

# Hydrological aspects of meteorological verification

F. Pappenberger,\* K. Scipal and R. Buizza  
*European Centre for Medium-Range Weather Forecasts, Reading, RG1 9AX, UK*

\*Correspondence to:  
F. Pappenberger, European  
Centre for Medium-Range  
Weather Forecasts, Reading,  
RG1 9AX, UK.  
E-mail:  
florian.pappenberger@ecmwf.int

## Abstract

All major weather forecast centres verify meteorological forecasts as a normal part of their operational duties, with verifications usually based on variables and methods of meteorological relevance. These forecasts are then often used to drive hydrological models, and this article demonstrates that a considerable gap exists between current meteorological practice and hydrological needs. This article discusses this gap in terms of the type of variables; the domain and resolution; the importance of choosing appropriate thresholds; and the smoothing and accumulation period. A list of recommendations for a user-focused evaluation is given in the conclusions. We suggest that the meteorological community, and specifically the forecast centres, should consider making these adjustments and producing additional products suitable for hydrological applications. Copyright © 2008 Royal Meteorological Society

**Keywords:** hydrology; verification; flood; Danube; evaluation; uncertainty

Received: 28 September 2007  
Revised: 11 December 2007  
Accepted: 7 January 2008

## 1. Introduction

Meteorological forecasts are regularly verified to make the model outputs meaningful (Uebel, 2003), monitor forecast quality over time and space, compare forecast systems or discover model errors to improve forecast quality (Ghelli and Ebert, 2008). Meteorological forecasts are often used to drive hydrological models, and in these cases, the evaluation of the quality of the meteorological forecast has to be specific to the hydrological application. However, there are strong differences between the epistemic cultures of hydrology and meteorology practise, for example, dissimilar attitudes to risk, uncertainty and error (Demeritt *et al.*, 2007). A Successful verification of meteorological forecasts for hydrological modelling needs requires a profound understanding of these differences.

This article will first discuss the current practice of meteorological verification from a hydrological perspective. Examples of forecast for the Danube catchment for the period from 2002 to 2006 will be used to illustrate (1) the importance of the variables used for verification; (2) the domain and resolution on which verification is performed; (3) the importance of choosing correct thresholds; and (4) the smoothing and accumulation period of each forecast.

This is followed by an introduction of some novel verification methods for meteorological forecasts based on hydrological properties. These methods are then illustrated by the example of the forecast of the July and August 2002 floods in the Danube.

## 2. Current practice

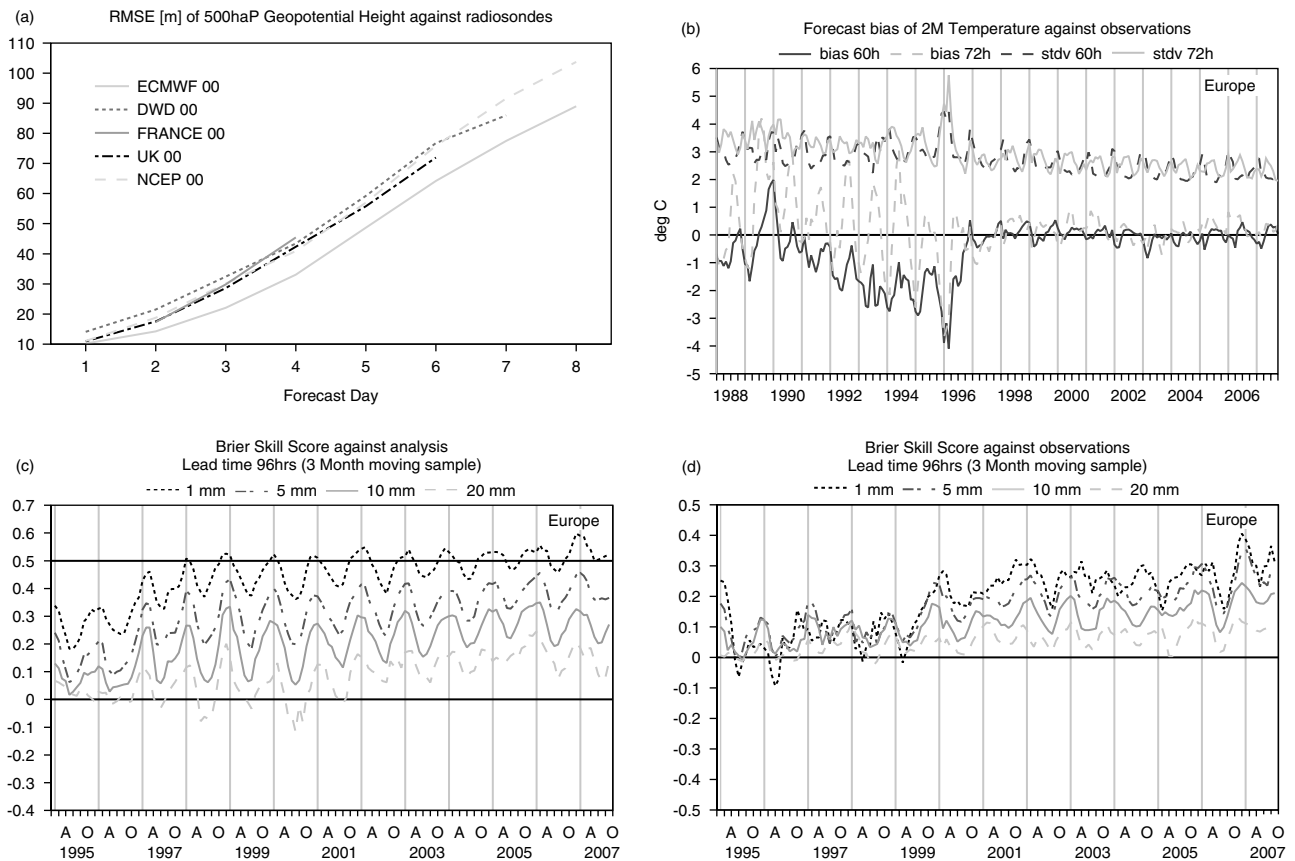
Forecast verification plays an important role in operational weather prediction, and there are many products

that are calculated on a day-to-day basis in order to verify the forecasts. In Figure 1 some typical examples of the verification products of the European Centre for Medium-Range Weather Forecasts (ECMWF) are shown.

The panel (a) in Figure 1 shows the Root Mean Squared Error (RMSE) of different forecast centres of 500 hPa Geopotential Height according to the latest WMO/CBS recommendations. The top right panel (b) shows the bias and standard deviation of the ECMWF-deterministic forecast for 2-m temperature for a 60- and 72-h lead time. The bottom panels (c) and (d) display the Brier Skill Score of the Ensemble Prediction System of ECMWF for four different thresholds of 24 h accumulated precipitation for a 96-h lead time. All panels show verification results for the European Domain. A more detailed explanation of these figures is given on the ECMWF Web pages (<http://www.ecmwf.int/products/forecasts/d/charts/medium/verification/>). These figures will be used to discuss the issues of hydrological aspects of meteorological scoring in what follows.

## 3. Discussion of hydrological aspects of meteorological forecast evaluation

In this section, current practise in meteorological verification will be discussed from a hydrological perspective. To do this effectively, it is important to recognize that hydrologists and meteorologists often use different terminologies to denote similar concepts or practices. For example, in meteorology the word 'verification' is commonly used to define the comparison of a model forecast with some perceived truth to establish qualitative or quantitative information of the forecast



**Figure 1.** Typical forecast verification products of the European Centre for Medium-Range Weather Forecasts. (a) This panel shows the Root Mean Squared Error of different forecast centres of 500 hPa Geopotential Height according to the latest WMO/CBS recommendations. The top right panel (b) shows the bias and standard deviation of the ECMWF-deterministic forecast for 2-m temperature for a 60 and 72-h lead time. The bottom panels (c) and (d) display the Brier Skill Score of the Ensemble Prediction System of ECMWF for four different thresholds of 24 h accumulated precipitation for a 96-h lead time: The bottom left panel (c) illustrates verification against observations, and the bottom right panel (d) illustrates verification against analysis. The results are filtered by a 3-monthly moving average. All panels show verification results for the European Domain.

error and is considered impossible (see discussions by Oreskes *et al.*, 1994; Oreskes, 2000; Beven, 2001a,b, 2006). It is usually replaced by ‘evaluation’ or a similar term in hydrology. In fact, many meteorologists are aware of this issue and also choose to define ‘verification’ as ‘evaluation’ (in its hydrological definition). In this article, the meteorological definitions are used (unless indicated otherwise).

### 3.1. Type of observations

One of the most important considerations when using meteorological forecasts in hydrological models is the use of observations of hydrologically relevant variables in the verification process. Here we regard the nature of the observations that are required. Many verifications in meteorology are performed on a synoptic scale, considering variables such as the 850 hPa temperature or the 500 hPa pressure (see Figure 1), and on this scale weather forecasts have increased in quality by roughly 1 day per decade (Grazzini, 2007). These fields are chosen because they give a useful view of large-scale flow, are essential for the assessment of the weather, and are available for many different models over a long time (Buizza *et al.*, 2005).

These synoptic patterns have correlations (sometimes stronger, sometimes weaker) to hydrologically relevant surface variables. For example, variance in the occurrence of synoptic patterns can be significantly related to variance in snow pack (Romolo *et al.*, 2006); surface temperature can be best described by 500 hPa and 700 hPa (Post *et al.*, 2002); or typical synoptic patterns can cause heavy rainfall (Nishiyama *et al.*, 2007). Large-scale fields can be used to estimate, for example, daily occurrence and amount of precipitation (Harpham and Wilby, 2005; Haylock *et al.*, 2006).

However, hydrological models are usually forced by surface variables, and in many cases the relationship between skill on synoptic scale and skill on the surface is weak (see for example Huth, 2004; Scherrer *et al.*, 2004). Hydrological applications demand both the verification of synoptic scale and surface variables in order for a forecast to be useful. The type of surface variables that need to be evaluated depends on the dominant process in each catchment and is a function of, for example, antecedent conditions and catchment characteristics (see uniqueness of place argument by Beven, 2000). The significance of inputs (and their uncertainties) on runoff estimation remains complex (Freer *et al.*, 2003, 2004) and research can show

contradictory results (Segond, 2006). Moreover, the importance of various variables depends strongly on the hydrological application (for example, the correctness of heavy precipitation will be very important for flood forecasts), and it is thus impossible to derive a general rule about which variables to evaluate. However, we suggest that precipitation and 2-m temperature should be part of the minimum requirements as they are most likely to be influential quantities for hydrological modelling results.

The type of variables against which the forecasts are compared is also important. There is a substantial difference between the use of analysis fields and actual observations. Analysis fields are derived from different sources of observations, which have been merged and corrected through data assimilation, and thus, are strongly dependent also on the weather model and on the data assimilation used to construct the analysis. Analysis fields are often used for convenience as they are already on the 'correct' model grid and cover areas and time without missing data. However, they suffer from errors due to reliability in observations, introduce additional uncertainties and approximations and can produce artificial skill (Simmons, 2001; Casati, 2004). Indeed, the error structure of an analysis field can be very different from observations (especially when the uncertainties introduced by scaling point measurements to grid representations and vice versa is considered). Therefore, they can produce vastly different skill scores. For example, in Figure 1, the two panels, (c) and (d), show that the analysis field produces significantly higher skill scores than the comparison against observations. Moreover, the analysis field has been optimized for the individual meteorological models, and thus, does not 'fit' to any particular hydrological model due to grid size and differences in process representation. One additional problem of using an analysis field as verification is that some variables are not produced by a data assimilation procedure: for example, the ECMWF data assimilation system does not produce an analysis of precipitation, and what is named precipitation analysis is actually a 24-h model forecast.

In many cases, hydrological models are calibrated and optimized with observations and, therefore, an evaluation of quality and skill of a meteorological model used in a hydrological application has to be performed using observations.

### 3.2. Benchmark and skill

It is usual model evaluation practise in meteorology to calculate forecast performance against a benchmark to calculate skill, in the form of climatology or persistence. However, in hydrology, the calculation of climatology in streamflow is not necessarily comparable with a meteorological climatology due to the influence of structural measures and modifications of river-bed through flood and human intervention. An example of a score commonly used in meteorology, which uses

climatology is the Brier Skill Score. Figure 1 shows an example of using the Brier Skill Score in the bottom panels, (c) and (d). The formulation of the Brier Skill Score is given in Jolliffe and Stephenson (2003). The Nash-Sutcliffe efficiency criterion is an example for hydrology (Schaeffli and Gupta, 2007). Schaeffli and Gupta (2007) discuss the importance of the benchmark model being a real test for the model, rather than just climatology (see also Mason, 2004; Bartholmes, 2007). For example, it is common practice to use long-term climatology when computing the Brier Skill Score (see Rodwell, 2005). This is sensible when the performance of the forecast model itself is of interest. However, if the model is coupled to a flood prediction system then the forecast has to be skillful for the dominating hydrological processes (and not necessarily for any long-term trends). In this case, the benchmark would be better derived from a short-term persistency or climatology of similar flood events. This is difficult to achieve in the case of flood forecasting due to the low frequency of extreme events, and the spatial and temporal correlation of flood events (Merz and Blöschl, 2003), therefore, the usage of short-term persistency is the best alternative option.

### 3.3. Analysis based on thresholds

It is also common meteorological practice to analyse forecasts based on thresholds of precipitation. The spatial and temporal dependency of floods led Merz and Blöschl (2003) to compile a typology of regional floods. This typology is partially based on precipitation thresholds and the time above those thresholds (see also Blazkova and Beven, 2002; Kusumastuti *et al.*, 2006; Struthers and Sivapalan, 2006). Standard meteorological evaluation thresholds are, for example, exceedance of 1 mm/24 h, 5 mm/24 h, 10 mm/24 h and 20 mm/24 h (see Figure 1, bottom two panels, (c) and (d)). However, from a hydrological view-point, these thresholds are rather arbitrary as they are not linked to any dominant hydrological processes. Any analysis for hydrological applications with thresholds has to be based on physical reasoning.

### 3.4. Multiple suitable performance measures

It is well known that use of a single evaluation method can cause misleading interpretation of forecast ability (Murphy and Winkler, 1987; Murphy, 1993, 1991; Casati, 2004). The use of multiple methods is advised, and frameworks for choosing methodologies are promoted (Murphy, 1996). A hydrologically specific methodology for choosing performance measures is demonstrated by Cloke and Pappenberger (2008) in which the properties of various performance measures are evaluated for application-specific suitability by a six-step framework. Fortunately, the meteorological as well as the hydrological community already embrace the usage of multiple measures as good practice (van Griensven and Bauwens, 2003; Vrugt *et al.*,

2003; Jolliffe, 2007; Schumann *et al.*, 2007; Pappenberger *et al.*, 2008). The performance measures have to be suitable for the event and application in question. For example, the Brier Skill Score may be unsuitable for low-frequency events (Casati, 2004).

### 3.5. Size of verification area

The size of the area used for verification strongly influences the skill score calculated. All panels in Figure 1 are based on the verification area of Europe, which allows the capture of large-scale synoptic patterns and reflects the model resolution. The individual river's catchment area is a hydrologically more relevant scale. It should be noted that skill over a catchment area can be lower than skill over Europe. However catchment based skill scores reflect more accurately the skill that local users experience. This is because of the smaller climatic range over the catchment area. The size of the catchment area is also important because larger catchments usually have a larger density of data, and thus, more of a smoothing effect. The resolution of the weather forecast model is often too coarse to resolve the small catchment scale.

As an example, Figure 2 shows the Brier Skill Score of the 'probabilistic prediction of precipitation in excess of 1 mm/24 h' at a 4-day lead time averaged over two spatial scales: the Danube catchment (817,000 km<sup>2</sup>) and the whole of Europe (Lat: 35°–70°, Long: 15°–35°).

The figure clearly shows that forecasts in the Danube area are on average less skillful, have a higher variance and significantly larger peaks and troughs. The above analysis seems to suggest that larger verification areas are sensible. However, just the opposite

is true if we are to analyse forecasts in a hydrologically relevant way. We must know the skill of our forecasts on a hydrological relevant scale – which is the catchment or sub-catchment scale.

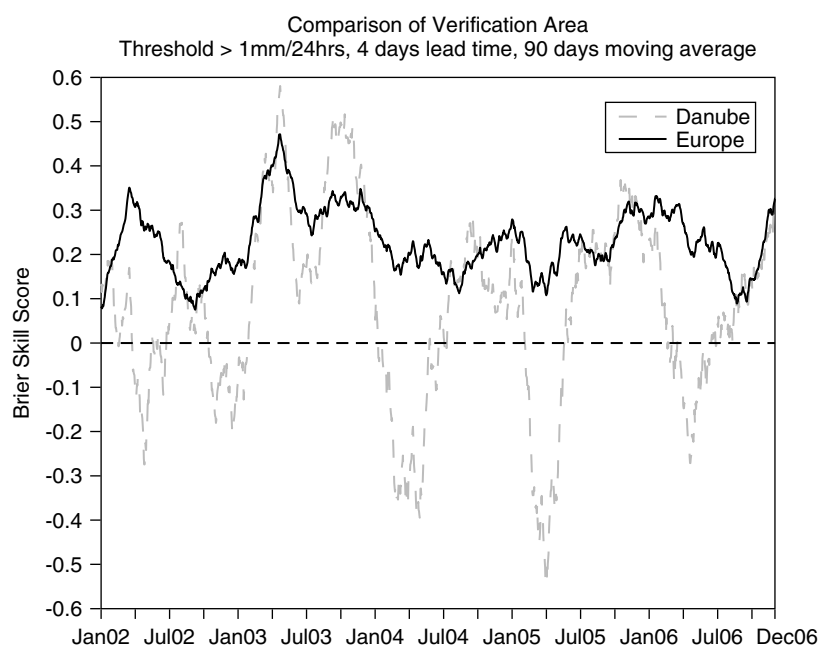
### 3.6. Averaging, smoothing and accumulation

Averaging, smoothing and accumulation are used in several different ways when analysing meteorological forecasts. In this section, we discuss how these analysis processes must remain hydrologically relevant.

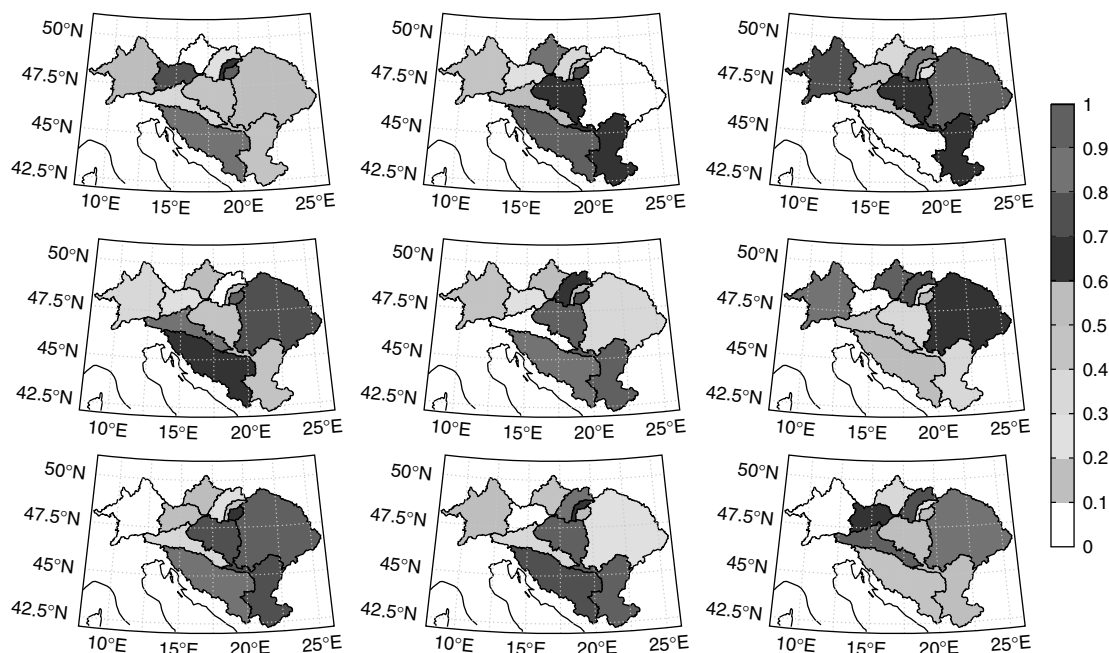
#### 3.6.1. Spatial averaging and interpolation

The effect of spatial smoothing is demonstrated in Section 3.5 and Figure 2. The larger a catchment or verification area, the more climatological regions are covered and the more data are available to build reliable averages. Kann and Haiden (2005) observed a reduction in Mean Absolute Error with increasing verification area. Thus, as noted above, catchment size has an effect on model performance. The total and main sensitivity of discharge predictions will depend on the spatial co-variance structure of the variables and the non-linear transformation through the hydrological model. For example, in the case of precipitation, the sensitivity of the river flow hydrograph towards the uncertainty in precipitation on catchment response decreases with catchment scale (Rodriguez-Iturbe and Mejia, 1974; Sivapalan and Bloschl, 1998; Woods and Sivapalan, 1999; Segond, 2006).

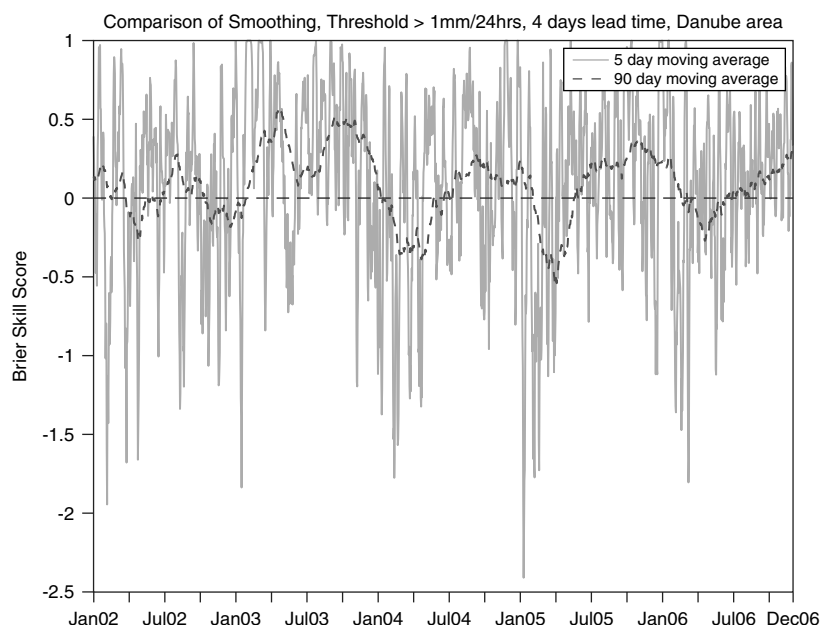
In order to be used in the hydrological model, the forecast and observations can be aggregated or interpolated in three different forms: grid; catchment average (see, for example, Figure 3); or point data. The choice of one particular representation may depend on



**Figure 2.** Comparison of the verification areas of Europe and the Danube by a Brier Skill Score of precipitation with a threshold of > 1 mm/24 h and 4-day lead time.



**Figure 3.** Stamp maps of mean precipitation forecasts of the ECMWF EPS system of major Danube catchments on 4 August 2002, with a lead time of 1 day.



**Figure 4.** Comparison of different time averaging windows for the verification areas of the Danube catchment by a Brier Skill Score of precipitation with a threshold of > 1 mm/24 h and 4-day lead time.

the type of hydrological model used (e.g. distributed vs lumped) and the catchment characteristics (Naden, 1992; Obled *et al.*, 1994; Andreassian *et al.*, 2004; Smith *et al.*, 2004; Dodov and Foufoula-Georgiou, 2005).

**3.6.2. Time averaging**

Averaging of model performance by using a time averaging window is commonly used in meteorology to show long-term trends. Panels (c) and (d) in Figures 1 and 2 are displayed with a 90-day filter. In Figure 4 the difference between a 90-day moving window and

a 5-day moving window for the Danube area is shown. The differences can be seen to be large, and the 5-day moving average can be significantly above or below the 90-day average for a significant amount of time (see, for example, October 2004). In a hydrological model, short-term variations are of interest as they directly influence the hydrological processes. Kann and Haiden (2005) have shown that an increased time averaging window decreases a Mean Absolute Error due to the auto-correlation in the observations. Averaging over a time scale that is significantly larger than, for example, the hydrological response time of a

catchment does not allow any conclusions on the quality of the forecast system for a hydrological application. Indeed use of long-term trends can be misleading for any short-term use of forecasts such as a single storm event.

### 3.6.3. Accumulation period and time-step

The accumulation period and time-step used in a forecast analysis may strongly affect the hydrological processes operating in the hydrological model.

The smoothing with a time averaging window in Figure 4 is over a forecast period for precipitation accumulated in 24 h for this particular forecast step. However, such an evaluation is in contrast to the usage of a precipitation forecast in hydrological models, where it is used as accumulation over the forecast lead time. For a 4-day forecast, the error accumulation from 0 to 96 h is important for a hydrological model, as the error is non-linearly transformed by the hydrological model, and different dominant processes may be triggered. For example, any large error at the beginning of a forecast may trigger the hydrological process of saturation excess and overland flow at a lead time of 96 h, whereas the forecast on day 4 alone (ignoring antecedent conditions) may lead to precipitation to infiltrate rather than flow overland. The error increases with forecast range due to the auto-correlation of the variable (see e.g. Kann and Haiden, 2005). This is less of an issue for variables for which the correlation to the previous day is negligible such as, potential evapotranspiration, but extremely important for a variable such as, precipitation.

It is also important to consider the time step of a meteorological model from a hydrological perspective. For example, most hydrological processes of interest operate on a short time scale of, for example, 1 and 24 h, when accumulated precipitation is too long a timescale to be useful for evaluating these. At the very least, evaluation should be performed on the time scale of the meteorological forecast. In the case of ECMWF EPS forecasts, this is currently 3 h for the first 72 forecast hours and 6 h thereafter, although achieving of smaller time steps would be preferable.

## 4. Methods for a hydrologically focused evaluation of meteorological forecasts

The previous section discussed a hydrological viewpoint of current practice in meteorological verification analysis. In this section, we propose alternative approaches, which support a hydrologically oriented verification of meteorological forecasts. We use the specific example of the July and August 2002 floods in the Danube to illustrate how these novel evaluation measures can be used to evaluate a meteorological forecast.

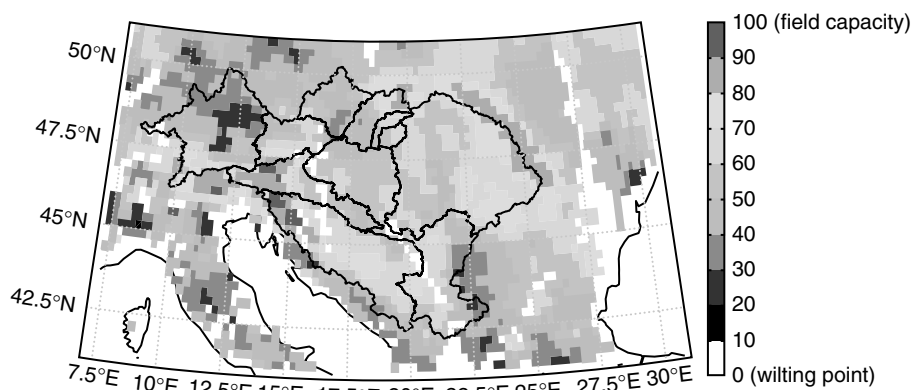
### 4.1. Coupled model evaluation

The coupling of hydrological models and meteorological models is a valuable method to evaluate application-specific performance of meteorological predictions (for an example of such an evaluation, see Balsamo *et al.*, 2008). Ahrens and Jaun (2007) argue that such an integrated approach overcomes scale issues and allows evaluation of high-resolution precipitation forecasts by utilizing, for example, discharge predictions and discharge observations (rather than precipitation forecasts and observations). The methodology respects the importance of dominant hydrological processes (see discussion on skill above, in Section 3.2) and the non-linear error transformation by the hydrological model (for an example of this, see Gurtz *et al.*, 2003; Verbunt *et al.*, 2006). However, it neglects the fact that many hydrological models are calibrated on observations or observational fields which have an error structure incompatible with that of the forecasts of the meteorological model (see discussion on the use of analysis fields above, Section 3.1). Moreover, the uncertainty of the hydrological model itself has to be acknowledged in this model cascade (Krzysztofowicz and Herr, 2001; Krzysztofowicz, 2002a,b; Gourley and Vieux, 2005; Pappenberger *et al.*, 2005).

The European Flood Forecasting System (de Roo *et al.*, 2003) successfully provided 7-day forecasts for the July and August 2002 floods in the Danube. This suggests that the meteorological forecast has been adequate in timing and precipitation quantities for this forecast period.

### 4.2. Process-/vulnerability-weighted evaluation

Coupling models is time and skill intensive, and is therefore, not always a suitable method. We suggest a novel process-/vulnerability-weighted approach as an alternative. It assumes that physical performance is not the only quantity, which should be used to calibrate and evaluate models. Pappenberger *et al.* (2006) proposed a model evaluation and calibration which included a risk component (population vulnerability to flooding). In this article, soil moisture has been selected to define vulnerability to flooding. The basic idea of the method is that during the verification process more weight is given to vulnerable areas, i.e. where flooding is more likely to occur because the water content in the soil is already high. The simplest way to integrate vulnerability is by computing a weighted mean based on the soil moisture distribution [which is done in this article with the (RMSE)]. We assume that saturation resulting in excess overland flow is the dominant process of flood generation in the area. This is, of course, a rather flawed assumption, as catchment hydrology is more complex, but we use this for illustration purposes. Soil moisture was derived from scatterometer observations from the ERS-2 satellite mission (Wagner *et al.*, 1999; Ceballos *et al.*, 2005). Figure 5 shows an example of the distribution of soil moisture in sub-catchments of the

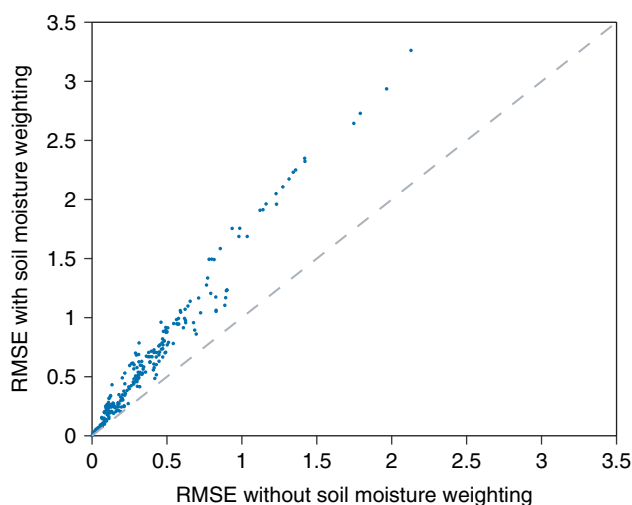


**Figure 5.** Soil moisture of the Danube catchment for 20 April 2004 derived from the ERS-2 satellite mission on a 0.25° latitudinal and longitudinal grid. The soil-moisture values are used as weighting matrix for the computation of the performance measure.

Danube area. Values range from wilting level (0) to field capacity (100) on a 0.25° grid. The weighting has been carried out according to these soil-moisture values, which have been normalized. Areas which are left blank are either covered by snow or have no available data and are excluded from the analysis.

In Figure 6, this effect is demonstrated for the RMSE measure for verification of the mean ECMWF EPS precipitation forecast of the July and August 2002 flooding in the Danube catchment.

Soil-moisture weighting leads to a magnitude of differences in performance measures. It is, for example, possible to get an un-weighted RMSE of 1 (see abscissa of Figure 6) and a weighted RMSE of 1.5 (see ordinate of Figure 6). The RMSE with soil-moisture weighting leads to larger errors than the scheme without weighting, which is an indication that the model produces larger errors in areas with higher soil moisture. This demonstrates that the usage of hydrologically relevant scores can reveal additional properties of a numerical prediction weather system.



**Figure 6.** Comparison of the RMSE performance measure and the RMSE with soil-moisture weighting performance measure for raster-based evaluation.

### 4.3. Hydrologically focused performance measures (feature-based verification)

The previous two methodologies relied on additional information in order to perform a hydrologically driven evaluation of a meteorological forecast. However, it is also possible to design performance measures, which are even more closely linked to hydrologically relevant properties. Feature-based (or entity-based) verification allows to focus on properties which are of hydrological relevance (for a discussion see Ebert and McBride, 2000). Paulat *et al.* (2007) have presented a measure, which includes proportional coverage and centre of gravity.

#### 4.3.1. Coverage

The percentage coverage of a forecast variable such as precipitation can be computed for each catchment and standard probabilistic or deterministic performance measures applied. It is an excellent measure to evaluate a key property of a predicted field, which is relevant to hydrological processes. In Table I the ratio between the mean forecasted coverage and the observed coverage is plotted for the Lower Morava catchment (a sub-catchment of the Danube). The size of the forecast cell plays an important role, as a large forecast-cell to catchment-size ratio will impact on this analysis. The table indicates that on average a larger precipitation cover has been forecasted than observed for this event.

#### 4.3.2. Centre of gravity

The centre of gravity of a system is a specific point at which, for many purposes, the system's mass behaves as if it were concentrated. The centre of mass is

**Table I.** Ratio between mean forecasted and observed precipitation coverage for the Lower Morava catchment

Lead time	42 h	66 h	90 h	114 h	138 h	162 h
Ratio	2.6	2.4	2.5	2.1	2.1	2.7

a function only of the positions and masses of the particles that comprise the system. A score could be computed as:

$$CoG = \frac{\sqrt{\text{dist}(c_f, loc_{outlet})}}{\sqrt{\text{dist}(c_{obs}, loc_{outlet})}} \quad (1)$$

where,

*CoG*: centre of gravity measure  
*c*: centre of gravity  
*dist*: distance after Vincenty (1975)  
*f*: forecast  
*loc<sub>outlet</sub>*: location of catchment outlet  
*obs*: observed

In Figure 7, a rose diagram of the azimuth in degrees of observed to modelled CoG is displayed (on the left). And on the right is a histogram of CoG distance measure for the Lower Morava catchment. This example is again for the July and August 2002 Danube floods. The rose diagram indicates that in the majority of the cases the modelled precipitation forecast was too far in the north-northeast. However, the distance fraction achieves a large proportion of values around 1, which indicates that the centre of the predicted and observed storm have similar distances to the catchment outlet. Thus, the shifts must have been very small.

## 5. Conclusions

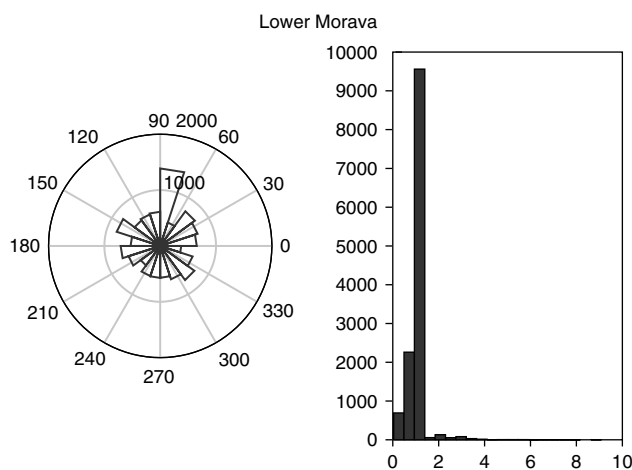
This article discusses hydrological aspects of meteorological verification of meteorological forecast models. Verification of forecasts has a long tradition in meteorology, and scores or skills are published regularly by all major forecast centres. However, in order to be valuable to hydrologists, such evaluations must consider the needs of the end-users and the constraints under which they work. We propose the following

recommendations for a comprehensive hydrologically focused evaluation of meteorological models:

1. Any verification should always include variables of interest to hydrology specific to the catchment characteristics and antecedent conditions. The absolute minimum requirements should include precipitation and temperature.
2. Comparison should be performed against actual observations in addition to analysis fields.
3. The benchmark model for computing skill should be short-term persistency or a probability distribution based on the dominant hydrological processes relevant for a particular hydrological application.
4. The choice of thresholds has to reflect the threshold behaviour of the catchments.
5. The verification area should be catchment-based.
6. Spatial averaging and interpolation of the evaluation process needs to match the use of the data in the hydrological model.
7. Smoothing performance measures or physical quantities with large time averaging windows (for example, 90 days) are useful for seeing long-term trends, but unacceptable when trying to understand performance on a hydrologically relevant time scale. Forecasts should be evaluated on a hydrologically sensible time scale, which is usually much lower than the standard 24-h time steps.
8. Antecedent conditions are important. Forecasts are used continuously, meaning that a forecast for day 4 has to be evaluated on, for example, accumulated precipitation until day 4. The memory of the catchment has to be respected in such an accumulated verification.

In order to enhance the usability of meteorological forecast for hydrological applications, additional methods to verification analysis are advocated in this article: (1) analysis of coupled modelling systems; (2) vulnerability-weighted (for example, soil moisture) performance analysis; (3) hydrologically focused performance measures such as, coverage and centre of gravity. These measures have been applied on forecasts of the July and August 2002 Danube floods. Results have shown that (1) a coupled forecast system indicates a successful meteorological forecast, (2) the forecast error predominantly occurred in areas with high soil moisture, and (3) the precipitation coverage was over-predicted, and the centre of gravity of rainfall field was slightly wrongly predicted as far to the northeast.

We suggest that the meteorological community, and specifically the forecast centres, should consider making these adjustments and producing additional products suitable for hydrological applications. Also, other user communities could be included with other specific product requirements. The end-user focus of some meteorological forecast services is encouraging, for example, demonstrated by their involvement in many hydrological/meteorological cross-cutting initiatives such as, HEPEx (Schaafe *et al.*, 2006).



**Figure 7.** On the left, a rose diagram of the azimuth in degrees of observed to modelled CoG. On the right, a histogram of CoG distance measure for the Lower Morava catchment.

## Acknowledgements

This work is funded by the EC PREVIEW (FP6 - Work Package: Plain Floods) program (<http://www.preview-risk.com>). We thank Hannah Cloke (King's College, London), Anna Ghelli (ECMWF), Antje Weisheimer (ECMWF) and two anonymous reviewers for their insightful reviews which improved the manuscript considerably.

## References

- Ahrens B, Jaun S. 2007. On evaluation of ensemble precipitation forecasts with observation-based ensembles. *Advances in Geosciences* **10**: 139–144.
- Andreassian V, Oddos A, Michel C, Ancil F, Perrin C, Loumagne C. 2004. Impact of spatial aggregation of inputs and parameters on the efficiency of rainfall-runoff models: a theoretical study using chimera watersheds. *Water Resources Research* **40**: DOI:10.1029/2003WR002854.
- Balsamo G, Beljaars ACM, Viterbo P, van der Hurk M, Hirschi A, Betts AK. 2008. Terrestrial water storage in NWP: Are we gaining a better understanding. In *At International Workshop on Catchment-scale Hydrological Modelling and Data Assimilation*, Melbourne.
- Bartholmes JC, Thielen J, Ramos MH, Gentilini S. 2008. The European Flood Alert System EFAS Part 2: Statistical skill assessment of probabilistic and deterministic operational forecasts. *Hydrol. Earth Syst. Sci. Discuss.* **5**: 289–322.
- Beven KJ. 2000. Uniqueness of place and process representations in hydrological modelling. *Hydrology and Earth System Sciences* **4**: 203.
- Beven KJ. 2001a. On explanatory depth and predictive power. *Hydrological Processes* **15**: 3069–3072.
- Beven KJ. 2001b. On hypothesis testing in hydrology. *Hydrological Processes* **15**: 1655–1657.
- Beven KJ. 2006. A manifesto for the equifinality thesis. *Journal of Hydrology* **320**: 18–36.
- Blazkova S, Beven K. 2002. Flood frequency estimation by continuous simulation for a catchment treated as ungauged (with uncertainty). *Water Resources Research* **38**(8): DOI:10.1029/2001WR000500.
- Buizza R, Houtekamer PL, Toth Z, Pellerin G, Wei MZ, Zhu YJ. 2005. A comparison of the ECMWF, MSC, and NCEP global ensemble prediction systems. *Monthly Weather Review* **133**: 1076–1097.
- Casati B. 2004. New approaches for the verification of spatial precipitation, PhD thesis, University of Reading, Reading.
- Ceballos A, Scipal K, Wagner W, Martinez-Fernandez J. 2005. Validation of ERS scatterometer-derived soil moisture data in the central part of the Duero Basin, Spain. *Hydrological Processes* **19**: 1549–1566.
- Cloke HL, Pappenberger F. 2008. Evaluating forecasts of extreme events for hydrological applications: an approach for screening unfamiliar performance measures. *Meteorological Applications* **15**: 181–190.
- 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**: 115–127.
- de Roo A, Gouweleeuw B, Thielen J, Bartholmes J, Bongioannini-Cerlini P, Todini E, Bates P, Horritt M, Hunter N, Beven KJ, Pappenberger F, Heise E, Rivin G, Hills M, Hollingsworth A, Holst B, Kwadijk J, Reggiani P, van Dijk M, Sattler K, Sprokkereef E. 2003. Development of a European flood forecasting system. *International Journal of River Basin Management* **1**: 49–59.
- Dodov B, Foufoula-Georgiou E. 2005. Fluvial processes and stream-flow variability: interplay in the scale-frequency continuum and implications for scaling. *Water Resources Research* **41**: DOI:10.1029/2004WR003408.
- Ebert EE, McBride JL. 2000. Verification of precipitation in weather systems: determination of systematic errors. *Journal of Hydrology* **239**: 179–202.
- Freer J, Beven KJ, Peters N. 2003. In *Calibration of Watershed Models*, Duan QY, Gupta H, Sorooshian S, Rousseau A, Turcotte R (eds). American Geophysical Union: Washington, DC: 69–88.
- Freer JE, McMillan H, McDonnell JJ, Beven KJ. 2004. Constraining dynamic TOPMODEL responses for imprecise water table information using fuzzy rule based performance measures. *Journal of Hydrology* **291**: 254–277.
- Ghelli A, Ebert EE. 2008. Editorial: Special issue on forecast verification. *Meteorological Applications* **15**: 1.
- Gourley JJ, Vieux BE. 2005. A method for evaluating the accuracy of quantitative precipitation estimates from a hydrologic modeling perspective. *Journal of Hydrometeorology* **6**: 115–133.
- Grazzini F. 2007. Predictability of a large-scale flow conducive to extreme precipitation over the western Alps. *Meteorology and Atmospheric Physics* **95**: 123–138.
- Gurtz J, Zappa M, Jasper K, Lang H, Verbunt M, Badoux A, Vitvar T. 2003. A comparative study in modelling runoff and its components in two mountainous catchments. *Hydrological Processes* **17**: 297–311.
- Harpham C, Wilby RL. 2005. Multi-site downscaling of heavy daily precipitation occurrence and amounts. *Journal of Hydrology* **312**: 235–255.
- Haylock MR, Cawley GC, Harpham C, Wilby RL, Goodess CM. 2006. Downscaling heavy precipitation over the United Kingdom: a comparison of dynamical and statistical methods and their future scenarios. *International Journal of Climatology* **26**: 1397–1415.
- Huth R. 2004. Sensitivity of local daily temperature change estimates to the selection of downscaling models and predictors. *Journal of Climate* **17**: 640–652.
- Jolliffe IT. 2007. Playing the score – exploring beyond the hedge. In *Third International Verification Methods Workshop*, Reading.
- Jolliffe IT, Stephenson DB. 2003. *Forecast Verification: A Practitioner's Guide in Atmospheric Science*. John Wiley and Sons: Chichester.
- Kann A, Haiden T. 2005. The August 2002 flood in Austria: sensitivity of precipitation forecast skill to areal and temporal averaging. *Meteorologische Zeitschrift* **14**: 369–377.
- Krzysztofowicz R. 2002a. Bayesian system for probabilistic river stage forecasting. *Journal of Hydrology* **268**: 16–40.
- Krzysztofowicz R. 2002b. Probabilistic flood forecast: bounds and approximations. *Journal of Hydrology* **268**: 41–55.
- Krzysztofowicz R, Herr HD. 2001. Hydrologic uncertainty processor for probabilistic river stage forecasting: precipitation-dependent model. *Journal of Hydrology* **249**: 46–68.
- Kusumastuti DI, Struthers I, Sivapalan M, Reynolds DA. 2006. Threshold effects in catchment storm response and the occurrence and magnitude of flood events: implications for flood frequency. *Hydrology and Earth System Sciences Discussions* **3**: 3239–3277.
- Mason SJ. 2004. On using “climatology” as a reference strategy in the Brier and ranked probability skill scores. *Monthly Weather Review* **132**: 1891–1895.
- Merz R, Blöschl G. 2003. A process typology of regional floods. *Water Resources Research* **39**: DOI: 10.1029/2002WR001952.
- Murphy AH. 1991. Forecast verification – its complexity and dimensionality. *Monthly Weather Review* **119**: 1590–1601.
- Murphy AH. 1993. What is a good forecast – an essay on the nature of goodness in weather forecasting. *Weather and Forecasting* **8**: 281–293.
- Murphy AH. 1996. The finley affair: a signal event in the history of forecast verification. *Weather and Forecasting* **11**: 3–20.
- Murphy AH, Winkler RL. 1987. A general framework for forecast verification. *Monthly Weather Review* **115**: 1330–1338.
- Naden PS. 1992. Spatial variability in flood estimation for large catchments – the exploitation of channel network structure. *Hydrological Sciences Journal-Journal Des Sciences Hydrologiques* **37**: 53–71.
- Nishiyama K, Endo S, Jinno K, Uvo CB, Olsson J, Berndtsson R. 2007. Identification of typical synoptic patterns causing heavy rainfall in the rainy season in Japan by a Self-Organizing Map. *Atmospheric Research* **83**: 185–200.
- Obléd 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.

- Oreskes N. 2000. In *The Earth Around Us*, Schneider van J (ed.). W.H. Freeman and Company: New York.
- Oreskes N, Shrader-Frechette K, Belitz K. 1994. Verification, validation, and confirmation of numerical-models in the earth-sciences. *Science* **263**: 641–646.
- Pappenberger F, Beven KJ, Ratto M, Matgen P. 2008. Multi-method global sensitivity analysis of flood inundation models. *Advances in Water Resources* **31**: 1–14.
- Pappenberger F, Beven KJ, Frodsham K, Romanovicz R, Matgen P. 2006. 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 Sciences* **9**: 381–393.
- Paulat M, Frei C, Hagen M, Wernli H. 2007. SAL – a novel error measure for the verification of precipitation forecasts. In *Third International Workshop on Verification Methods*, Reading.
- Post P, Truija V, Tuulik J. 2002. Circulation weather types and their influence on temperature and precipitation in Estonia. *Boreal Environment Research* **7**: 281–289.
- Rodriguez-Iturbe I, Mejia JM. 1974. The design of rainfall network in time and space. *Water Resources Research* **10**: 713–728.
- Rodwell MJ. 2005. Comparing and combining deterministic and ensemble forecasts: how to predict rainfall occurrence better. *ECMWF Newsletter* **1006**: 1–6.
- Romolo L, Prowse TD, Blair D, Bonsal BR, Martz LW. 2006. The synoptic climate controls on hydrology in the upper reaches of the Peace River Basin. Part I: snow accumulation. *Hydrological Processes* **20**: 4097–4111.
- Schaake J, Franz K, Bradley A, Buizza R. 2006. The hydrological ensemble prediction experiment (HEPEX). *Hydrology and Earth System Sciences Discussions* **3**: 3321–3332.
- Schaefli B, Gupta H. 2007. Do Nash values have value? *Hydrological Processes* **21**: 2075–2080.
- Scherrer SC, Appenzeller C, Eckert P, Cattani D. 2004. Analysis of the spread-skill relations using the ECMWF ensemble prediction system over Europe. *Weather and Forecasting* **19**: 552–565.
- Schumann G, Matgen P, Hoffmann L, Hostache R, Pappenberger F, Pfister L. 2007. Deriving distributed roughness values from satellite radar data for flood inundation modelling. *Journal of Hydrology* **344**: 96–111.
- Segond M-L. 2006. *Stochastic Modelling of Space-time Rainfall and the Significance of Spatial Data for Flood Runoff Generation*. Imperial College London: London; 222.
- Simmons A. 2001. Representation of stratosphere in ECMWF operations and ERA-40, In *ECMWF/SPARC Workshop on Modelling and Assimilation for the Stratosphere and Tropopause*, ECMWF, Reading.
- Sivapalan M, Blöschl G. 1998. Transformation of point rainfall to areal rainfall: intensity-duration frequency curves. *Journal of Hydrology* **204**: 150–167.
- Smith MB, Koren VI, Zhang Z, Reed SM, Pan JJ, Moreta F. 2004. Runoff response to spatial variability in precipitation: an analysis of observed data. *Journal of Hydrology* **298**: 267–286.
- Struthers I, Sivapalan M. 2006. Theoretical investigation of process controls upon flood frequency: role of thresholds. *Hydrology and Earth System Sciences Discussions* **3**: 3279–4419.
- Uebel T. 2003. In *Logical Empiricism: Historical and Contemporary Perspectives*, Parrini P, Salmon WC, Salmon MH (eds). University of Pittsburgh Press: Pittsburgh, PA; 368.
- van Griensven A, Bauwens W. 2003. Multiobjective autocalibration for semidistributed water quality models. *Water Resources Research* **39**: DOI: 10.1029/2003WR002284.
- Verbunt M, Zappa M, Gurtz J, Kaufmann P. 2006. Verification of a coupled hydrometeorological modelling approach for alpine tributaries in the Rhine basin. *Journal of Hydrology* **324**: 224–238.
- Vincenty T. 1975. Direct and inverse solutions of geodesics on the ellipsoid with application of nested equations. *Survey Review* **23**: 88–93.
- Vrugt JA, Gupta HV, Bastidas LA, Bouten W, Sorooshian S. 2003. Effective and efficient algorithm for multiobjective optimization of hydrologic models. *Water Resources Research* **39**. DOI: 10.1029/2002WR001746.
- Wagner W, Lemoine G, Borgeaud M, Rott H. 1999. A study of vegetation cover effects on ERS scatterometer data. *IEEE Transactions on Geoscience and Remote Sensing* **37**: 938–948.
- Woods R, Sivapalan M. 1999. A synthesis of space-time variability in storm response: rainfall, runoff generation, and routing. *Water Resources Research* **35**: 2469–2485.