

The Skill of Probabilistic Precipitation Forecasts under Observational Uncertainties within the Generalized Likelihood Uncertainty Estimation Framework for Hydrological Applications

F. PAPPENBERGER, A. GHELLI, AND R. BUIZZA

European Centre for Medium-Range Weather Forecasts, Reading, United Kingdom

K. BÓDIS

Joint Research Centre of the European Commission, Ispra, Italy

(Manuscript received 15 August 2007, in final form 7 November 2008)

ABSTRACT

A methodology for evaluating ensemble forecasts, taking into account observational uncertainties for catchment-based precipitation averages, is introduced. Probability distributions for mean catchment precipitation are derived with the Generalized Likelihood Uncertainty Estimation (GLUE) method. The observation uncertainty includes errors in the measurements, uncertainty as a result of the inhomogeneities in the rain gauge network, and representativeness errors introduced by the interpolation methods. The closeness of the forecast probability distribution to the observed fields is measured using the Brier skill score, rank histograms, relative entropy, and the ratio between the ensemble spread and the error of the ensemble-median forecast (spread–error ratio). Four different methods have been used to interpolate observations on the catchment regions. Results from a 43-day period (20 July–31 August 2002) show little sensitivity to the interpolation method used. The rank histograms and the relative entropy better show the effect of introducing observation uncertainty, although this effect on the Brier skill score and the spread–error ratio is not very large. The case study indicates that overall observation uncertainty should be taken into account when evaluating forecast skill.

1. Introduction

The quality of meteorological forecasts can be a key controlling factor for the quality of a hydrological forecast. Pappenberger et al. (2005) have shown that any low-quality meteorological forecast leads to large uncertainties not only in the coupled surface runoff model but also in the subsequent flood inundation predictions. However, the effect of precipitation uncertainty on runoff estimation is complex and it sometimes shows contradictory results (Segond 2006). Nevertheless, the importance of acknowledging this uncertainty in applying any hydrological model is unquestionable. In the last decade, ensemble prediction systems have substantially improved their precipitation probability forecasts (Palmer 2002). It has been shown that such systems are superior to single deterministic forecasts for certain time ranges

(Palmer 2000; Zhu et al. 2002). Therefore, ensemble predictions are increasingly used as input into real-time flood forecasting systems (de Roo et al. 2003; Gouweleeuw et al. 2005; Pappenberger et al. 2005). It is important that these forecasts are evaluated on the scale of interest for hydrological applications, in particular catchment scale (see, e.g., Ahrens and Jaun 2007).

In this paper, all evaluations are performed on catchment averages. Verifications need to acknowledge the uncertainty in the observations as well as in the forecast. Wilson et al. (1999) have shown how scores vary if the point forecast for temperature, precipitation, wind speed, cloud amounts, and visibility is verified against observations that have been fitted with theoretical distributions. Ahrens and Jaun (2007) evaluated forecast accuracy of precipitation on a catchment scale and investigated the importance of the uncertainty resulting from the spatial interpolation of precipitation data. They derived an ensemble of interpolated rain gauge data using stochastic simulations with kriging. Their research demonstrated that the influence of the uncertainties is

Corresponding author address: Florian Pappenberger, ECMWF, Shinfield Park, Reading RG2 9AX, United Kingdom.
E-mail: florian.pappenberger@ecmwf.int

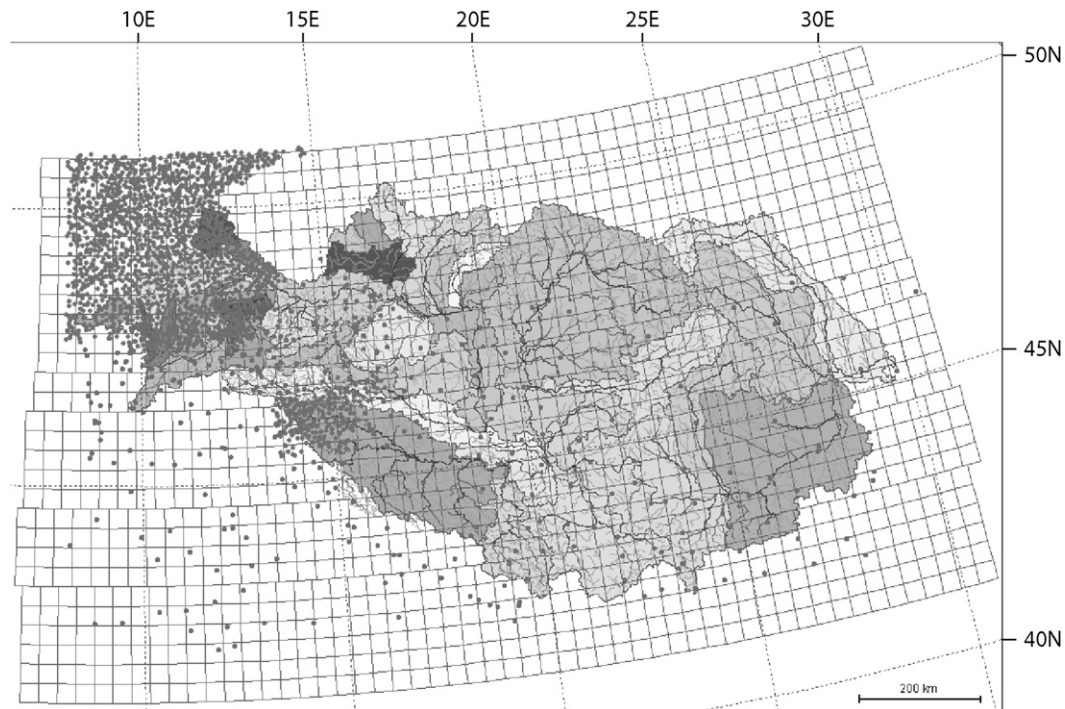


FIG. 1. The subdivision of the Danube catchment into 45 subcatchments (catchments are overlapping). The overlaying grid is the forecast grid of the ensemble forecasts. The dots indicate measurement stations. Catchment areas with fewer than five stations have been excluded from further analysis.

substantial. Ahrens and Jaun (2007) neglected the uncertainty due to the interpolation method and the error in the measurements.

In this paper, a methodology that takes into account measurement errors, inhomogeneities of the rain gauges network, and representativeness of spatial interpolation as source of uncertainty are presented. This is achieved by applying a generalized Bayesian method based on Monte Carlo (MC) analysis, which leads to probability distributions of catchment averages. The method embeds Brier skill score, rank histogram, relative entropy, and the ratio between the ensemble spread and the error of the ensemble-median forecast (spread–error ratio) into a framework in which both observation and forecast uncertainties are acknowledged. The value of this methodology will be demonstrated using a case study of flooding that happened in July and August 2002 in the Danube catchment. The flooding affected more than 600 000 people and caused damage in excess of 15 million U.S. dollars (USD) (WHO 2007). Since the major cause of this flooding was heavy rain, attention will be focused on this weather variable.

The paper begins with a description of the catchment and flood event (section 2). Section 3 describes the forecast system, the Generalized Likelihood Uncertainty Estimation (GLUE) method, which has been used to

compute probability distributions of catchment mean precipitation. A description of the methodology for incorporating observation uncertainty is included as well as a summary of the scores used in this study. Section 4 presents the results of the case study, while conclusions are drawn in section 5.

2. Description of catchment

The Danube River has an approximate catchment area of 817 000 km², with a river length of 2857 km that is shared among 18 countries. Around 82 million people live within the catchment area of the Danube River, which is the second largest river in Europe. The source is in the Black Forest (Germany, 8°09′) and its mouth is at the Black Sea (29°05′). It has an uneven orography, with heights up to 4000 MSL and an average elevation of 475 MSL. (UNESCO 2006).

The catchment has been subdivided into 45 subcatchments (Fig. 1) upstream of gauging stations used within the European Flood Alert System, whose aim is to provide medium-range flood alerts across Europe with a lead time between 3 and 10 days (de Roo et al. 2003). The subcatchments range from very small and compact, such as the River Hron, to basins that spread nearly over the entire domain, such as that of the upstream Iron

Gate Reservoir. Thus, different numbers of forecast grid points and observations are contained in each catchment. Catchment areas range from 1600 to 132 000 km², with a mean of 22 333 km². The catchments overlap each other, in the case of gauging stations upstream on the same river. For example, the catchment upstream of Iron Gate also contains many smaller catchments from the upper Danube, as catchments are nested into each other.

3. Methodology

All analyses are performed on catchment averages of 24 h of accumulated precipitation. The description of the methodology is subdivided into three parts: The first part concentrates on the representation of the forecast uncertainty. The second part describes the methodology used to derive probability distributions from uncertain observations by using a generalized Bayesian framework. The third part discusses the scores used to compare forecast probability distribution and observed average catchment precipitation.

a. Forecast uncertainty

The rainfall forecasts used in this study are based on the European Centre for Medium-Range Weather Forecasts (ECMWF) ensemble forecasts [known as the Ensemble Prediction System (EPS)]. These provide 50 realizations for a 15-day lead time. A detailed description of the ECMWF numerical weather prediction model can be found in Molteni et al. (1996), Persson (2001), Buizza et al. (2001), and Gouweleeuw et al. (2005). For this study, a set of precipitation forecasts in the period from 20 July to 31 August 2002 with a 7-day lead time has been used. The forecasts, at TL399L40 resolution (i.e., spectral triangular truncation 399 with linear grid and 40 vertical levels, which correspond to about 80-km resolution at midlatitude), start at 1200 UTC of the first day of each period. Each forecast grid contributes to the catchment averages proportionally to its fractional area within each catchment.

b. Observation uncertainty

The precipitation measurements from this study have been obtained from (i) a high-density network of rain gauges owned by countries that are part of the Danube catchment (Ghelli and Lalaurette 2000) and (ii) Synoptic Ocean Prediction Experiment (SYNOP) stations available on the Global Telecommunication System (GTS). The observed accumulated precipitation is for a period of 24 h starting at 0600 UTC to match observed and forecast quantities; the selected forecast lead times are 42, 66, 90, 114, 138, and 162 h. Figure 1 shows the uneven spatial distribution of the observations: high-density coverage on the top western part of the Danube and very low density on the eastern part.

Therefore, observations are highly uncertain. The influence of this uncertainty has been split into three categories: (i) uncertainty due to measurement error, uncertainty due to inhomogeneities of the network density, and (iii) uncertainty introduced by the application of an interpolation method. The GLUE framework (Beven and Binley 1992) has been used for this analysis. This methodology will first be explained, then followed by a discussion of the three sources of uncertainty in deriving catchment averages from observations.

1) GLUE

A common feature of environmental models is that values of specific input variables (e.g., precipitation) can rarely be known sufficiently well enough to reproduce model predictions that agree unequivocally with the available verification dataset (Beven 2006; Beven and Binley 1992; Beven and Freer 2001). This uncertainty needs to be taken into account (Pappenberger and Beven 2006). Model predictions that cannot reproduce the observations sufficiently are classed as nonbehavioral. Additional uncertainties in model predictions may also arise from the structure of the chosen model, from the nature of imposed boundary conditions, and from errors linked to approximation processes used in the numerical solution. In this paper, the MC framework is used to analyze model uncertainty. The MC approach consists of running repeated simulations of a model using a range of values for each uncertain input parameter or factor. The concept of equifinality (Beven 2006) is based on the assumptions that there may be a large number of parameter sets across the parameter space that are able to map model predictions to the observed data to an acceptable level of performance. Uncertainty analysis techniques, such as the GLUE methodology of Beven and Binley (1992), accept the notion of equifinality and attempt to estimate the level of confidence that can be placed upon a range of model predictions rather than concentrating on a single "optimum" prediction. The detailed description of the GLUE procedure is given in Beven and Binley (1992).

2) UNCERTAINTY AS A RESULT OF MEASUREMENT ERRORS

Comprehensive reviews of errors in precipitation measurements are presented by Sevruk (2005) and Willems (2001), who also point out that it is very difficult to measure precipitation without introducing systematic errors and biases (Sevruk 1986; Sevruk and Klemm 1989). Sevruk (1986) specified and quantified five main factors that introduce a systematic error into the measurements (see Table 1).

Correction procedures involve the knowledge of meteorological conditions (such as wind speed) and type of

TABLE 1. Main components of systematic error in precipitation measurements.

Error	Magnitude (%)
Loss due to wind field deformation above the gauge orifice	2–10 (10–50 in snow conditions)
Losses from wetting on internal walls of the collector and in the container when it is emptied	2–10
Loss due to evaporation from the container	0–4
Splash out and splash in	1–2
Blowing and drifting snow	–

gauge. This study uses some generalized assumptions about the measurement errors, as not enough information was available to attribute individual measurement errors to each rain gauge. The error distribution is assumed to be represented by the following equation:

$$\tilde{P}_{N,t} = P_{N,t} - \left(w_{N,t} + \frac{r_{N,t}}{100} + \frac{\text{ews}_{N,t}}{100} \right) P_{N,t}, \quad (1)$$

where P_N is measured precipitation at location N at time t ; \tilde{P}_N is estimated precipitation at location N at time t ; $w_{N,t}$ is error induced by wind at location N at time t ; $r_{N,t}$ is random error at location N at time t ; and $\text{ews}_{N,t}$ is error induced by evaporation, wetting, and splashing at location N at time t .

The wind-induced error is represented (Nespor and Sevruk 1999) by

$$w_{N,t} = (b_1 P_{N,t}^{b_2} + b_3)^{0.14}, \quad (2)$$

where $b_1 \sim U(0.0154, 10)$, $b_2 \sim U(-4.392, -0.0879)$, $b_3 \sim U(0, 0.3)$, and U is uniform distribution. Equation (2) is constrained by the envelope curve of all results presented in the paper by Nespor and Sevruk (1999). The equation, as in Nespor and Sevruk (1999), assumes random wind. The random error distribution was estimated from results by Ciach (2002) as

$$r_{N,t} \sim N\left(0, \sqrt{0.1095 P_{N,t}^{-0.725}}\right), \quad (3)$$

where $N(a_1, a_2)$ is a normal distribution with mean a_1 and standard deviation a_2 .

The error induced by evaporation, wetting, and splashing is based on the results presented in wide range of studies (Lewis and Harrison 2007; Michelson 2004; Molini et al. 2005; Ren and Li 2007; Sevruk 1986, 1996, 2005; Sevruk and Klemm 1989; Sieck et al. 2007; Strangeways 2004; Tartaglione et al. 2005):

$$r_{N,t} \sim \frac{U(0.05, 0.3)^{0.16 P_{N,t}}}{P_{N,t}}. \quad (4)$$

In Fig. 2, the various error sources are presented for a MC experiment for precipitations up to 20 mm. The figure illustrates that errors can be very large, especially at low precipitation amounts and that wind and wetting–evaporation–splashing errors contribute the most to the total error.

3) UNCERTAINTY AS A RESULT OF INHOMOGENEOUS DENSITY

Inhomogeneities in the network density are a major source of uncertainty and taking account of this uncertainty may prove difficult. On the one hand, one wants to use the entire sample to generate the best possible precipitation field. On the other hand, the errors in the network densities have to be quantified. In this analysis, 20% of all stations are randomly selected and omitted from the computation of interpolated precipitation fields. Care has been taken to ensure that no area ends up with a network density of zero (e.g., in some of the Danube area, the five nearest neighbors of the omitted station could not be left out as well). The remaining stations are used to compute the interpolated fields.

4) SKILL OF INTERPOLATED FIELD

The performance is assessed using a fuzzy framework. The framework (Ebert 2008) assigns a membership value between 0 and 1 that should be interpreted as likelihood measures of the interpolated field. The membership function is based on the MC experiment presented in Fig. 2 and Eq. (5). The mean and maximum errors are computed for each precipitation amount, and they define the performance function as (for a comparable application see Pappenberger et al. 2006b):

$$\begin{aligned} a &= \overline{t_{P_{N,t}}}, \\ b &= \max(t_{P_{N,t}}), \quad \text{and} \\ \mu_{N,t}(\tilde{P}_{N,t}, P_{N,t}, b, a) &= \begin{cases} 0, & \text{for } P_{N,t} - \tilde{P}_{N,t} > bP_{N,t}, P_{N,t} - \tilde{P}_{N,t} < 0 \\ \frac{(P_{N,t} - \tilde{P}_{N,t}) - (bP_{N,t})}{(aP_{N,t}) - (bP_{N,t})}, & \text{for } P_{N,t} - \tilde{P}_{N,t} \leq bP_{N,t}, P_{N,t} - \tilde{P}_{N,t} \geq aP_{N,t} \\ 1, & \text{for } P_{N,t} - \tilde{P}_{N,t} < aP_{N,t}, \end{cases} \end{aligned} \quad (5)$$

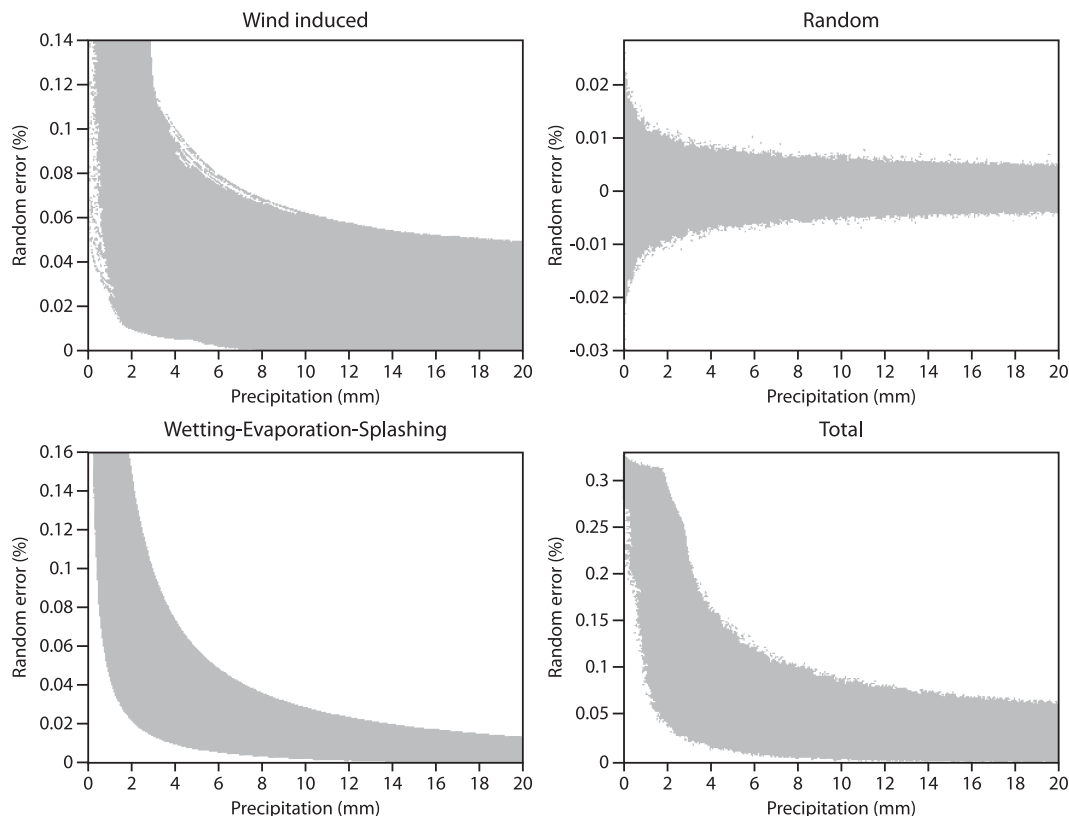


FIG. 2. Results of an MC experiment for precipitation up to 20 mm and the same error sources as in Eq. (1).

where μ is the membership value and t is the total error [Eq. (1)].

The average membership value is used as the likelihood measure for the interpolated field. An interpolated field is rejected as nonbehavioral if more than 50% of all stations within one of the subcatchments of the omitted 20% total stations had a performance [as in Eq. (5)] of zero.

5) INTERPOLATION UNCERTAINTY

Several studies (Busuioc et al. 2006; Coelho et al. 2006; Khan et al. 2006a,b; Rebora et al. 2006) have described upscaling or downscaling methods for precipitation. Uncertainties arising from these scaling methods cannot be ignored. Tustison et al. (2001) pointed out that there is always a representativeness error (or commensurability error; see Beven 2006) when point data are matched to a model scale or vice versa. The magnitude of this error depends on measurement and density ambiguities as well as on the statistical structure of the underlying field, the interpolation scheme, and the scale that is used for averaging (Tustison et al. 2001). For example, the magnitude of a variance reduction factor in producing catchment-based intensity–duration–

frequency (IDF) curves depends on the correlation structure of the precipitation [as well as on catchment size and shape (Sivapalan and Bloschl 1998)]. Point-to-area scaling is central to this study, as the scope of the paper is to assess the quality of the forecast on a catchment scale. Additionally, earlier analysis comparing precipitation forecasts to observed precipitation has been performed using gridded data that represent areal quantities (Cherubini et al. 2002; Ghelli and Lalaurette 2000). Therefore, in this framework, the interpolation scheme has significant influence on the analysis. Syed et al. (2003) and Creutin and Obled (1982) review different interpolation methods, such as Thiessen polygons (Creutin and Obled 1982; Dirks et al. 1998; McCuen 1998; Tabios and Salas 1985; Thiessen 1911), inverse square distance (Bedient and Huber 1988), isohyetal methods (McCuen 1998), kriging, nearest neighbor or arithmetic mean (Creutin and Obled 1982), or splines and other multiquadratic methods (Shaw and Lynn 1972). Moreover, secondary variables, such as topography (Goovaerts 2000) or distance from the coastline (Marquinez et al. 2003), may be included in the interpolation scheme. Many authors argue that kriging is superior to other commonly used techniques (Abtey

et al. 1993; Creutin and Obled 1982; Shaw and Lynn 1972; Tabios and Salas 1985), provided all assumptions are met (Borga and Vizzaccaro 1997). Skok and Vrhovc (2006) have shown that errors up to 50% are introduced by any interpolation method.

In this study, four different interpolation methods from the list above have been chosen as representative for interpolating a $100\text{ m} \times 100\text{ m}$ grid over the Danube catchment, which is then aggregated to the catchment scale:

- Quasi kriging: a simplified kriging approach that uses topography as a secondary variable (Heise and Rivin 2001; Shepard 1968)
- Linear: triangle-based linear interpolation (Watson 1994)
- Cubic: triangle-based cubic interpolation (Yang 1986) (all negative values have been set to zero, as negative rainfall is physically impossible, so this invalidates some of the assumptions of this interpolation method)
- Nearest neighbor: nearest-neighbor interpolation (Watson 1994)

Interpolation methods that not only recognize spatial correlation but also temporal correlation (Segond 2006) are a natural extension for the future progression of this study.

Figure 3 summarizes the different steps for deriving the average catchment precipitation probability distributions. These probability distributions are used to compute skill scores.

c. Performance measures/skill scores

Many different skill scores and performance measures have been developed in meteorology (see, e.g., references in Gandin and Murphy 1992; Gober et al. 2004; Venugopal et al. 2005; Weisheimer et al. 2005) as well as in many other related disciplines (Hagen-Zanker 2006; Hagen-Zanker et al. 2006; Pappenberger et al. 2006a). Most of the measures developed do not allow probabilistic inference [refer to Woodhead (2007) for an excellent discussion], although they can be interpreted in a generalized Bayesian likelihood framework (Smith 2006).

In this study, four different evaluation measures have been used: the Brier skill score, the rank histogram, the relative entropy, and the ratio between the ensemble spread and the error of the ensemble median forecast (hereafter spread–error ratio).

- The Brier score (Brier 1950) has traditionally only been used to evaluate probability of occurrence of the forecast system. The observational probability is usually binary (1 if the event occurs and 0 if it does

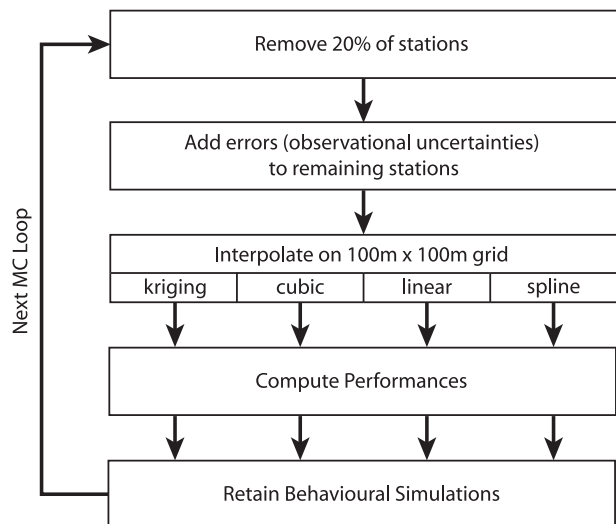


FIG. 3. GLUE framework for the analysis of the observational uncertainties.

not). The method outlined above allows for replacing this binary classification by the probability of occurrence of the observed system. The skill score has been computed using a climatological forecast as reference, whereby the climatology is derived from the average frequency of the event within the 43 days of forecast. The thresholds for the score have been chosen to reflect the 10th, 35th, 65th, and 90th percentiles of all observations within the evaluation period, based on an interpolation using the full set of observations and no additional error sources. Therefore, the corresponding precipitation thresholds are 0.001 , 0.12 , 1.9 , and $9.7\text{ mm (24 h)}^{-1}$. The rank probability score is also computed as the average over all the thresholds. Although these thresholds reflect the characteristics of the event, note that EPS has a negative bias at small thresholds (rains too little) and a positive bias at large thresholds (rains too much).

- The rank histogram, also known as the Talgrand histogram (Hamill et al. 2001; Talagrand 1997), shows the probability of the observation falling between any two adjacent members of the forecasting system. This probability will be either 1 or 0 if the observation is one value, but it could take any values between 0 and 1 if the observations are expressed in a probabilistic fashion. In fact, a probabilistic precipitation observation can contribute to several ranks according to its probability at the respective locations.
- The relative entropy is a measure of the distance between the observation and the forecast probability distributions. The ignorance score can be derived from the relative entropy, as shown in Roulston and Smith (2002). The formulation of either of these measures

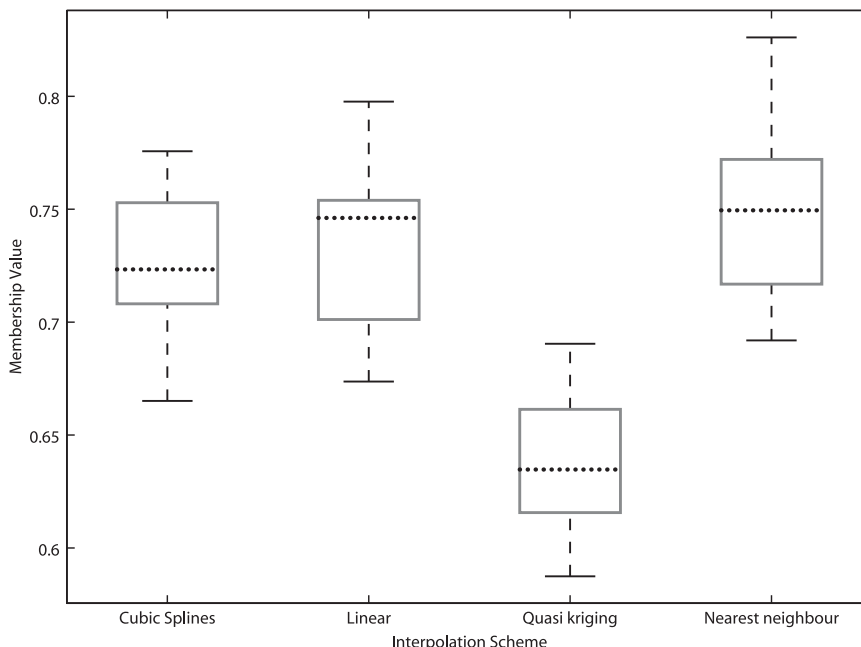


FIG. 4. Box-and-whisker plots of the performance of the four different interpolation schemes. The box marks the 25th and 75th percentiles. The lines extend to the 5th and 95th percentiles. The median is marked with a dotted line.

demands that the probability of any observation or forecast always be greater than zero. Roulston and Smith (2002) have argued that a “forecast probability of zero is impossible, unless it is truly impossible.” Thus, any precipitation that is not coherent with our conceptual understanding of the physics is truly impossible (although even that understanding is limited). Assigning a probability and a cut-off point is crucial for this performance measure, as it is highly sensitive to the tails of the distribution. The physical necessity to specify a cut-off point also makes the ignorance score improper. In this paper, we assumed a minimum probability of 0.1×10^{-5} for all values between 90% of the minimum and 110% of the maximum of all forecast and observed values (distributions are linearly extrapolated to those points, if necessary). Even though the ignorance score (and with it the relative entropy) sensitivity to extreme events and its robustness have been questioned (Gneiting et al. 2007; Gneiting and Raftery 2007), the score is used in this paper on the basis that it is representative for this application.

- Additionally, the spread-error ratio is used in this paper. From each distribution (observations and forecast), the difference between the 90% and 10% percentile will be computed, and the value of the forecasts subtracted from the observations. The error

is calculated as the difference between the median of the observations minus the forecasts.

A more detailed description of the scores can be found in Jolliffe and Stephenson (2003).

4. Results of case study

The analysis of the results is split into two parts: (i) the different interpolation schemes are compared and then (ii) a discussion of results looking at the different performance measures introduced previously.

a. Comparison of interpolation schemes

The relative merits of various interpolation schemes have been examined at length in the past. A definite ranking cannot be easily established a priori, because it highly depends on the area and event interpolated; therefore, in the present paper, the four selected schemes have been compared, and Fig. 4 shows the results. The boxplots show the mean (dotted line), 25th, and 75th percentiles (lower and upper side of the box) of the membership value (the larger the membership value, the better the system). Differences among the various interpolation methods are small. The nearest-neighbor interpolation scheme has the best average performance, followed by the cubic spline and linear interpolation

TABLE 2. Average rank of each interpolation method derived from the same set of data (see Fig. 2).

Method	Cubic splines	Linear	Quasi kriging	Nearest neighbor
Average rank	2.1	1.9	3.7	1.9

methods. However, the uncertainties are very large and the differences are within these ranges, which suggests an equifinality of the interpolation schemes.

It is also possible to perform a direct comparison between the four schemes. Figure 3 shows that a set of four maps is derived from one sample by applying the four interpolation methods. The average ranks that each scheme achieves within one set of realizations are displayed in Table 2.

In this comparison, the nearest-neighbor and the linear interpolation schemes are better than the other two methods, but criteria that distinguish between behavioral and nonbehavioral simulations are not based on these ranks [for the definition of “behavioral,” refer to Eq. (5) in section 3b(4)]. Thus, hereafter, no further distinction among the interpolation methods is made and they are analyzed together.

b. Performance assessment

A positive Brier skill score indicates whether a forecast is more skilful than a reference forecast (in this case, the mean over the period). In Table 3, the average skill scores for all catchments and the four thresholds are shown [thresholds are chosen according to percentiles (refer to the definition of skill scores for more details)]. It can be seen that the skills are fairly low and they decrease with lead time. The skill of the control experiment, the interpolated field with all available data points and no errors added to it, is included in the table under the column labeled 100. The Brier skill score for the control is generally lower than for the uncertain observations. This indicates that the methodology is able to capture some of the errors inherent in the observations and is able to make a fairer comparison with the model forecasts.

Table 3 summarizes all catchments, although catchments contributed to a different extent to the flood

event. The table shows a decrease of the skill score with lead time and significant differences in the skill at different thresholds. In Fig. 5, the Brier skill scores for different lead times are compared to each other for each catchment for the $1.3 \text{ mm (24 h)}^{-1}$ threshold. The histograms show the Brier skill score distribution. The figure shows a scatterplot for each combination of lead times, which compares the skills. Dots indicate catchments with a center of gravity of less than 16° west, and all other catchments are plotted as crosses. The figure indicates that not all forecasts for all catchments are skill full (also valid for all other thresholds) for lead times up to day 7 and that many of the Brier skill scores are not very high. In general, the skill of the forecast deteriorates with increasing lead time, as the majority of the forecasts is below the gray line (dots on the gray line have the same skill for two consecutive lead times). The histograms show no dominant distributions.

Figure 6 depicts the longitude of the catchment center of gravity versus the relative entropy for a lead time of 42 h. The relative entropy clearly decreased from west to east, as eastern catchments have fewer rain gauges and therefore, on average, a larger variability. Roulston and Smith (2002) have shown the sensitivity of the ignorance score to variance, which explains the behavior in this study. This example also shows that different performance measures exhibit different sensitivity to observed variability.

It has been speculated that smaller catchments have a lower skill score, as they contain fewer model grid points and thus experience a lower smoothing effect. An analysis of the rank probability score against catchment size (not shown) demonstrates that there is no relationship between catchment size and skill for this particular event.

Table 4 summarizes the forecast outliers for different lead times and three different catchment categories: small (less than 7018 km^2), medium (between 7018 and $24\,448 \text{ km}^2$), and large (greater than $24\,448 \text{ km}^2$). This subdivision allows for 15 catchments in each category.

Table 4 indicates over- and underprediction outliers in the early time steps and a more even distribution at longer lead times. The differences between the three different catchment sizes are small. However, smaller

TABLE 3. Average Brier skill scores for six different lead times.

Threshold/lead time [mm (24 h)^{-1}]	0.001		0.12		1.9		9.7	
Stations used in interpolation (%)	80	100	80	100	80	100	80	100
42 h	0.12	0.09	0.30	0.2	0.37	0.17	0.21	-0.2
66 h	0.11	0.02	0.29	0.2	0.33	0.15	0.13	-0.23
90 h	0.10	0.01	0.27	0.19	0.30	0.15	0.10	-0.23
114 h	0.10	0.01	0.26	0.19	0.24	0.14	0.09	0.25
138 h	0.08	-0.02	0.20	0.16	0.23	0.14	0.09	-0.28
162 h	0.06	-0.05	0.19	0.12	0.22	0.11	0.06	-0.3

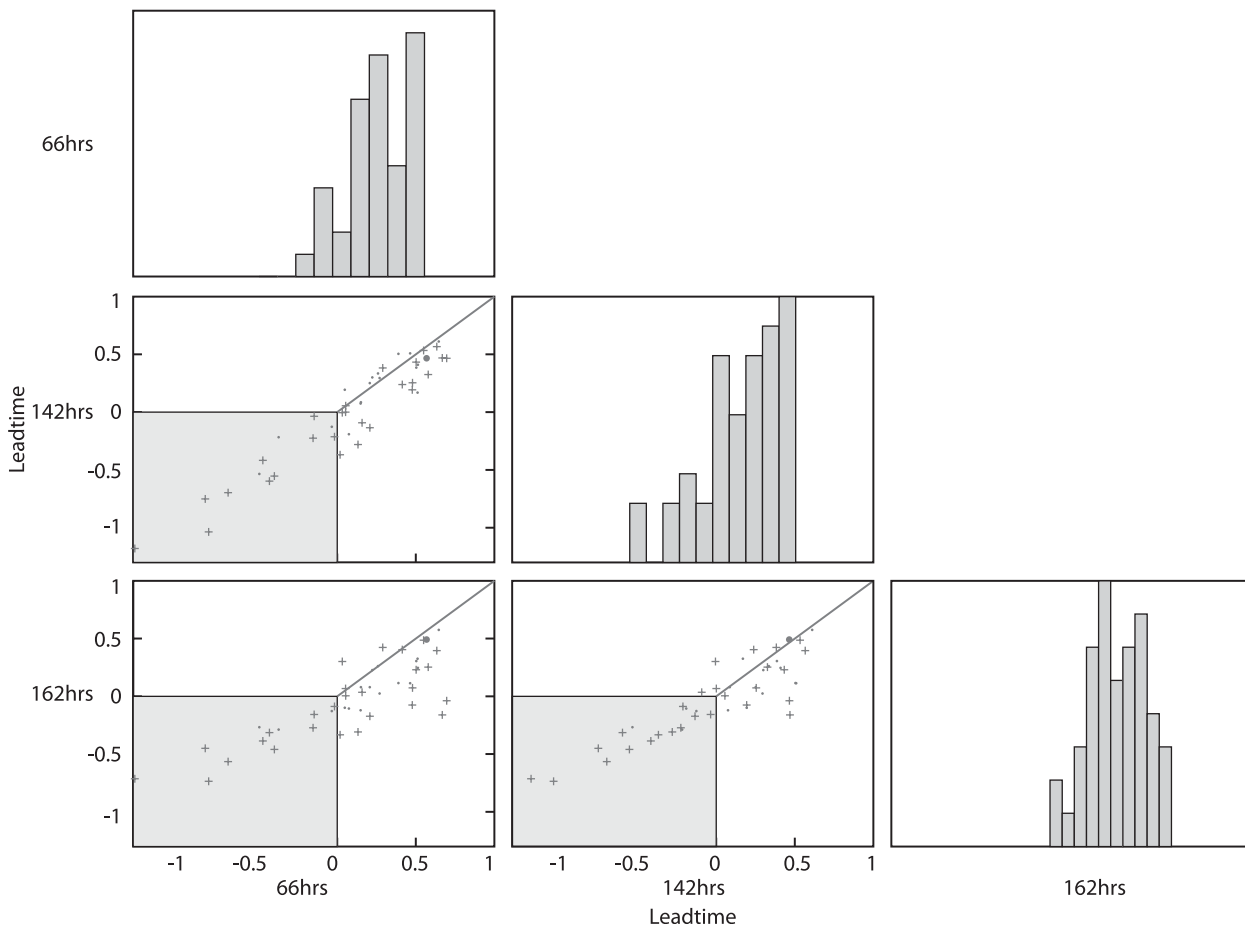


FIG. 5. Comparison of Brier skill scores with different lead times for a threshold of $1.2 \text{ mm} (24 \text{ h})^{-1}$ in scatter diagrams and histograms. In the scatter plots the lead times are plotted against each other for the same catchment. Histograms of the distribution of the Brier skill score are plotted on the diagonal of this plot. The shaded areas indicate negative Brier scores.

catchments seem to have the tendency to proportionally overpredict at the longer lead times, whereas this signal is less evident in larger catchments. The EPS has shown the tendency to overpredict at longer lead times; this signal is stronger for the smaller catchments, which may be explained by the smoothing effect mentioned. There is no sensitivity to the location of the catchment, as eastern and western catchments have similar entropy scores.

It can be observed that uncertainties in the observations lead to flatter rank histograms. While in a deterministic evaluation, one observation can only contribute to one bar of the histogram—for example, an observation can only contribute to one rank despite being only a small amount larger than the respective forecast; in a probabilistic evaluation, the contribution of each observation is spread out, thus leading to a more uniform appearance of the rank histogram.

The effect of observational error to the skill can also be seen by looking at the spread–error ratio. Figure 7 shows the absolute error of the mean observation minus the

mean forecast against the spread of the forecast subtracted from the spread of the observation for all lead times. Catchments with a center of gravity of less than 16° longitude are plotted as open black circles (western catchments) and all others as closed gray squares. Additional information is given to the percentage of catchments in each quadrant of the figures. For example, at a lead time of 42 h, 7% of the eastern catchments have a positive spread difference and a negative mean difference; 23% have a positive spread difference and a positive mean difference; 47% have a negative spread difference and a negative mean difference; and 23% have a negative spread difference and a positive mean difference. The following four conclusions can be drawn from Fig. 7:

- i) Eastern catchments have on average a smaller spread difference and mean absolute error in comparison to western catchments.
- ii) The mean absolute error and the spread of mean distance increase with lead time.

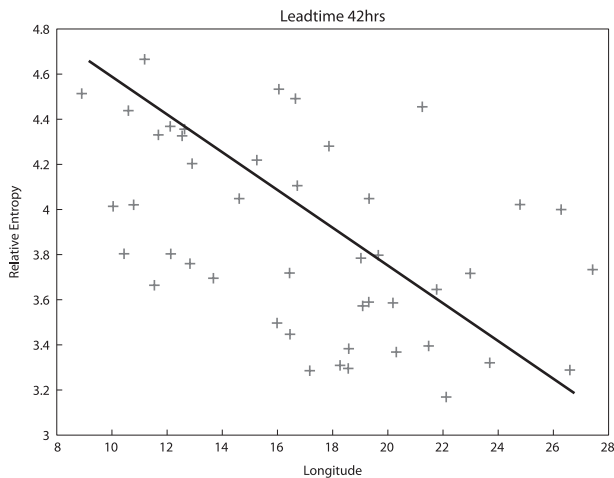


FIG. 6. Longitude of the catchment center of gravity vs relative entropy for a lead time of 42 h.

- iii) Negative spread difference is a dominant feature.
- iv) All conclusions are subjected to high uncertainties, as sample size is relatively small, which is inevitable with extreme events.

The first conclusion shows that the uncertainty in the forecast is dominant if the observed precipitation is low. The increase in the spread of the mean absolute error is explained by the increasing spread of the forecast, which leads to smaller errors between the two means. The dominance of the negative spread difference suggests that the spread of the forecast is larger than the spread produced by the uncertain observations; thus, the spread is too large. However, the high variability in the results also indicates the importance of recognizing the uncertainty in the observations. The importance of the uncertainty in the observations is clearly shown in the range of the spread differences, which in some cases exceeds the mean difference. In catchments with the mean difference larger than the spread error difference, the uncertainty in the observations has a negligible effect.

5. Conclusions

This paper introduces a methodology for evaluating ensemble forecasts using uncertain observations for catchment-based precipitation averages. The forecast uncertainty is generated by the Ensemble Prediction System of the European Centre for Medium-Range Weather Forecasts. The observations include uncertainty introduced by error in the measurements, inhomogeneities in the density of the rain gauges network, and representativeness error introduced by the interpolation method. Probability distributions for mean catchment precipitation are derived using the Gener-

TABLE 4. Outliers forecast vs lead time for three different catchment sizes.

Lead time (h)	Small (%)	Medium (%)	Large (%)
42	20	18	18
66	15	13	13
90	14	11	10
114	11	9	8
138	11	8	7
162	11	7	5

alized Likelihood Uncertainty Estimation framework with a five-step Monte Carlo loop. The first 20% of the stations are removed randomly, and the error is added to the remaining stations. The observed quantities are interpolated on a $100 \text{ m} \times 100 \text{ m}$ grid using four different interpolation methods (quasi kriging, cubic spline, linear, and nearest neighbor). A skill score based on fuzzy membership functions is computed for the remaining 20% of the stations with the four precipitation fields. Fields that underperform according to these skill score are classified as nonbehavioral and are not included in any further analysis. The skill score is also used to generate probabilistic distributions for catchment averages (which are aggregated from the finer grid). The probability distribution of the forecast and the observed are compared using four different scores: Brier skill score, rank histograms, relative entropy, and spread-error ratio. The methodology is applied to 43 cases (from 20 July to 31 August 2002). The four performance indicators highlight the sensitivity of the skills to catchment size and location and observation uncertainty. Five main conclusions can be drawn from this work.

- The sensitivity study suggests that there are small differences between the different interpolation methods for these 43 cases.
- The Brier skill score is always positive—however, not very large—suggesting skill for catchment average forecasts.
- Catchments in the east have worse performances than in the west due to lower precipitation and higher variability (fewer observation stations, smoother topography) in the catchment averages. A comparison of the spread differences shows that (i) eastern catchments have on average a smaller spread differences and mean absolute error in comparison to western catchments; (ii) mean absolute error and the spread of the mean differences increases with lead time, and uncertainty in the forecast dominates the results; and (iii) the spread of the ensemble forecasts is larger than the spread of the probability distribution of the observations.

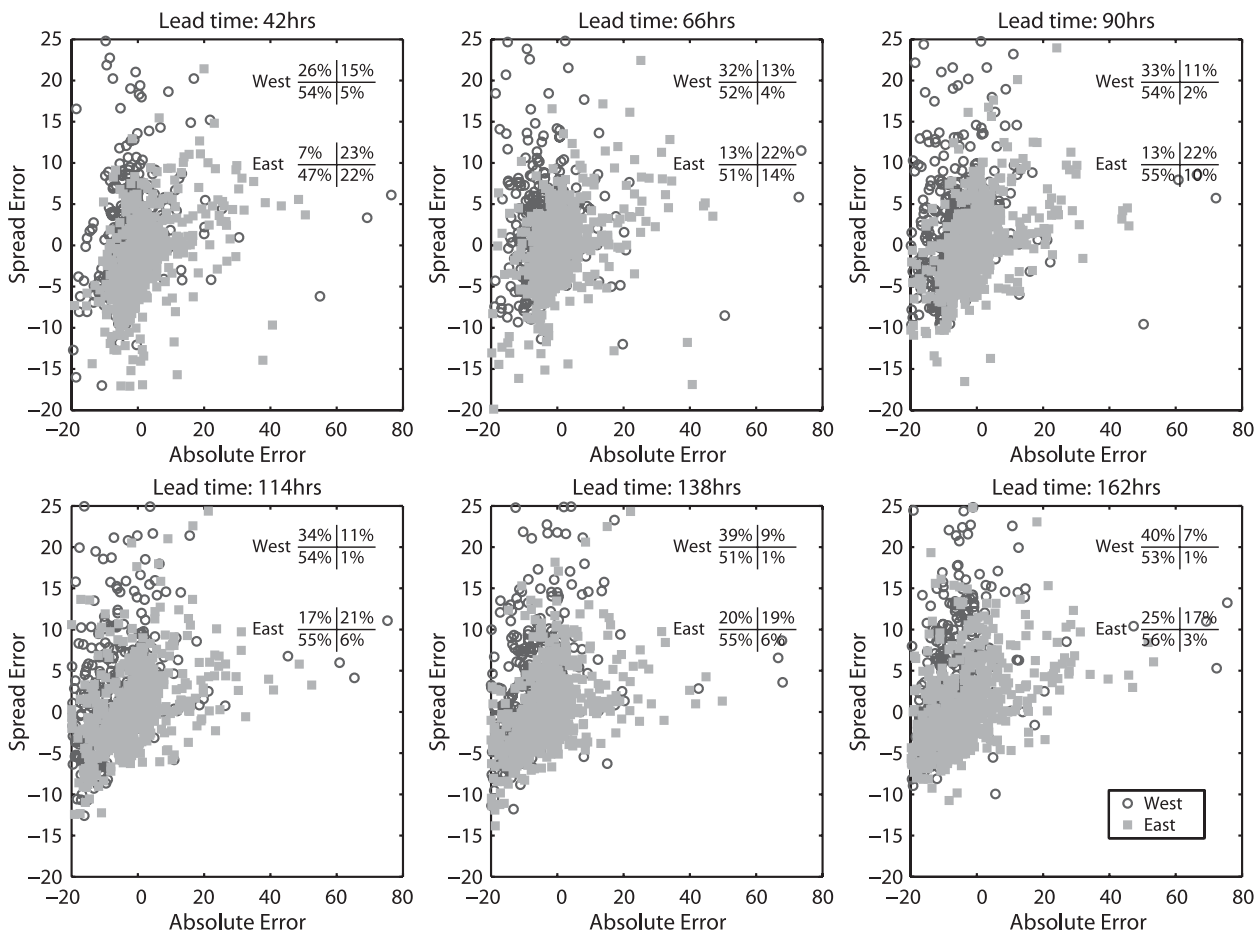


FIG. 7. Plot of the absolute error of observation mean minus forecast mean against the spread of the forecast subtracted from the spread of the observation for all lead times. Catchments with a center of gravity of less than 16° longitude are plotted as open black circles (western catchments) and all others as closed gray squares (eastern catchments). Additional information is given to the percentage of catchments in each quadrant of the figures. For example, at a lead time of 42 h, 7% of all eastern catchments have a positive spread difference and a negative mean difference; 23% have a positive spread difference and a positive mean difference; 47% have a negative spread difference and a negative mean difference; and 23% have a negative spread difference and a positive mean difference.

- The effect of the uncertainty in the observations can be clearly seen in the rank histograms: the uncertainty flattens the histogram, reducing the number of outliers. The relative entropy is also influenced by the uncertainty in the observations but dominated by an east–west divide.
- Different skill measures have different sensitivity to observation variability and thus uncertainty in the observations. Entropy measures have a higher sensitivity, whereas the Brier skill score is less sensitive.

Even though the analysis is based on 24 h of accumulated precipitation, which may not be the preferred use for hydrological purposes, it clearly points out that model evaluation should include observation uncertainty.

Acknowledgments. Florian Pappenberger is funded by the PREVIEW (FP6 work package: plain floods) pro-

gram (available online at <http://www.preview-risk.com>). We thank the Deutscher Wetter Dienst (Germany), the Zentralanstalt für Meteorologie und Geodynamik (Austria), the Agencija Republike Slovenije za okolje (Slovenia), and the Országos Meteorológiai Szolgálat (Hungary) for providing data of the high-density data network. Moreover, the comments by Hannah Cloke (King’s College London) and Phil Younger (Lancaster University) greatly helped to improve the quality of this paper.

REFERENCES

Abtew, W., J. Obeysekera, and G. Shih, 1993: Spatial-analysis for monthly rainfall in south Florida. *Water Resour. Bull.*, **29**, 179–188.

Ahrens, B., and S. Jaun, 2007: On evaluation of ensemble precipitation forecasts with observation-based ensembles. *Adv. Geosci.*, **10**, 139–144.

- Bedient, P. B., and W. C. Huber, 1988: *Hydrology and Floodplain Analysis*. Addison-Wesley, 650 pp.
- Beven, K. J., 2006: A manifesto for the equifinality thesis. *J. Hydrol.*, **320** (1–2), 18–36.
- , and A. Binley, 1992: The future of distributed models: Model calibration and uncertainty prediction. *Hydrol. Processes*, **6** (3), 279–298.
- , and J. Freer, 2001: Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology. *J. Hydrol.*, **249**, 11–29.
- Borga, M., and A. Vizzaccaro, 1997: On the interpolation of hydrologic variables: Formal equivalence of multiquadratic surface fitting and kriging. *J. Hydrol.*, **195**, 160–171.
- Brier, G. W., 1950: The statistical theory of turbulence and the problem of diffusion in the atmosphere. *J. Meteor.*, **7**, 283–290.
- Buizza, R., D. S. Richardson, and T. N. Palmer, 2001: The new 80-km high-resolution ECMWF EPS. *ECMWF Newsletter*, No. 90, ECMWF, Reading, United Kingdom, 2–9.
- Busuioc, A., F. Giorgi, X. Bi, and M. Ionita, 2006: Comparison of regional climate model and statistical downscaling simulations of different winter precipitation change scenarios over Romania. *Theor. Appl. Climatol.*, **86**, 101–123.
- Cherubini, T., A. Ghelli, and F. Lalaurette, 2002: Verification of precipitation forecasts over the alpine region using a high-density observing network. *Wea. Forecasting*, **17**, 238–249.
- Ciach, G. J., 2003: Local random errors in tipping-bucket rain gauge measurements. *J. Atmos. Oceanic Technol.*, **20**, 752–759.
- Coelho, C. A. S., D. B. Stephenson, F. J. Doblas-Reyes, M. Balmaseda, A. Guetter, and G. J. van Oldenborgh, 2006: A Bayesian approach for multi-model downscaling: Seasonal forecasting of regional rainfall and river flows in South America. *Meteor. Appl.*, **13** (1), 73–82.
- Creutin, J. D., and C. Obled, 1982: Objective analyses and mapping techniques for rainfall fields: An objective comparison. *Water Resour. Res.*, **18**, 413–431.
- de Roo, A., and Coauthors, 2003: Development of a European Flood Forecasting System. *Int. J. River Basin Manage.*, **1**, 49–59.
- Dirks, K. N., J. E. Hay, C. D. Stow, and D. Harris, 1998: High-resolution studies of rainfall on Norfolk Island. Part II: Interpolation of rainfall data. *J. Hydrol.*, **208**, 187–193.
- Ebert, E. E., 2008: Fuzzy verification of high resolution gridded forecasts: A review and proposed framework. *Meteor. Appl.*, **15**, 51–64.
- Gandin, L. S., and A. H. Murphy, 1992: Equitable skill scores for categorical forecasts. *Mon. Wea. Rev.*, **120**, 361–370.
- Ghelli, A., and F. Lalaurette, 2000: Verifying precipitation forecasts using up-scaled observations. *ECMWF Newsletter*, No. 87, ECMWF, Reading, United Kingdom, 9–17.
- Gneiting, T., and A. E. Raftery, 2007: Strictly proper scoring rules, prediction, and estimation. *J. Amer. Stat. Assoc.*, **102**, 359–378.
- , F. Balabdaoui, and A. E. Raftery, 2007: Probabilistic forecasts, calibration and sharpness. *J. Roy. Stat. Soc.*, **69B**, 243–268.
- Gober, M., C. A. Wilson, S. F. Milton, and D. B. Stephenson, 2004: Fairplay in the verification of operational quantitative precipitation forecasts. *J. Hydrol.*, **288**, 225–236.
- Goovaerts, P., 2000: Geostatistical approaches for incorporating elevation into the spatial interpolation of rainfall. *J. Hydrol.*, **228**, 113–129.
- Gouweleeuw, B. T., J. Thielen, G. Franchello, A. P. J. de Roo, and R. Buizza, 2005: Flood forecasting using probabilistic weather predictions. *Hydrol. Earth Syst. Sci.*, **9**, 365–380.
- Hagen-Zanker, A., 2006: Comparing continuous valued raster data: A cross disciplinary literature scan. Research Institute for Knowledge Systems, 34 pp.
- , G. Engelen, J. Hurkens, R. Vanhout, and I. Uljee, 2006: Map Comparison Kit version 3.0: User manual. Research Institute for Knowledge Systems, 73 pp.
- Hamill, T. M., J. S. Whitaker, and C. Snyder, 2001: Distance-dependent filtering of background error covariance estimates in an ensemble Kalman filter. *Mon. Wea. Rev.*, **129**, 2776–2790.
- Heise, E., and G. Rivin, 2001: Precipitation analysis and prediction. Final report on the DWD contribution to the EU Project ‘An European Flood Forecasting System,’ German Meteorological Service Research and Development Division Rep. 80, ISSN 1430-0281, 97 pp.
- Jolliffe, I. T., and D. B. Stephenson, 2003: *Forecast Verification: A Practitioner’s Guide in Atmospheric Science*. J. Wiley, 254 pp.
- Khan, M. S., P. Coulibaly, and Y. Dibiye, 2006a: Uncertainty analysis of statistical downscaling methods. *J. Hydrol.*, **319**, 357–382.
- , —, and —, 2006b: Uncertainty analysis of statistical downscaling methods using Canadian Global Climate Model predictors. *Hydrol. Processes*, **20** (14), 3085–3104.
- Lewis, H. W., and D. L. Harrison, 2007: Assessment of radar data quality in upland catchments. *Meteor. Appl.*, **14**, 441–454.
- Marquinez, J., J. Lastra, and P. Garcia, 2003: Estimation models for precipitation in mountainous regions: The use of GIS and multivariate analysis. *J. Hydrol.*, **270**, 1–11.
- McCuen, R. H., 1998: *Hydrologic Analysis and Design*. 2nd ed. Prentice Hall, 867 pp.
- Michelson, D. B., 2004: Systematic correction of precipitation gauge observations using analyzed meteorological variables. *J. Hydrol.*, **290**, 161–177.
- Molini, A., L. G. Lanza, and P. La Barbera, 2005: Improving the accuracy of tipping-bucket rain records using disaggregation techniques. *Atmos. Res.*, **77**, 203–217.
- Molteni, F., R. Buizza, T. N. Palmer, and T. Petroliagis, 1996: The ECMWF Ensemble Prediction System: Methodology and validation. *Quart. J. Roy. Meteor. Soc.*, **122**, 73–119.
- Nespor, V., and B. Sevruk, 1999: Estimation of wind-induced error of rainfall gauge measurements using a numerical simulation. *J. Atmos. Oceanic Technol.*, **16**, 450–464.
- Palmer, T. N., 2000: Predicting uncertainty in forecasts of weather and climate. *Rep. Prog. Phys.*, **63** (2), 71–116.
- , 2002: The economic value of ensemble forecasts as a tool for risk assessment: From days to decades. *Quart. J. Roy. Meteor. Soc.*, **128**, 747–774.
- Pappenberger, F., and K. J. Beven, 2006: Ignorance is bliss: Or seven reasons not to use uncertainty analysis. *Water Resour. Res.*, **42**, W05302, doi:10.1029/2005WR004820.
- , —, N. M. Hunter, P. D. Bates, B. T. Gouweleeuw, J. Thielen, and A. P. J. de Roo, 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). *Hydrol. Earth Syst. Sci.*, **9**, 381–393.
- , K. Frodsham, J. Beven, K. Frodsham, R. Romanovicz, and P. Matgen, 2006a: Fuzzy set approach to calibrating distributed flood inundation models using remote sensing observations. *Hydrol. Earth Syst. Sci. Discuss.*, **3**, 2243–2277.

- , P. Matgen, K. J. Beven, J.-B. Henry, L. Pfister, and P. de Fraipont, 2006b: Influence of uncertain boundary conditions and model structure on flood inundation predictions. *Adv. Water Resour.*, **29**, 1430–1449.
- Persson, A., 2001: User guide to ECMWF forecast products. ECMWF Meteorological Bulletin M3.2, 153 pp.
- Rebora, N., L. Ferraris, J. von Hardenberg, and A. Provenzale, 2006: Rainfall downscaling and flood forecasting: A case study in the Mediterranean area. *Nat. Hazards Earth Syst. Sci.*, **6**, 611–619.
- Ren, Z. H., and M. L. Li, 2007: Errors and correction of precipitation measurements in China. *Adv. Atmos. Sci.*, **24**, 449–458.
- Roulston, M. S., and L. A. Smith, 2002: Evaluating probabilistic forecasts using information theory. *Mon. Wea. Rev.*, **130**, 1653–1660.
- Segond, M.-L., 2006: Stochastic modelling of space-time rainfall and the significance of spatial data for flood runoff generation. Ph.D. thesis, Imperial College London, 222 pp.
- Sevruk, B., 1986: Correction of precipitation measurements. *Proc. Int. Workshop on the Correction of Precipitation Measurements*, Zurich, Switzerland, ETH/IAHS/WMO.
- , 1996: Adjustment of tipping-bucket precipitation gauge measurements. *Atmos. Res.*, **42**, 237–246.
- , 2005: Rainfall measurement: Gauges. *Encyclopedia of Hydrological Sciences*, M. G. Anderson and J. J. McDonnell, Eds., Wiley, 529–536.
- , and S. Klemm, 1989: Catalogue of national standard precipitation gauges. WMO Instruments and Observing Methods Rep. 39, 50 pp.
- Shaw, E. M., and P. P. Lynn, 1972: Areal rainfall evaluation using two surface fitting techniques. *Bull. Int. Assoc. Hydrol. Sci.*, **12** (4), 419–433.
- Shepard, D., 1968: A two-dimensional interpolation function for irregularly-spaced data. *Proc. 23rd ACM*, Princeton, New Jersey, Association of Computing Machinery, 517–524.
- Sieck, L. C., S. J. Burges, and M. Steiner, 2007: Challenges in obtaining reliable measurements of point rainfall. *Water Resour. Res.*, **43**, W01420, doi:10.1029/2005WR004519.
- Sivapalan, M., and G. Bloschl, 1998: Transformation of point rainfall to areal rainfall: Intensity–duration–frequency curves. *J. Hydrol.*, **204**, 150–167.
- Skok, G., and T. Vrhovec, 2006: Considerations for interpolating rain gauge precipitation onto a regular grid. *Meteor. Z.*, **15**, 545–550.
- Smith, P. J., 2006: Likelihood measures for environmental modelling. Ph.D. dissertation, Lancaster University, 312 pp.
- Strangeways, I., 2004: Improving precipitation measurement. *Int. J. Climatol.*, **24**, 1443–1460.
- Syed, K. H., D. C. Goodrich, D. E. Myers, and S. Sorooshian, 2003: Spatial characteristics of thunderstorm rainfall fields and their relation to runoff. *J. Hydrol.*, **271**, 1–21.
- Tabios, G. Q., and J. D. Salas, 1985: A comparative analysis of techniques for spatial interpolation of precipitation. *Water Resour. Bull.*, **21**, 365–380.
- Talagrand, O., 1997: Assimilation of observations, an introduction. *J. Meteor. Soc. Japan*, **75**, 191–209.
- Tartaglione, N., S. Mariani, C. Accadia, A. Speranza, and M. Casaioli, 2005: Comparison of rain gauge observations with modeled precipitation over Cyprus using contiguous rain area analysis. *Atmos. Chem. Phys.*, **5**, 2147–2154.
- Thiessen, A. H., 1911: Precipitation averages for large areas. *Mon. Wea. Rev.*, **39**, 1082–1084.
- Tustison, B., D. Harris, and E. Foufoula-Georgiou, 2001: Scale issues in verification of precipitation forecasts. *J. Geophys. Res.*, **106** (D11), 11 775–11 784.
- UNESCO, cited 2006: The Danube catchment. [Available online at http://portal.unesco.org/fr/ev.php-URL_ID=27239&URL_DO=DO_TOPIC&URL_SECTION=201.html.]
- Venugopal, V., S. Basu, and E. Foufoula-Georgiou, 2005: A new metric for comparing precipitation patterns with an application to ensemble forecasts. *J. Geophys. Res.*, **110**, D08111, doi:10.1029/2004JD005395.
- Watson, D. F., 1994: *Contouring: A Guide to the Analysis and Display of Spatial Data*. Computer Methods in the Geosciences, Vol. 10, Pergamon Press, 321 pp.
- Weisheimer, A., L. A. Smith, and K. Judd, 2005: A new view of seasonal forecast skill: Bounding boxes from the DEMETER ensemble forecasts. *Tellus*, **57A**, 265–279.
- WHO, cited 2007: The OFDA/CRED International Disaster Database. World Health Organization Collaborating Centre for Research on the Epidemiology of Disasters. [Available online at www.emdat.be.]
- Willems, P., 2001: A spatial rainfall generator for small spatial scales. *J. Hydrol.*, **252**, 126–144.
- Wilson, L. J., W. R. Burrows, and A. Lanzinger, 1999: A strategy for verification of weather element forecasts from an ensemble prediction system. *Mon. Wea. Rev.*, **127**, 956–970.
- Woodhead, S. P. B., 2007: Bayesian calibration of flood inundation simulators using an observation of flood extent. Ph.D. thesis, University of Bristol, 219 pp.
- Yang, T. Y., 1986: *Finite Element Structural Analysis*. International Series in Civil Engineering and Engineering Mechanics, Prentice-Hall, 543 pp.
- Zhu, Y. J., Z. Toth, R. Wobus, D. Richardson, and K. Mylne, 2002: The economic value of ensemble-based weather forecasts. *Bull. Amer. Meteor. Soc.*, **83**, 73–83.