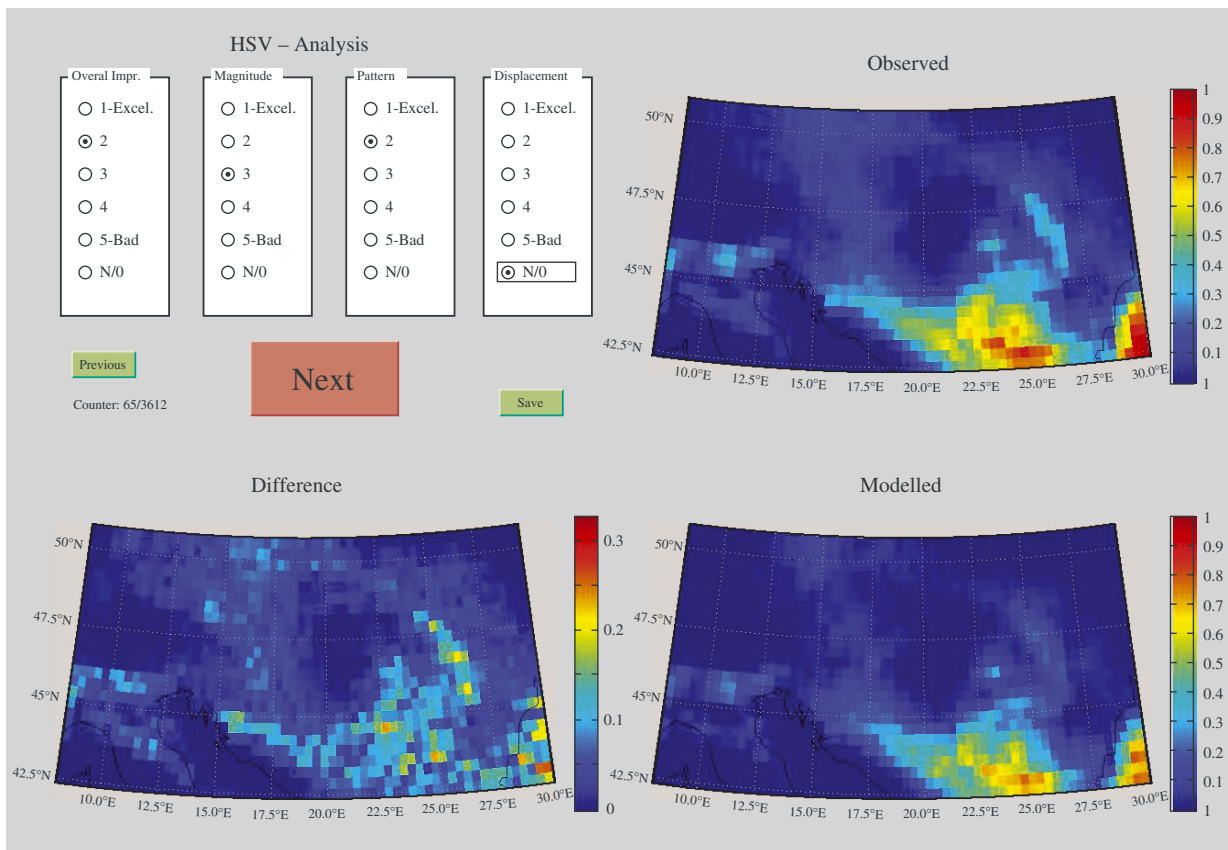
Figure 1. Human visual system interface for the comparison of observed and forecasted data.. This figure is available in colour online at www.interscience.wiley.com/ma

Table IV. Classification of performance measures used in this study. A closed circle means that this measure is primarily designed with this property. An open circle means that this measure can be easily adapted to contain this property. No entry means that this measure is unsuitable for this property. Details of the performance measures can be found in Table III.

Category	FNS	IQA	IWC	LM	BSA	SVD	WAV	MSE	MAE	RMSE	PSNR	HaD	JaD	CoD	Corr
Deterministic	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●
Probabilistic (1 M/O-Way)	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
Probabilistic (2 MO-Way)															
Continuous	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●
Categorical															
Object oriented								●	●	●		●	●	●	●
Raster oriented	●	●	●	●	●	●	●								
Hazard	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●
Risk	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
Spatially ignorant			●					●	●	●	●	●	●	●	●
Spatially aware	●	●	○	●	●	●	●	○	○	○	○	○	○	○	○
Temporally ignorant	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●
Temporally aware	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
Scale ignorant	●	●	●	●	●	●		●	●	●	●	●	●	●	●
Scale aware	○						●								
Formal															
Strictly proper															
Not strictly proper	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●

are no probabilistic 2-way measures. This is because we are merely seeking to establish a framework for performance measures. Probabilistic 2-way measures are easily incorporated into this approach but were not included so that testing of the human interface could be done in a timely fashion.

The classification of the measures in this way has given an overview of the capabilities of different measures. For example, IQA is clearly classified as spatially ignorant and so could not be expected to provide information about quality of the spatial distribution of the precipitation field.

5.2.   STEP 2: results of scatterplot analysis

In Figure 2 a plot matrix of the different performance measures is shown. In the scatterplots the different measures are plotted against each other for the same evaluation dataset. Histograms of the distribution of each measure are plotted on the diagonal of this plot. The evaluation measures have been normalized (with the results of the magnitude and displacement tests). The construction of this figure helps to get a 'feeling' for the different performance measures.

Some clear linear relationships can be observed (e.g. between LM and IWC). Such a definite linear relationship suggests that it is unnecessary to continue analysis with

both measures (as the value of one measure can be expressed by a simple regression with the value of the other measure).

It is apparent in Figure 2 that some performance measures do not cover the entire spectrum of normalized values. For example, the JaD is only observed between 0.95 and 1, and this is because the values have been normalized with the results of the displacement and magnitude experiment, and the JaD is relatively insensitive to magnitude errors (for more discussion, see Section 5.3). The image resolution would not show any dots if this adjustment were not made.

The histograms on the plot allow important conclusions to be drawn regarding the sensitivities to the type of error. The experiment has been designed so that there are equal numbers of percentual errors (e.g. there is an equal number of simulations which have an error of 40 percent points). Therefore, the distributions would be uniform if all errors were equal, for example, if an error between the 10th and 50th percentile were equal to an error between the 20th and 50th percentile. The FNS measure is the closest in achieving this ideal. The PSNR has a clear peak and a near-normal distribution, indicating that it weights towards small errors. In general, one cannot discriminate against a measure purely on these grounds as this can be an attractive property in
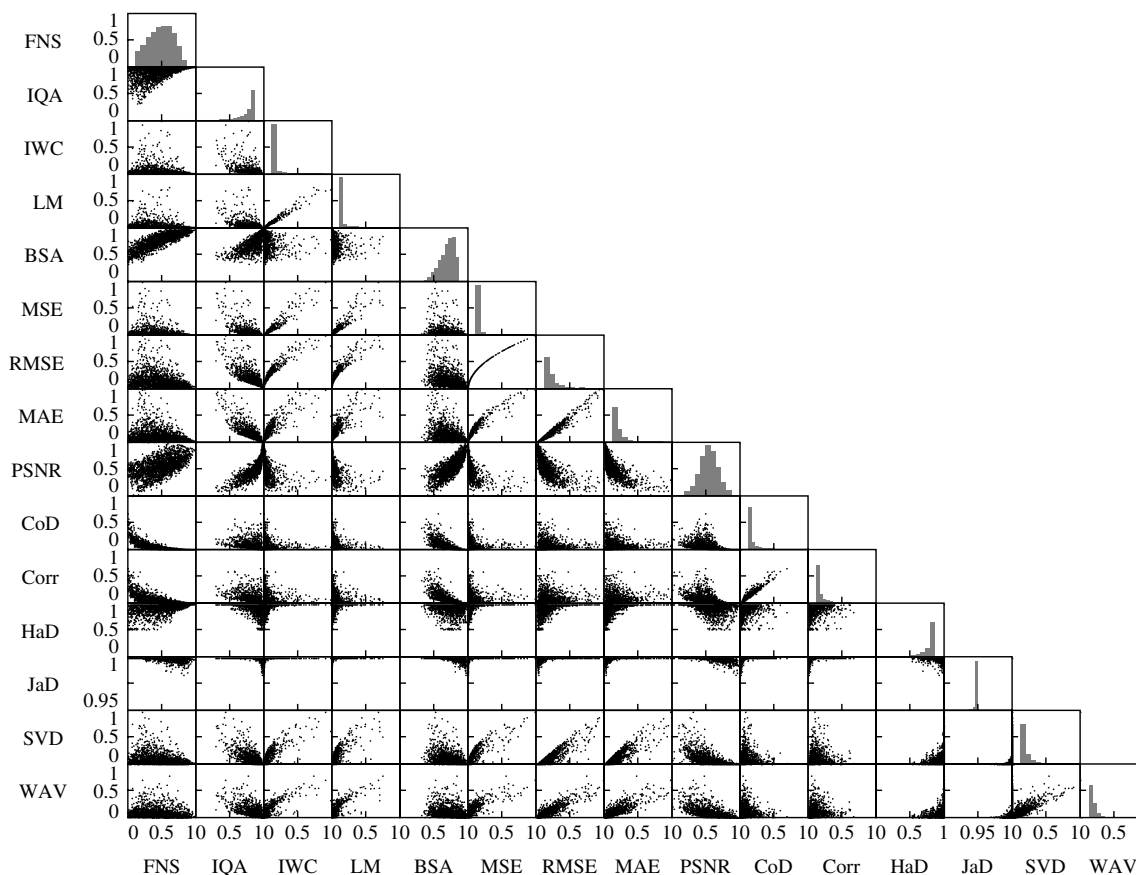

Figure 2. Comparison of various continuous measures in scatterplots and histograms. In the scatterplots the different measures are plotted against each other for the same evaluation dataset. Histograms of the distribution of each measure are plotted on the diagonal of this plot. The evaluation measures have been normalized between 0 and 1 (with the results of the magnitude and displacement tests). Details of the performance measures can be found in Table III.

applications different to the one discussed here. However, here a flood forecast is considered and thus the response will not be dominated by small errors. Small errors will not make much difference in the actual occurrence of the flood. Therefore, this measure is suitable for the present application.

In summary, for this case it is seen that the IWC can be removed from further consideration. A measure of how the various measures respond to the case study has also been obtained. For example, the IQA will be to some extent linearly related to the MAE, but will have no clear relationship to LM. Again, it is recommended that a substantial amount of time is spent exploring the relationships between different methods for any study, e.g by studying scatterplots, in order to gain the craft skill needed to use measures effectively.

5.3. STEP 3: results of the magnitude analysis

Figure 3 shows the box-and-whisker plots of the magnitude experiment in which the percentiles of the distribution of EPS forecast are compared. Each box has lines at the lower quartile, median and upper quartile. The whiskers are lines extending from each end of the box to show the extent of 1.5 quartiles. Outliers are data with values beyond the end of the whiskers (marked as a cross). It would be expected that performance measures

show sensitivity to the decreasing quality of the forecasts. For example, the average value of the FNS performance measure decreases with increasing error. The whisker indicates the magnitude of the uncertainty; the longer the bars the higher the uncertainty created by the individual measures. Certain measures can now be eliminated based on this analysis shown in Figure 3, because they are unsuitable for this case. For example, HaD exhibits far too large uncertainties to be useful for further analysis spreading the entire evaluation range from 0 to 1 and showing no distinction between 10, 20, 30 and 40%. The JaD measure shows nearly no sensitivity towards the increasing error (besides the outliers), and similar reasons disqualify LM, MSE and IWC (the latter has already been disqualified from the previous analysis).

5.4. STEP 4: results of the displacement analysis

In Figure 4, the box-and-whisker plots for the sensitivity of the measures towards displacement is shown.

Performances should decrease with increasing distance of displacement if the optimum performance of a measure is indicated by 1. Moreover, in the same scenario, the uncertainty bounds should increase with increasing distance as the correlation between the fields decreases. According to this analysis, the IWC, LM, MSE, HaD, JaD are not suitable for further use, as they do not behave as
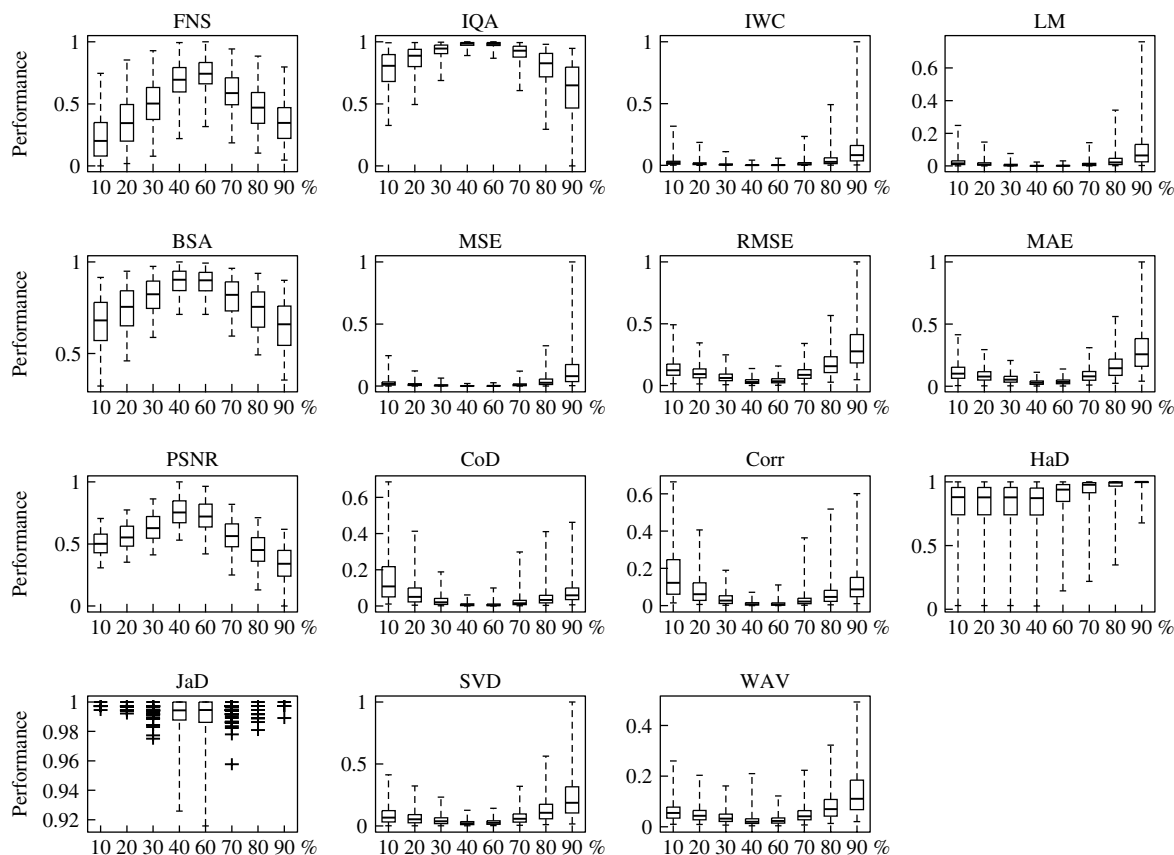

Figure 3. Box-and-whisker plots of the magnitude experiment. Each box has lines at the lower quartile, median and upper quartile. The whiskers are lines extending from each end of the box to show 1.5 times the inter-quartile range. Outliers are data with values beyond the end of the whiskers (marked as a cross). Details of the performance measures can be found in Table III.
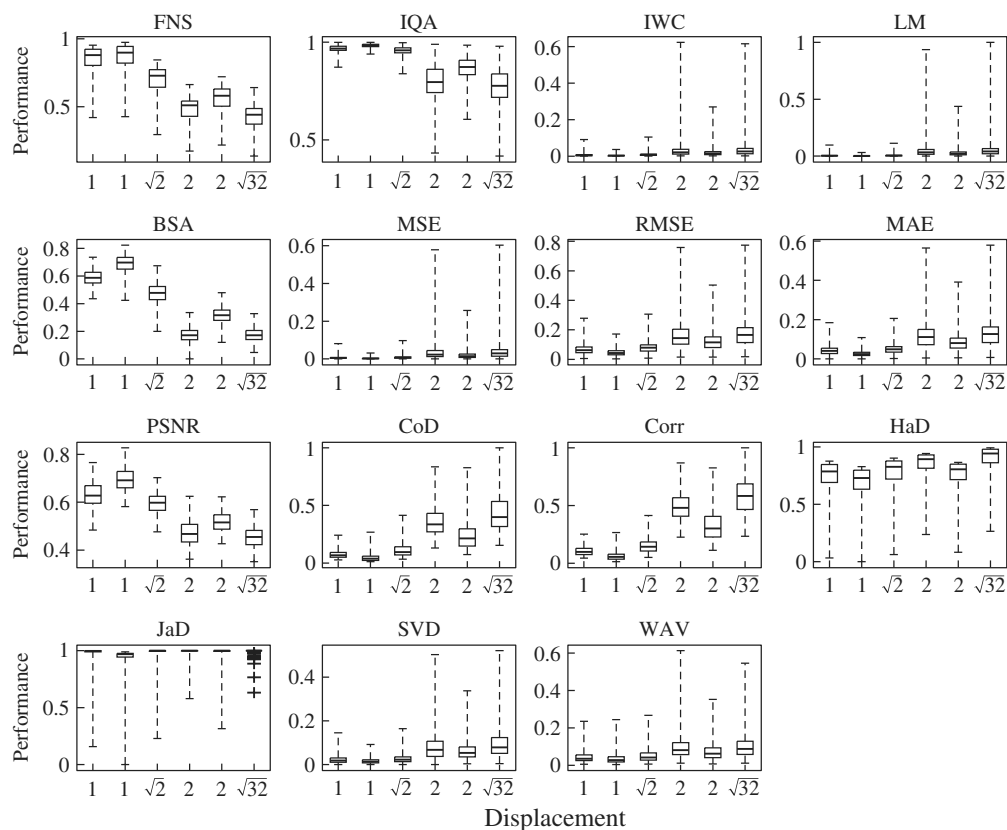
Figure 4. Box-and-whisker plots of the displacement experiment. Each box has lines at the lower quartile, median and upper quartile. The whiskers are lines extending from each end of the box to show 1.5 times the inter-quartile range. Outliers are data with values beyond the end of the whiskers (marked as a cross). On the x-axis is the Euclidian distance from the origin. Details of the performance measures can be found in Table III.

expected. In this case, these are the same measures that could be discarded based on the magnitude analysis, however, it is stressed that this may not always be the case.

5.5. STEP 5: results of the analysis of spatial dependency

In this analysis none of the measures are sensitive to a displacement of any of the maps created by the sequential Gaussian simulation and so results have not been shown graphically. This could be because variations in the range of the variogram are significantly larger than the steps of the displacement experiment. Smaller ranges cannot be chosen as this would interfere with the nugget effect (variance due to noise). This illustrates that for this case study all the measures chosen are robust towards large-scale changes of the geostatistical properties in the underlying fields and thus can be used for further analysis.

5.6. STEP 6: results of the human visual experiment

The human visual experiment, or eyeball verification, has been conducted with the same dataset as above. The interface shown in Figure 1 has been used by a human interpreter to rate various images. All image combinations have been analysed at least twice by two different human interpreters. The displacement category

in the interface was seen as problematic very early in the test as the interpreters were not able to make a clear distinction between the test images, and used primarily the 'no opinion' category. This may be partially the fault of the design of the experiment, which does not allow for enough spread.

In Figure 5 the results of the magnitude and displacement experiment are shown in a two-dimensional histogram. In the magnitude experiment one would expect a triangular shape in this diagram, with small errors (40 and 60%) getting high performance values and larger errors getting lower performance values. To improve the visual comparison, the frequencies have been computed based along the experimental designs, e.g. all percentages in the 10% column add up to 100%. The results shown in Figure 5 indicate that the variations of the frequencies become larger with larger errors. Small errors are not predominantly recognized and images with these errors are predominantly rated as excellent. However, an overall shape is visible. This is far less the case for the displacement experiment. A decreasing subjective performance rating would be expected with increasing displacement. This pattern cannot be seen, which indicates that the displacements are not large enough to have any visual impact. The fact that the magnitude shows a stronger signal than the displacement may also be due to the scientific training of the evaluators. Magnitude changes
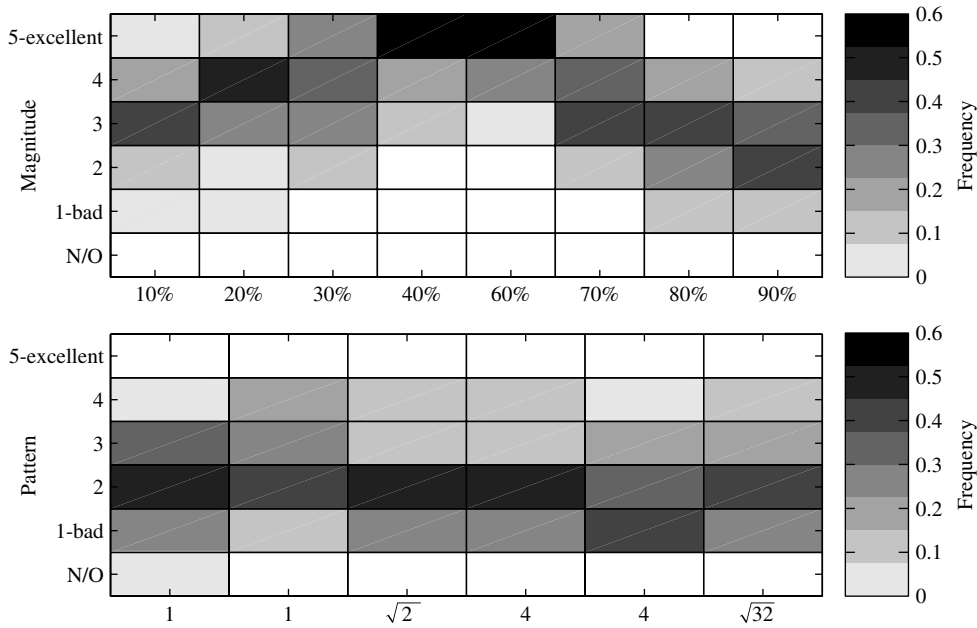
Figure 5. Comparison of eyeball tests to magnitude and displacement experiment in a two-dimensional frequency histogram (frequency computed along the experimental, y-axis, designs).

are very important in most flood applications and the displacements may have been seen as too small on the scale in question for this flood event.

In Figure 6 the pattern and magnitude rating are compared to the overall impression. All images show a positive correlation. The correlation of the magnitude to the overall impression seems to be stronger (less spread) than the evaluation values based on the pattern criteria.

The latter can be expected from the previous figures (Figures 3 and 4). The magnitude and pattern also show a positive correlation, which is probably less expected as no structure has been introduced into the experiment. This indicates that there is a visual correlation, which is not anticipated and cannot yet be fully explained. It is possible that if the human interpreter has a good opinion of the magnitude, pattern errors are seen more positively.
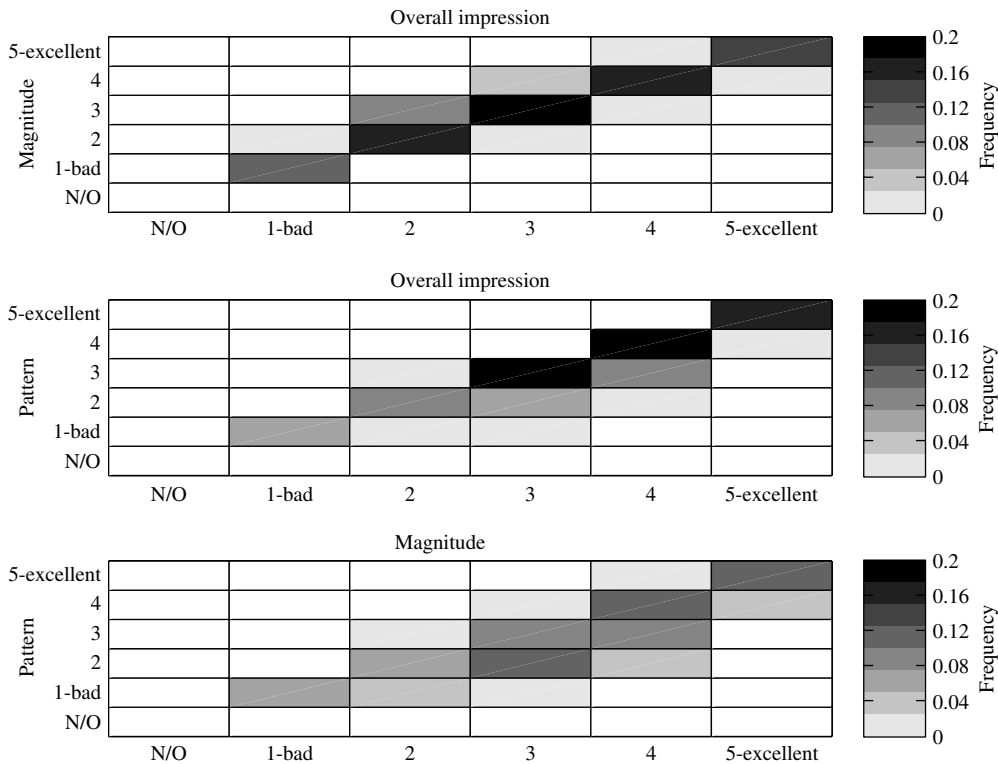

Figure 6. Comparison of eyeball categories in a two-dimensional frequency histogram (frequency computed to the number of total experiments).
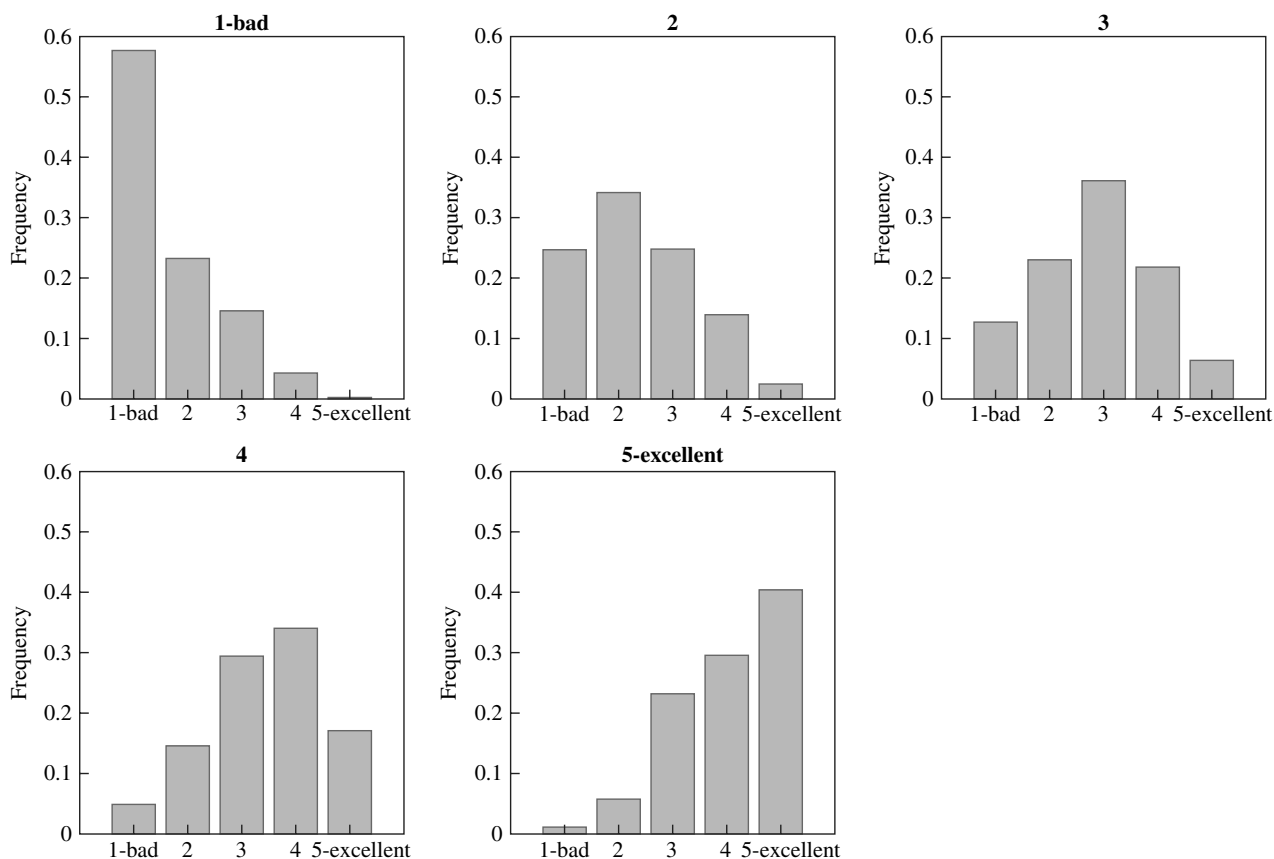
Figure 7. Histograms of the verification attempt to predict eyeball measures (overall impression) with continuous discrete performance measures. The bold text (figure titles) indicates results of the eyeball verification. The normal text (legend on x-axis) shows the measures predicted with the random forest.

The values of the performance measures are then compared to the criteria used in the eyeball verification directly. In this case, no correlation between the performance measures and the individual criteria has been found, and higher-dimensional relationships are tested for with the help of a regression tree (as described in Section 2.6). In Figure 7 the results of the regression tree analysis are displayed. The figure shows histograms of the verification of the attempt to predict eyeball measures (overall impression) with continuous discrete performance measures. The bold italic text (figure titles) stands for the results of the eyeball verification. The normal text (x-axis description) shows the measures predicted with the random forest. A clear relationship can be seen and there is a clear peak in each histogram at the location of the result of the eyeball experiment. Therefore, it is possible to predict the outcome of the eyeball verification exercise with the computed performance measures (incorporating uncertainties). The maximum frequencies are comparable, although there is a clearer peak at the top left plot at the 'bad' eyeball experiment results. It is obviously easier to make a distinction between clearly underperforming predictions compared to better predictions.

The performance measures can be ranked (Table V) according to their importance in predicting the outcome of the eyeball experiment (see Pappenberger *et al.*

(2006a) for details). The influence of all measures, which are not listed in Table V, is negligible. This ranking can be used to narrow the numbers of performance measures from those listed in Table V, which should be carried over for further use for the case in question. An important practical property for the evaluation criteria is the average computation time (Table VI).

At the end of step 6, there are five suitable measures, of which four are computationally feasible: IQA, PSNR, CoD and Corr are seen as the most appropriate measures for a further analysis of precipitation amounts in this region for this particular event. Table IV shows that our chosen performance measures cover all possible classes.

Table V. Ranking of performance measures according to importance in predicting the overall impression of the eyeball verification.

Rank	Measure
1	Corr
2	PSNR
3	IQA
4	CoD
5	FNS

Table VI. Average execution time on an Intel Pentium 4, 3.4 Ghz, 2Gb RAM using Matlab of selected deterministic continuous performance measures.

Measure	FNS	IQA	PSNR/CoD/Corr[a]
Time (s)	36	2	0.4

[a] Total time for all measures. Computed in one function due to high similarity.

For example, the IQA is raster oriented and spatially aware in contrast to the PSNR, which is not raster oriented and spatially ignorant.

5.7. Further analysis: comparison of magnitude and displacement experiment

In Figure 8 the results of the magnitude and displacement experiment are compared in a quantile–quantile plot of the distributions of all computed performance measures. From this it can be seen whether the change in magnitude or the shift of the different distributions has a greater influence (assuming that the size of the spatial shift and the size of magnitude shift have equal importance). If the dark thick line lies on the thin straight line, then a spatial shift and the magnitude shift have equal weights. This could be used as a criterion for measure selection, unless the difference in the reaction to magnitudinal/spatial shifts is required in the analysis. If the dark thick line is above the thin straight line then the magnitudinal errors have a greater influence on the performance measure and *vice versa*. No measure will be discarded based on this particular analysis. It allows quantification that the PSNR measure is more sensitive to magnitude in the upper end of the distribution. The final selection of measures used in any case should have dominant sensitivities towards both magnitude and displacement experimental setups. This is more important at the upper end of the performance distribution as good model that perform well are being sought.

The relationships depicted in Figure 8 should always be considered if the performance measures are applied in evaluating forecast fields.

6. Discussion

It has been shown in the previous Sections that for any particular analysis, one can easily narrow down the types of performance measures, which are suitable. Measures are excluded if they show no sensitivity towards the two experiments, or exhibit large uncertainty bounds. This
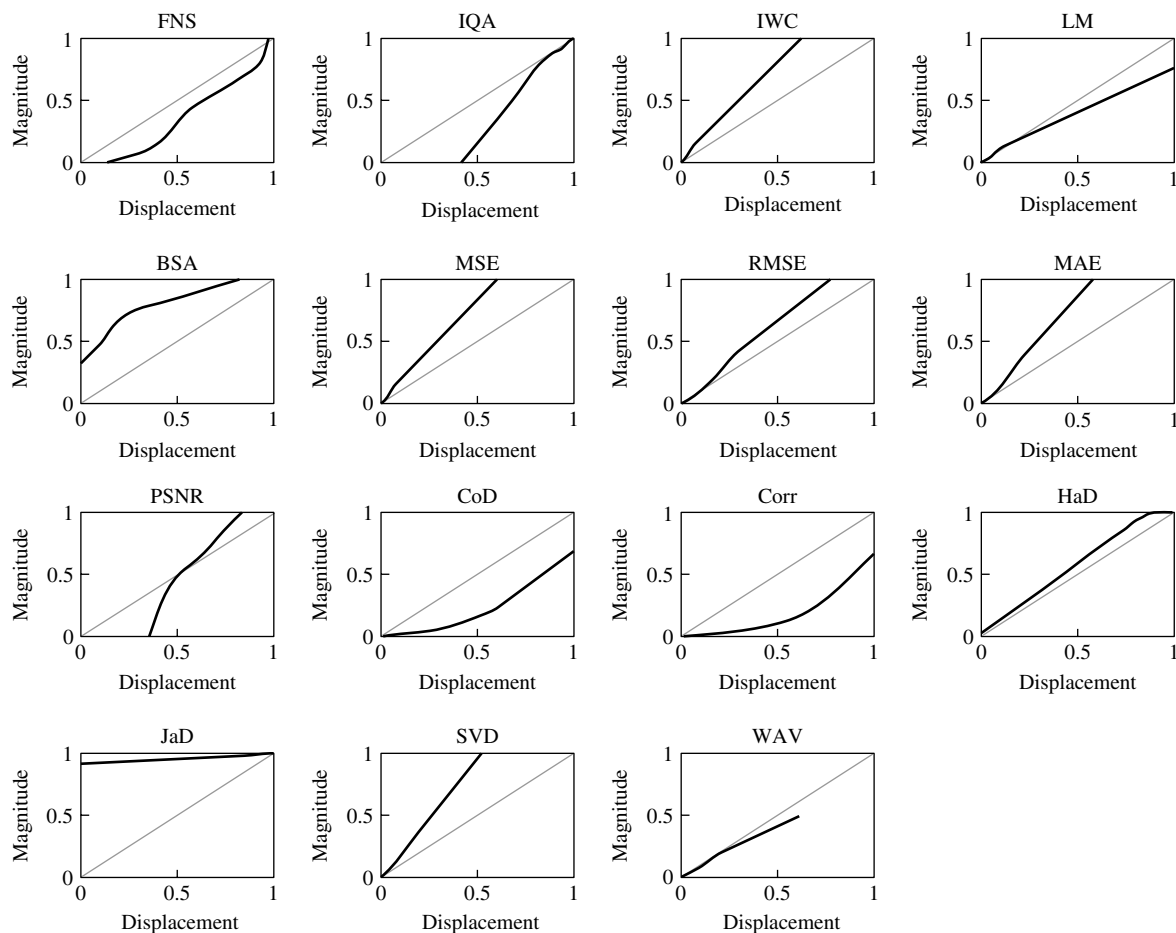

Figure 8. Comparison of the magnitude and displacement experiment in quantile–quantile plots for the continuous deterministic measures. Details of the performance measures can be found in Table III.

Table VII. Skill measures of precipitation for the July and August 2002 Danube floods. Details of the performance measures can be found in Table III.

Lead time (h)	IQA	PSNR	CoD	Corr
42	0.52	26	0.80	0.92
66	0.50	25	0.79	0.91
90	0.50	25	0.79	0.90
114	0.50	24	0.78	0.90
138	0.49	23	0.77	0.88
162	0.47	22	0.75	0.86

article recommends the use of IQA, PSNR, CoD and Corr for the deterministic assessment of forecasts for this particular variable (precipitation), region (Danube) and event (floods of July and August, 2002). The range of these four forecast measures for this event allows us to evaluate different aspects of the forecast quality, and thus summarize the performance of the forecasts more adequately (Murphy, 1991).

The average performance of the control forecast for the July and August 2002 Danube floods to observations can be seen in Table VII. Corr should be a familiar measure for most readers, and thus, the response of IQA, PSNR and CoD (which may be more unfamiliar) can be referenced against this measure. In addition, the reader is referred back to the relationships depicted in Figure 2. Each of the measures focuses on slightly different properties of the system, for example, a dataset can have a high correlation, but not necessarily matching in magnitude. The PSNR would pick up on the latter. The IQA is more sensitive to pattern displacement.

The combination of measures shown in Table VII shows that with increasing lead time the performance of the forecast drops. The values are comparably high and suggest a successful forecast. However, it is important to stress that 'skill' is always defined relative to another forecast in any particular application. The measures computed above do not indicate whether a forecast is sufficient enough, for example, to issue a flood warning, as only a coupled modelling approach would allow for such a conclusion (see, for example, Pappenberger *et al.*, 2005). Even a comparison with a long-term climatology does not necessarily lead to successful flood prediction (see discussion in Pappenberger *et al.*, 2007a).

Damrath *et al.* (2007) have pointed out that different users need different verification results. We recommend that studies using other events and variables should follow the process outlined in this article in order to understand which performance measures adequately describe the quality of those particular forecasts. The results presented here should not be taken as general conclusions, but may be used as the basis for similar studies.

7. Conclusions

Many different performance measures have been formulated in meteorology, each of which may be valuable in a different way when analysing forecasts. The use of many different measures in the same analysis is recommended, however the researcher needs a method of getting to grips with a new or unfamiliar method, and to choose between different performance measures and evaluate their suitability for the task at hand. In this article, an approach for comparing and evaluating performance measures, focussing on meteorological forecasts of hydrological extreme events, is described. A six-step process to narrow the number of measures which have to be computed in order to best evaluate such forecasts is postulated.

For step 1, each measure is categorized according to an overview of the various methods and sets them in context to each other. Measures which do not have the desired properties are excluded at this step. Step 2 is a scatterplot analysis. If measures show simple (e.g. linear) relationships between each other when they are plotted against each other, then one of them can be excluded. In step 3, the sensitivity of the measures towards changes in magnitude is evaluated, as this is one important property to be expected from such a measure. Step 4 is an experimental set-up, which tests measures regarding their robustness towards displacement. Step 5 evaluates the behaviour of each performance measure with respect to fields with known geostatistical properties. Step 6 is a comparison with an eyeball verification (where a human interpreter had to classify a large number of images).

This evaluation process has been demonstrated on precipitation predictions by the ECMWF precipitation forecasts for the July and August 2002 floods in the River Danube. The experiment is initially performed with 15 different measures, which range from traditional approaches such as the RMSE to more novel methods such as fuzzy interference. The analysis cuts this down to four measures, which are then used to demonstrate that the forecast performs well on several fronts.

Researchers and practitioners are encouraged to try new or unfamiliar performance measures in their analyses, and to use the six-step approach to 'get to grips' with them.

Appendix: Summary of symbols used

Symbol	Explanation
a	Range of variogram
A	Map A (observed)
A_x	Window used for map A

Appendix (*Continued*)

Symbol	Explanation
B	Map B (modelled)
B_x	Window used for map B
c	Sum of c_1 and c_0 with a maximum of 1
c_0	Spatially uncorrelated variance
c_1	Spatially correlated variance
C	Constant
Ct	Contrast
$f(a,b)$	Similarity measure
H	Lag (distance) between two locations
i, j	Cell index
I	Ignorance
IMSE	Mean weighted ignorance
l	Luminance
L	Range of pixel values (set to the absolute maximum)
MC	Weighted ME
ME	Local Mean Error
n	Number of cells
N	Number of cells in proximity
N/O	No opinion
$P(z)$	Frequency of value z
R	Pearson correlation
S	One-way similarity
s	Overall similarity
SSIM	Similarity index of luminance, contrast and structure
SSS	Spatial smoothing scalar
St	Structure
$W(d)$:	Distance weight
w_G	Distance weight based on a normalized Gaussian function with a standard deviation of 1.5
σ	Standard deviation
μ	Mean

References

Arnaud P, Bouvier C, Cisneros L, Dominguez R. 2002. Influence of rainfall spatial variability on flood prediction. *Journal of Hydrology* **260**: 216–230.

Bobbin J, Recknagel F. 2001. Knowledge discovery for prediction and explanation of blue- green algal dynamics in lakes by evolutionary algorithms. *Ecological Modelling* **146**: 253–262.

Breiman L, Cutler A, 2004. Random Forests, http://oz.berkeley.edu/users/breiman/RandomForests/ [accessed 03 Dec 07].

Breiman L, Friedman J, Olshen R, Stone C. 1984. *Classification and Regression Trees*. Wadsworth Pub. Co: Belmont, CA.

Briggs WM, Levine RA. 1997. Wavelets and field forecast verification. *Monthly Weather Review* **125**: 1329–1341.

Brocker J, Smith LA. 2007. Scoring probability forecasts: on the importance of being proper. *Weather and Forecasting* **22**(2): 382–388.

Buizza R, Bidlot J-R, Wedi N, Fuentes M, Hamrud M. Holt G, Palmer T, Vitart F. 2006. The new ECMWF Variable Resolution Ensemble Prediction System (VAREPS), ECMWF Technical Memorandum 500.

Casati B. 2004. *New approaches for the verification of spatial precipitation*. PhD Thesis. University of Reading: Reading.

Damrath U, Brown B, Nurmi P. 2007. Different types of verification results required by different users. In *Third International Verification Methods Workshop*, Reading.

Demeritt D, Cloke H, Pappenberger F, Thielen J, Bartholmes J, Ramos M-H. 2007. Ensemble predictions and perceptions of risk, uncertainty, and error in flood forecasting. *Environmental Hazards* **7**(2): 115–127.

Deutsch C, Journel A. 1998. *GSLIB: Geostatistical Software Library and User's Guide*. Oxford University Press: New York.

Dietterich TG. 2000. An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization. *Machine Learning* **40**: 139–157.

Ebert EE, McBride JL. 2000. Verification of precipitation in weather systems: determination of systematic errors. *Journal of Hydrology* **239**: 179–202.

EM-DAT. 2007. The OFDA/CRED International Disaster Database, www.em-dat.net, Université Catholique de Louvain, Brussels.

Fairhurst AM, Lettington AH. 2000. The effect of visual perception on the required performance of imaging systems. *Journal of Modern Optics* **47**: 1435–1446.

Freund Y. 2001. An adaptive version of the boost by majority algorithm. *Machine Learning* **43**: 293–318.

Gandin LS, Murphy AH. 1992. Equitable skill scores for categorical forecasts. *Monthly Weather Review* **120**: 361–370.

Ghelli A, Lalaurette A. 2000. Verifying precipitation forecasts using up-scaled observations. *ECMWF Newsletter* **87**: 9–17.

Gober M, Wilson CA, Milton SF, Stephenson DB. 2004. Fairplay in the verification of operational quantitative precipitation forecasts. *Journal of Hydrology* **288**: 225–236.

Grieb TM, Hudson RJM, Shang N, Spear RC, Gherini SA, Goldstein RA. 1999. Examination of model uncertainty and parameter interaction in a global carbon cycling model (GLOCO). *Environment International* **25**: 787–803.

Hagen A. 2003. Fuzzy set approach to assessing similarity of categorical maps. *International Journal of Geographical Information Science* **17**: 235–249.

Hagen-Zanker A. 2006. *Comparing Continuous Valued Raster Data: A Cross Disciplinary Literature Scan Maastricht*. Research Institute for Knowledge Systems: Maastricht.

Hagen-Zanker A, Engelen G, Hurkens J, Vanhout R, Uljee I. 2006. *Map Comparison Kit 3: User Manual*. Research Institute for Knowledge Systems: Maastricht.

Ho TK. 2002. A data complexity analysis of comparative advantages of decision forest constructors. *Pattern Analysis and Applications* **5**: 102–112.

Hoffman RN, Liu JF, Grassotti C. 1995. Distortion representation of forecast errors. *Monthly Weather Review* **123**: 2758–2770.

Jewson S. 2004. Probabilistic forecasting of temperature: comments on the Bayesian Model Averaging Approach. http://www.citebase.org/abstract?id=oai:arXiv.org:physics/0409127.

Jolliffe IT. 2007. Playing the score – exploring beyond the hedge. In *Third International Verification Methods Workshop*, Reading.

Joliffe IT, Stephenson DB. 2003. *Forecast verification: a practitioner's guide in atmospheric science*. John Wiley & Sons, Ltd.,: Chichester; 240pp.

Kelly KS, Krzysztofowicz R. 2000. Precipitation uncertainty processor for probabilistic river stage forecasting. *Water Resources Research* **36**: 2643–2653.

Krzysztofowicz R. 2002. Bayesian system for probabilistic river stage forecasting. *Journal of Hydrology* **268**(1–4): 16–40.

Lee SI. 2001. Developing a bivariate spatial association measure: an integration of Pearsons's r and Moran's I. *Journal of Geophysical Systems* **3**: 369–385.

Mariani S, Casaioli M, Accadia C, Llasat MC, Pasi F, Davolio S, Elementi M, Ficca G, Romero R. 2005. A limited area model intercomparison on the "Montserrat-2000" flash-flood event using statistical and deterministic methods. *Natural Hazards and Earth System Sciences* **5**: 565–581.

Mason SJ. 2007. Do high skill scores mean a good forecast. In *Third International Verification Methods Workshop*, Reading.

Michaelides K, Wilson MD. 2007. Uncertainty in predicted runoff due to patterns of spatially variable infiltration. *Water Resources Research* **45**: 1–14.

Murphy AH. 1991. Forecast verification: its complexity and dimensionality. *Monthly Weather Review* **119**: 1590–1601.

Murphy AH. 1996. General decompositions of MSE-based skill scores: measures of some basic aspects of forecast quality. *Monthly Weather Review* **124**: 2353–2369.

Murphy AH, Daan H. 1985. *Forecast evaluation. Probability, Statistics, and Decision Making in the Atmospheric Sciences*. Murphy AH, Katz RW (Eds.). Westview Press: Boulder, CO; 379–437.

Murphy AH, Winkler RL. 1987. A general framework for forecast verification. *Monthly Weather Review* **115**: 1330–1338.

Obled C, Wendling J, Beven K. 1994. The sensitivity of hydrological models to spatial rainfall patterns – an evaluation using observed data. *Journal of Hydrology* **159**: 305–333.

Olzak LA, Wickens TD. 1999. Paradigm shifts: new techniques to answer new questions. *Perception* **28**: 1509–1531.

Pappenberger F, Beven K. 2004. Functional classification and evaluation of hydrographs based on multicomponent mapping. *International Journal of River Basin Management* **2**: 89–100.

Pappenberger F, Iorgulescu I, Beven KJ. 2006a. Sensitivity analysis based on regional splits (SARS – RT). *Environmental Modelling & Software* **21**: 976–990.

Pappenberger F, Beven KJ, Frodsham K, Romanovicz R, Matgen P. 2006b. Fuzzy set approach to calibrating distributed flood inundation models using remote sensing observations. *Hydrology and Earth System Sciences – Discussions*.

Pappenberger F, Buizza R, Scipal K. 2007a. Hydrological perspective on meterological verification. *Atmospheric Science Letters* in press.

Pappenberger F, Beven K, Frodsham K, Romanowicz R, Matgen P. 2007b. Grasping the unavoidable subjectivity in calibration of flood inundation models: a vulnerability weighted approach. *Journal of Hydrology* **333**: 275–287.

Pappenberger F, Beven KJ, Hunter N, Gouweleeuw B, Bates P, de Roo A, Thielen J. 2005. Cascading model uncertainty from medium range weather forecasts (10 days) through a rainfall-runoff model to flood inundation predictions within the European Flood Forecasting System (EFFS). *Hydrology and Earth System Science* **9**: 381–393.

Paulat M, Frei C, Hagen M, Wernli H. 2007. SAL – a novel error measure for the verification of precipitation forecasts. *Third International Workshop on Verification Methods*, Reading, UK.

Polanyi M. 1967. *The Tacit Dimension*. Anchor Books, Doubleday & Co.: Garden City, NY.

Roulston MS, Smith LA. 2002. Evaluating probabilistic forecasts using information theory. *Monthly Weather Review* **130**(6): 1653–1660.

Shnayderman A, Gusev A, Eskicioglu AM. 2006. An SVD-based grayscale image quality measure for local and global assessment. *IEEE Transactions on Image Processing* **15**: 422–429.

SwissRe. 2003. Torrential rains cause major flooding across Europe, http://www.swissre.com/internet/pwswpspr.nsf/fmBookMarkFrame-Set?ReadForm&BM=./vwAllbyIDKeyLu/ulur-5cyj23?Open-Document.

Tan P-N, Steinbach M, Kumar V. 2006. *Introduction to Data Mining*. Pearson Education: New York.

Tompa D, Morton J, Jernigan E. 2000. Perceptually based image comparison. In *Proceedings of the International Conference on Image Processing*, Vancouver, BC, 1; 489–492.

van der Waerdens BL. 1953. Order tests for two-sample problem and their power. *Indagationes Mathematicae* **15**: 303–316.

Venugopal V, Basu S, Foufoula-Georgiou E. 2005. A new metric for comparing precipitation patterns with an application to ensemble forecasts. *Journal of Geophysical Research-Atmospheres* **110**: D08111. DOI: 10.1029/2004JD005395.

Wang Z, Bovik AC, Sheikh HR, Simoncelli EP. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing* **13**: 600–612.

Weisheimer A, Smith LA, Judd K. 2005. A new view of seasonal forecast skill: bounding boxes from the DEMETER ensemble forecasts. *Tellus Series A-Dynamic Meteorology and Oceanography* **57**: 265–279.

Wilson LJ, Gneiting T. 2007. Another look at proper scoring rules. *Journal of the American Statistical Association* **5**: 1–20.

Woodhead SPB. 2007. Bayesian calibration of flood inundation simulators using an observation of flood extent, PhD, University of Bristol, Bristol, 219.

Zang LJ, Gove JH. 2005. Spatial assessment of model errors from four regression techniques. *Forest Science* **51**: 334–346.