

This article was downloaded by: [European Commission]

On: 04 November 2011, At: 03:38

Publisher: Taylor & Francis

Informa Ltd Registered in England and Wales Registered Number: 1072954 Registered office: Mortimer House, 37-41 Mortimer Street, London W1T 3JH, UK



## International Journal of River Basin Management

Publication details, including instructions for authors and subscription information:  
<http://www.tandfonline.com/loi/trbm20>

### Improving the evaluation of hydrological multi-model forecast performance in the Upper Danube Catchment

K. Bogner<sup>a</sup>, H. L. Cloke<sup>b</sup>, F. Pappenberger<sup>c</sup>, A. De Roo<sup>d</sup> & J. Thielen<sup>e</sup>

<sup>a</sup> Joint Research Centre, IES, Ispra, Italy

<sup>b</sup> Department of Geography, King's College London, London, UK E-mail:  
hannah.cloke@kcl.ac.uk

<sup>c</sup> European Centre for Medium Range Weather Forecasts, Reading, UK E-mail:  
florian.pappenberger@ecmwf.int

<sup>d</sup> Joint Research Centre, IES, Ispra, Italy E-mail: ad.de-roo@jrc.ec.europa.eu

<sup>e</sup> Joint Research Centre, IES, Ispra, Italy E-mail: jutta.thielen@jrc.ec.europa.eu

Available online: 26 Sep 2011

To cite this article: K. Bogner, H. L. Cloke, F. Pappenberger, A. De Roo & J. Thielen (2011): Improving the evaluation of hydrological multi-model forecast performance in the Upper Danube Catchment, International Journal of River Basin Management, DOI:10.1080/15715124.2011.625359

To link to this article: <http://dx.doi.org/10.1080/15715124.2011.625359>



PLEASE SCROLL DOWN FOR ARTICLE

Full terms and conditions of use: <http://www.tandfonline.com/page/terms-and-conditions>

This article may be used for research, teaching, and private study purposes. Any substantial or systematic reproduction, redistribution, reselling, loan, sub-licensing, systematic supply, or distribution in any form to anyone is expressly forbidden.

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The accuracy of any instructions, formulae, and drug doses should be independently verified with primary sources. The publisher shall not be liable for any loss, actions, claims, proceedings, demand, or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.



Research paper

## Improving the evaluation of hydrological multi-model forecast performance in the Upper Danube Catchment

K. BOGNER, *Joint Research Centre, IES, Ispra, Italy. Email: konrad.bogner@jrc.ec.europa.eu (Author for correspondence)*

H.L. CLOKE, *Department of Geography, King's College London, London, UK. Email: hannah.cloke@kcl.ac.uk*

F. PAPPENBERGER, *European Centre for Medium Range Weather Forecasts, Reading, UK. Email: florian.pappenberger@ecmwf.int*

A. DE ROO, *Joint Research Centre, IES, Ispra, Italy. Email: ad.de-roo@jrc.ec.europa.eu*

J. THIELEN, *Joint Research Centre, IES, Ispra, Italy. Email: jutta.thielen@jrc.ec.europa.eu*

### ABSTRACT

Medium range flood forecasting activities, driven by various meteorological forecasts ranging from high resolution deterministic forecasts to low spatial resolution ensemble prediction systems, share a major challenge in the appropriateness and design of performance measures. In this paper possible limitations of some traditional hydrological and meteorological prediction quality and verification measures are identified. Some simple modifications are applied in order to circumvent the problem of the autocorrelation dominating river discharge time-series and in order to create a benchmark model enabling the decision makers to evaluate the forecast quality and the model quality. Although the performance period is quite short the advantage of a simple cost-loss function as a measure of forecast quality can be demonstrated.

*Keywords:* Floods; forecasting; quality; verification; cost-loss

### 1 Introduction

The European Flood Alert System (EFAS) developed at the Joint Research Centre (JRC) runs in pre-operational mode in order to provide national water authorities with early flood warnings based on medium range weather forecasts (Bartholmes *et al.* 2009, Thielen *et al.* 2009a, 2009b). EFAS uses the hydrological model LISFLOOD, which is a distributed, hydrological rainfall-runoff model. It is a hybrid between a conceptual and a physical rainfall-runoff model designed specifically to simulate the hydrological processes that occur in large catchments (Van Der Knijff *et al.* 2010). The model has been extensively tested and calibrated for various catchments across the globe (Mo *et al.* 2006, Feyen *et al.* 2007, He *et al.* 2009, Thiemiig *et al.* 2010, Bogner and Pappenberger 2011).

The operational set-up of EFAS covers Europe on a 5 km grid and simulates discharges on 6 hourly time-steps. The input

parameters for soil and land use are derived from European databases. Snow-melt is simulated using a degree-day approach, with a correction factor for higher snow-melt rates during rain events. The model parameters that control snowmelt rates, infiltration, overland and river flow, as well as residence time in the soil and subsurface reservoirs, were calibrated with the Shuffled Complex Evolution method developed at the University of Arizona (SCE-UA, Duan and Sorooshian 1992), replacing the original random sampling with a Latin Hypercube sampling scheme to generate the initial population. Following recommendations in Feyen *et al.* (2008), calibration was done in a semi-distributed way, dividing each catchment into sub-catchments (in total 231), based on available station data.

Within the EU Project PREVIEW (prevention, information and early warning) the LISFLOOD model has been calibrated for the Upper Danube Catchment with a 1 km<sup>2</sup> spatial resolution. Various weather forecast products from the European Centre for

Received 19 May 2011. Accepted 16 September 2011.

ISSN 1571-5124 print/ISSN 1814-2060 online  
<http://dx.doi.org/10.1080/15715124.2011.625359>  
<http://www.tandfonline.com>

Medium-Range Weather Forecasts (ECMWF), the German meteorological service (Deutscher Wetterdienst, DWD), the HydroMeteorological Service of the Emilia-Romagna (Agenzia Regionale per la Prevenzione e l'Ambiente - Servizio Idro Meteo, ARPA - SIM) and some experimental results from the Institute for Meteorology and Climate Research (IMK) have been compared. The aim of the PREVIEW project was to examine the added value of an extended 3–10 days flood forecasting system including probabilistic forecasts based on ensembles for the upper-Danube catchment area (upstream Bratislava). More detail about PREVIEW can be found at <http://www.preview-risk.com/>.

One of the major challenges in evaluating ensemble forecasts is the application of appropriate performance measures. Current practice varies and ranges from using measures based on deterministic streamflow prediction to those taken from meteorological forecasting (Cloke and Pappenberger 2009). It is clear that the simple mean squared error (MSE) based performance measures typically used in hydrological modelling suffer from a number of important deficiencies (e.g. discussion in Gupta *et al.* (2009)). They exhibit bias towards high flow events, are sensitive to timing errors, and do not account for autocorrelation and thus they should be used with particular care (McCuen *et al.* 2006, Schaeffli and Gupta 2007), especially in flood forecasting. In addition, typical time series available for evaluation of flood forecasts are nearly always too short (see discussion in Cloke and Pappenberger (2009)). Performance measures designed for different applications are often unfamiliar (Cloke and Pappenberger 2008) or may not be suitable for direct application in flood hydrology. Recognition of the deficiencies of traditional measures has led to the proposal and development of several alternative methodologies (e.g. Seibert 2001, Mathevet *et al.* 2006, Criss and Winston 2008, Weijs *et al.* 2010, Ehret and Zehe 2011, Liu *et al.* 2011).

Appropriate forecast systems should provide adequate forecasts for individual flood events as well as for other flow regimes. Therefore the main objective is to derive performance criteria not only for flood events, but also to evaluate the technical quality of the different forecasting systems continuously within the year. In this study some of the difficulties in evaluating such a complex forecast system with the 'blind' application of widely used forecast performance measures are explored and improvements to the measures suggested. The advantage of a simple cost-loss function as a measure of forecast quality will be demonstrated.

## 2 Data and model setup

Based on the system of available meteorological forecasts ensembles of river discharge series have been generated for the Danube catchment upstream Bratislava. The hydrological year 2002 has been chosen for this study, because two major flood events happened in the Upper Danube area, one event in the springtime triggered mainly by snow-melt processes and one summer flood event

caused by the overlapping of intense convective and long lasting stratiform precipitation. The necessary re-runs of the weather forecasts of this period with the latest available model versions of the different numerical weather prediction (NWP) systems was extremely computationally demanding. Therefore the evaluation time span has had to be restricted to the most important period of 2002, although ideally a longer time series should have been used in order for conclusions to be more statistically robust.

For some selected sub-catchments (Figure 1 and Table 1) the discharge prediction experiment has been performed in two steps. First, the river flow is simulated continuously for the hydrological year 2002 (10 January 2001–30 September 2002) using observed meteorological daily data from synoptic stations (provided by the JRC MARSSTAT Unit) and local (national) high density gauging networks. In a second step the daily output of the simulation run is taken as initial conditions and is combined with the meteorological 5–10 day forecast. Therefore the results of the LISFLOOD model are stored day by day and subsequently used as initial conditions for the forecast run for every day of the performance period taking the meteorological forecasts as input. Finally the observations, the simulations (i.e. the output of LISFLOOD given observed meteorological input) and the forecasts (i.e. the simulated discharge series driven by the meteorological forecasts) are evaluated for the stations Hofkirchen (Danube), Wiblingen (Iller), Schaerding (Inn) and for the outlet at Bratislava (Danube). The different temporal and spatial resolutions of the forecast products are summarized in Table 2. In order to have at least two different forecast systems to compare (ECMWF and DWD), the analysed forecast horizon was limited to 7 days, i.e. the maximum length of the global model of DWD (DWD-GME) forecasts.

## 3 Description of performance measures

For the evaluation of the hydrological forecast quality many different measures have been applied, but not all of them are suitable without further modifications. The aim of this paper is to outline some limitations of the most popular performance measures and to show some simple ways in order to make them applicable. Due to the limited data sets available and their difference in length, the evaluation study will be divided into two parts, i.e. for complete hydrological years ('year') and for individual flood events ('event'). For the 'year' forecasts classical deterministic performance measures for continuous variables can be calculated such as the mean error (ME), the root mean squared error (RMSE), the Nash–Sutcliffe measure (Nash and Sutcliffe 1970) and the persistence efficiency measure (PEM). For the verification of meteorological forecasts, continuous observations are usually converted into binary series. In this paper the economic value has been chosen as a hydrological verification measure for the binary discharge series (i.e. the series are divided into two categories, namely discharge above and below a threshold). For the evaluation of 'events'

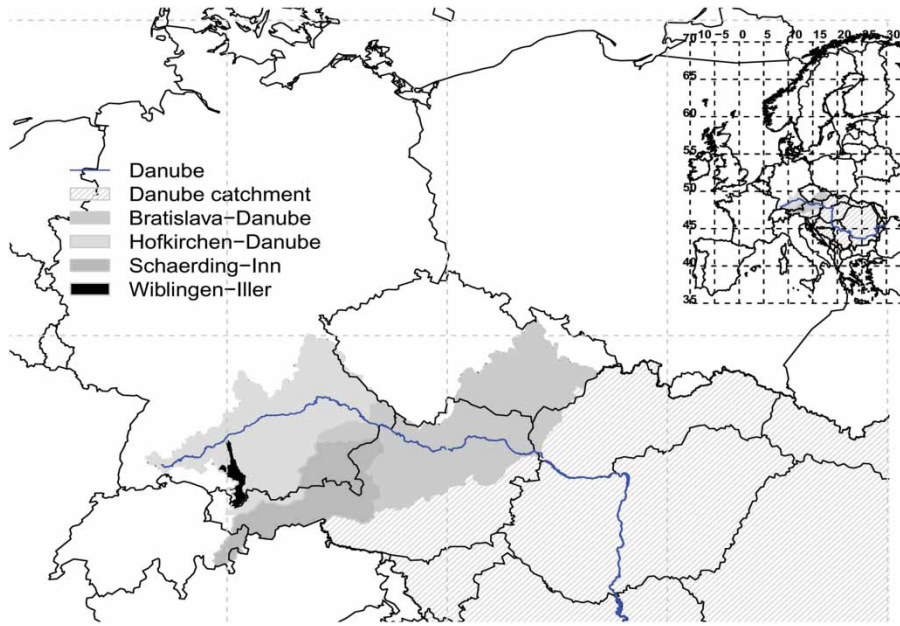


Figure 1 Subcatchments of the Upper Danube basin upstream Bratislava (SK)

Table 1 Gauging stations

ID	Station name	Area (km <sup>2</sup> )
1	Bratislava	131,000
2	Hofkirchen	47,500
3	Schaerding	25,600
4	Wiblingen	2000

and of the ‘year’ forecast a simple cost-loss function can be applied in order to compare the different forecast products (deterministic and probabilistic) and to estimate the expected costs of the forecast as a measure of forecast quality.

### 3.1 Continuous measures

The ME, the RMSE and the mean absolute error (MAE) are all standard performance measures and so their equations are omitted from this paper. The PEM is defined as a normalized statistic that quantifies the relative magnitude of the residual variance to the variance of the errors obtained by the use of a simple

persistence model (Gupta *et al.* 1999, Seibert 2001). The power of PEM is derived from its comparison of model performance with a simple benchmark model:

$$PEM = 1 - \left[ \frac{\sum_{i=1}^n (Y_i^{obs} - Y_i^{sim})^2}{\sum_{i=1}^n (Y_i^{obs} - Y_i^{bench})^2} \right], \quad (1)$$

where  $Y_i^{obs}$  is the  $i$ th observation,  $Y_i^{sim}$  is the  $i$ th forecasted value and  $Y_i^{bench}$  is the  $i$ th value of a benchmark model. The Nash–Sutcliffe efficiency is defined the same way, the only difference is the definition of the benchmark. For the Nash–Sutcliffe measure the benchmark model represents the climatology given by the mean observed discharge ( $(1/n) \sum_{i=1}^n Y_i^{obs}$ ), which neglects the seasonal effects of the discharge data. For the PEM the forecast precipitation has been set to zero as a benchmark, thus the PEM shows directly the effect of the meteorological forecast.

Using the PEM and ME in combination has significant advantages. The ME measures the bias and is defined as the ratio between forecast mean and observed mean, thus a negative or positive ME indicates under- or over prediction of the simulated values, respectively. From the combination of the PEM and the

Table 2 Available weather forecast products

Provider	Product	#	Forecast period	Resolution (km)	Forecast horizon (days)
ECMWF	EPS	51	10 January 2001–30 September 2002	80	10
	VAREPS	51	20 July 2002–20 August 2002	40	10
	Deterministic	1	10 January 2001–30 September 2002	40	15
DWD	GME	1	10 January 2001–30 September 2002	40	7
	COSMO-EU	1	20 July 2002–20 August 2002	7	3
ARPA-SIM	COSMO-LEPS	11	20 July 2002–20 August 2002	10	5.5
IMK	High resolution	1	3 August 2002–13 August 2002	1	1.5

ME conclusions can be drawn about the hydrological model quality and the meteorological forecast quality. For the range of lead times with increasing PEM values, the bias of the forecast is dominated by the hydrological model error, whereas for decreasing PEM ranges the bias given by the ME is caused by the meteorological forecast in addition to the hydrological model prediction error.

### 3.2 Categorical measures

For the purpose of verifying meteorological forecasts, categorical measures, like the Brier score (Jolliffe and Stephenson 2003), are widely used. The estimation of an economic value has been proposed for the evaluation of meteorological forecasts (Richardson 2000). The economic value is a measure of the reduction in expected expense  $E$  given a particular forecast. The relative economic value  $V$  compares this reduction in expense with the reduction due to a perfect forecast as the ratio:

$$V = \frac{E_{\text{climate}} - E_{\text{forecast}}}{E_{\text{climate}} - E_{\text{perfect}}} \quad (2)$$

In order to apply such a categorical measure for the continuous-valued hydrological forecasts of river flow series, the time series have to be converted first into a binary series by the use of a threshold filter resulting in a series of zeros, for discharge values below a certain threshold  $u$ , and a series of ones, for discharge values greater than a threshold  $u$  (see Figure 2). However applying categorical measures to these binary data directly would give wrong and misleading results, simply because of the fact that hydrological data are strongly autocorrelated. In order to avoid such autocorrelation problems, flood events are typically separated through the application of some subjective criteria to build clusters of (temporally) independent events. The same approach has been applied here by setting different minimum values of inter-event times. For example, at Hofkirchen (Figure 3) the event separation criterion was set to a minimum of 8 subsequent days below the threshold.

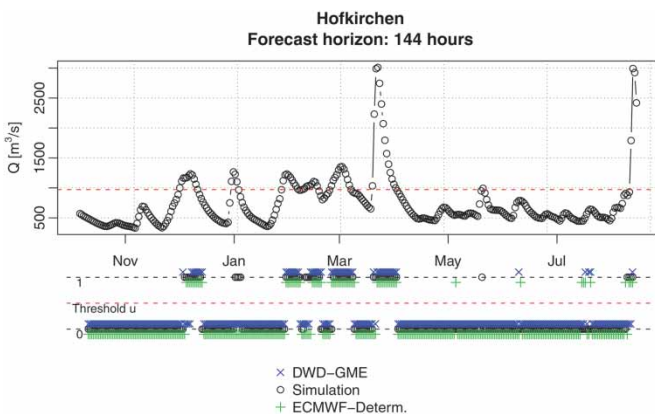


Figure 2 Discharge converted to binary series for the 6 days ahead predictions for the DWD-GME and the ECMWF deterministic forecast in comparison to the reference discharge (i.e. the simulation)

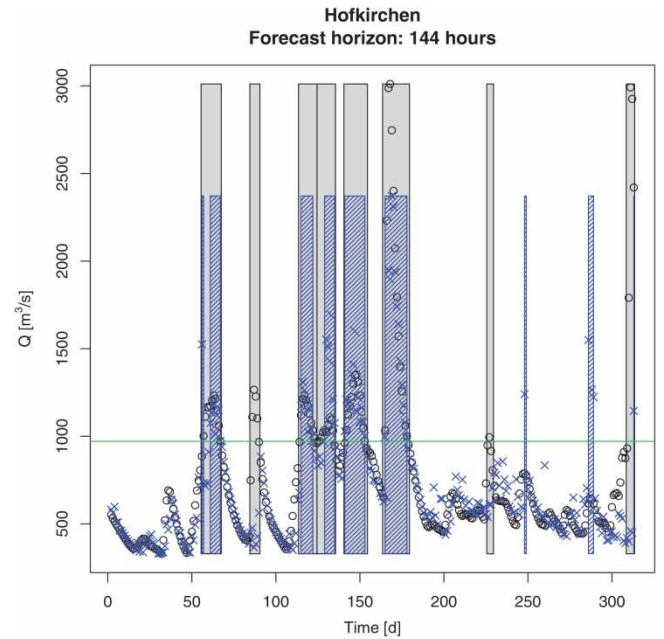


Figure 3 Clusters of events above threshold  $u$  separated by the criteria of temporal independence between events (with grey rectangular representing the observed clusters and the blue bars indicating the forecast exceeding the threshold). For the station Hofkirchen it is assumed that two events are independent, if the inter-event time is greater than 8 days

### 3.3 Cost loss

Since only some weather forecast data are available for the August 2002 flood event, the conversion detailed in the previous section would reduce the amount of information beyond a useful limit. Therefore a performance method based on the cost-loss function proposed by Laio and Tamea (2007) has been applied. The method is beneficial for several reasons: It can be applied to continuous variables, that means no loss of information. Furthermore, it is not necessary to choose criteria for defining independence between events, which would be impossible to apply uniformly for different catchments. In addition, deterministic and probabilistic forecasts can be compared directly, and also it combines different performance measures, like the MSE, ME and the ranked probability score (RPS). The following description of the cost-loss function follows closely the work of Laio and Tamea (2007).

The cost-loss function is the sum of the:

- (1) cost term  $C(\chi)$ , which is an increasing function of a design value fixed by a decision maker for a hypothetical event  $\chi$  and
- (2) the loss term  $L(x, \chi)$ , which is an increasing function of  $(x - \chi)$  when the observed event  $x$  is greater than the hypothetical event  $\chi$ ,  $x > \chi$ :

$$CL(x, \chi) = C(\chi) + L(x, \chi) \quad (3)$$

If the prediction is deterministic,  $\chi$  is equal to the point forecast,  $\chi = (\tilde{x})$ ; if the prediction is probabilistic, then the  $c$  value can be chosen among the forecast outcomes of the ensemble. Thus the decision maker is able to calculate the expected expenses  $\overline{CL}(\chi)$  according to Eq. 4 and to take the decision

$c^*$  that minimizes  $\overline{CL}(\chi)$ .

$$\overline{CL}(\chi) = \int_{\text{all } \hat{x}} CL(\hat{x}, \chi) p(\hat{x}) d\hat{x}. \quad (4)$$

The probabilistic forecast  $p(\hat{x})$  given in Eq. 4 can be used by the decision maker to represent the probability distribution of the future events,  $f(x)$ . In order to compare the deterministic and probabilistic forecasts based upon their operational value, a very simple cost-loss function is applied.  $C(\chi)$  should be a linear function,  $C(\chi) = c \cdot \chi$ , where  $c$  is a constant, and  $L(x, \chi)$  is a stepwise linear function,  $L(x, \chi) = H(x - \chi) \cdot l \cdot (x - \chi)$ .  $l$  is a constant ( $l > c$ ), and  $H(x - \chi)$  is the heavy-side function. Setting  $\xi = c/l < 1$  and after some transformations the following cost-loss function is obtained (see Figure 4):

$$\rho_{\xi}(x, \chi) = |\chi - x| + 2(\xi - 0.5)(\chi - x). \quad (5)$$

Of course it is questionable if this linear function approach is realistic, since the losses will be rather nonlinear, but for the purpose of comparing the different forecast systems this simplification should be adequate: On the one hand the same simple method will be applied for all systems in the same way and therefore the comparison should be unbiased; on the other hand no data are available for estimating a more realistic loss function. Given this loss function the optimal design value  $\chi^*$  can be found. By taking the expected value of Eq. 5, one obtains

$$\bar{\rho}_{\xi}(\chi) = 2\xi \left( \chi - \int_{\text{all } \hat{x}} \hat{x} p(\hat{x}) d\hat{x} \right) + 2 \int_{\chi}^{\infty} (\hat{x} - \chi) p(\hat{x}) d\hat{x}, \quad (6)$$

whose derivative with respect to  $\chi$ , equated to zero, provides the optimal decision value  $\chi^*$

$$P(\chi^*) = 1 - \xi \rightarrow \chi^* = P^{-1}(1 - \xi) \quad (7)$$

that depends only on the cumulative distribution function of the forecasts,  $P(\bullet)$ , and on the cost-loss ratio  $\xi < 1$ . According to

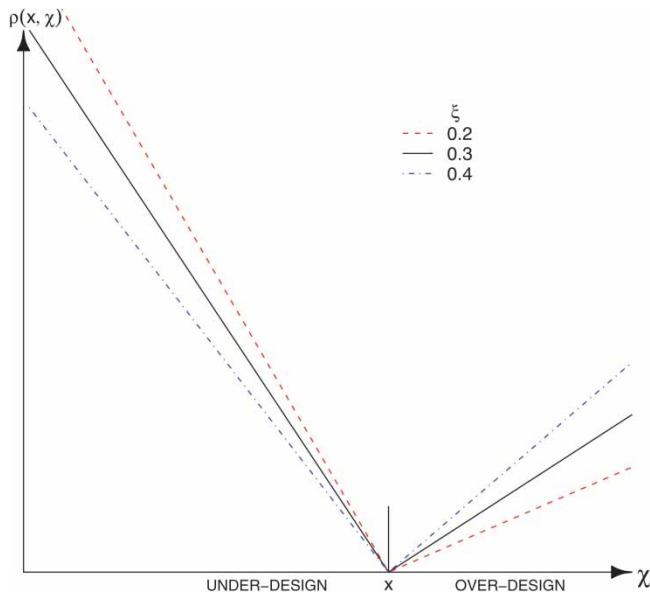


Figure 4 Cost-loss function given in Eq. 5 for different cost-loss ratios  $\xi = c/l$

Eq. 5 the operational value of different predictions can be found from the  $\rho(x, \chi^*)$  values averaged over  $n$  forecast discharge values for each forecast horizon:

$$\begin{aligned} EC(\xi) &= \frac{1}{n} \sum_{i=1}^n \rho_{\xi}(x_i, \chi_i^*) \\ &= \frac{1}{n} \sum_{i=1}^n \{ |P_i^{-1}(1 - \xi) - x_i| + 2(\xi - 0.5)(P_i^{-1}(1 - \xi) - x_i) \} \end{aligned} \quad (8)$$

When the prediction is deterministic, then  $\chi^* = \tilde{x}$  for any  $\xi$  and  $P(x) = H(x - \tilde{x})$ , therefore Eq. 8 will read as

$$EC_{\text{det}}(\xi) = \frac{1}{n} \sum_{i=1}^n \{ |\tilde{x}_i - x_i| + 2(\xi - 0.5)(\tilde{x}_i - x_i) \}. \quad (9)$$

The slope of the deterministic  $EC(\xi)$  curve is two times the bias of the prediction (ME) indicating overestimation (positive bias) or underestimation (negative bias), and its intercept with the  $\xi = 0.5$  vertical line is the MAE of the forecast. Furthermore the average of the  $EC(\xi)$  over the possible  $\xi$  values corresponds to the  $\overline{CRPS}$ , the mean continuous RPS (Hersbach 2000).

The lower the expected cost value  $EC(\xi)$ , the more valuable the forecast. The cost-loss ratios of  $\xi < 0.5$  represent situations where the losses are very relevant compared to the costs of precautionary actions, which is usually the case in a flooding situation.

The rescaling of the cost-loss ratio plays an important role. Here we rescale the  $EC(\xi)$  values in a different way from that proposed by Laio and Tamea (2007), who rescaled the curves with respect to the cost of a mean-value deterministic prediction. Instead of taking the mean of the observed discharge data, it seems to be more appropriate to rescale the curve using zero precipitation predictions, as was done for the PEM.

#### 4 Evaluation of the forecast performance

According to the difference in the length of the provided data sets different forecast evaluation measures will be applied. For the complete hydrological year 2002 the meteorological forecasts of DWD-GME, ECMWF deterministic and the Ensemble Prediction System (EPS) were available ('year'), whereas the forecast from COSMO-EU, COSMO-LEPS and VAREPS have been rerun only for the August 2002 flood event ('event'). For the 'year' forecasts classical performance measures for continuous variables can be calculated (see previous section). It is important to evaluate the impact of autocorrelation in this context. Different approaches have been applied to account for the problem of serial correlation in hydrology. For example, in Beven (2001), a likelihood measure for the evaluation of hydrological models is shown including a Gaussian autoregressive error model. In general the first order auto-regressive model (AR (1)) is the most widely used model (see, for example, Yang *et al.* (2007) and Mantovan and Todini (2006), proposing a multi-normal AR (1) model as likelihood function). Another

approach is to adjust the sample size for autocorrelation by estimating a so-called ‘effective sample size’. More on the adjustment can be found elsewhere (Dawdy and Matalas 1964, WMO 1966, Thiébaux and Zwiers 1984, McMillan *et al.* 2010).

In this study the ‘memory effects’ of hydrological time-series will not be handled directly either by error models or by the adjustment of the sample size, but indirectly by filtering methods and by simple modifications of the performance measures like the simple cost-loss function approach, that will be applied for the evaluation of the flood event (and for the ‘year’ forecast).

#### 4.1 Continuous measures

In Figure 5 the PEM and Nash–Sutcliffe efficiency are shown for the hydrological year 2002 at station Hofkirchen. The PEM increases between a lead time of 1 and 3 days for the deterministic DWD and the median EPS forecast. The PEM also increases between a lead time of 1 and 4 days for the deterministic ECMWF forecast. Thus this measure of forecast quality shows that for the very first lead times the gain from using the meteorological forecast as input is limited. Furthermore the peak in the PEM curve at lead time between 3 and 4 days indicates that the time of concentration of the runoff is about 3 days, which is in accordance with observations. The time of concentration is defined as the amount of time for the whole watershed to contribute to the outflow or the amount of time for the water to reach the outlet from the furthest point from the outlet. Runoff is assumed to reach a peak at the time of concentration (Spellman and Whiting 2004). Therefore the quality of the forecast at this station is dominated by routing effects for the forecast horizons of 1 and 2 days. From day 3 onward the effect of the routing decreases and the quality of the precipitation forecast dominates the PEM. In Tables 3 and 4 the performance results for the four

catchments are shown, applying the measure of ME and PEM. For the alpine Inn river (Schaerding) the time of concentration is quite small, which is also indicated by the maximum PEM value being between lead time of 1 and 2 days, whereas the time of routing of the flood peak downstream to Bratislava will take several days and the PEM does not reach its maximum within the analysed 7 days of lead time. For the station Wiblingen the results are not conclusive at all, since the catchment size is on the lower limit of applicability with just a few meteorological forecast grid cells covering the catchment.

Additionally in Table 5 the MAE and the RMSE, which are commonly understood statistical measures, are provided for comparison. However, analysing the quality of a forecast based on such scalar measures like the ME, MAE and RMSE separately should be avoided, since each of these measures has drawbacks and the interpretation could be misleading. In particular, the ME or bias says nothing about the accuracy of forecasts. It is obvious that a prediction system forecasting 25 times too high water levels and 25 times too low values will get the same ME as a prediction system making 50 forecasts that match the observations exactly. The MAE and the RMSE circumvent this problem, but show different sensitivity to large errors, thus the RMSE rewards the more consistent forecast. Especially the peak flow prediction will dominate the RMSE and a miss or time shift of the forecasted peak will increase the RMSE drastically, even though the compared forecast system could have the same MAE. Therefore it is necessary to look at the combination of the RMSE and MAE, which can provide more hydrologically relevant information. For example, at the station Hofkirchen (Table 5) it is interesting to see that the EPS-median forecast is better (i.e. lower) than the deterministic forecasts (DWD and ECMWF) with respect to the MAE from lead time of 2–7 days, whereas the RMSE of the EPS median is only slightly better (i.e. lower) for lead time of 2–3 days. That means that for longer lead times the peaks forecasted by EPS-median are probably too smooth, whereas the EPS-median forecast is better during mean flow conditions.

#### 4.2 Categorical measure

In this section the forecast quality based on a categorical time series with and without clustering is evaluated. The difference between the application of performance measures to the binary series (1) directly by neglecting the autocorrelations and (2) separating into clusters, can be demonstrated when calculating the relative economic value (see, for example, Roulin 2007, Buizza 2008). In Figure 6 it can be seen how the relative economic value  $V$  decreases for two different forecasts (deterministic DWD-GME and ECMWF) and how it is shifted to the lower cost-loss ratios by eliminating the autocorrelation through partitioning the observed and forecasted events into clusters that exceed specific thresholds. Since small values of the cost-loss ratio indicate that the costs to prepare are small in relation to the losses, a decision maker will benefit by taking actions even

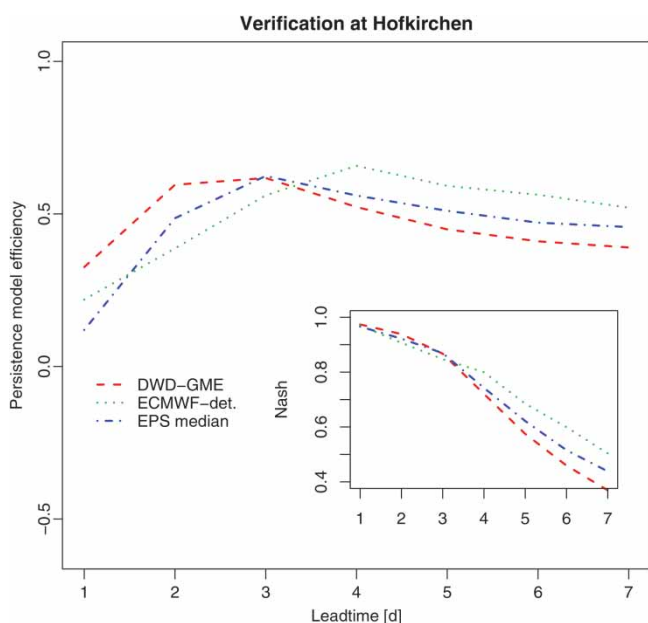


Figure 5 Persistence and Nash–Sutcliffe efficiency measure for the hydrological year 2002 at station Hofkirchen

Table 3 Mean error

Catchment	Forecast	Lead time (days)						
		1	2	3	4	5	6	7
Bratislava	DWD-GME	-261.96	-293.06	-334.75	-393.77	-447.57	-471.78	-479.01
	ECMWF-det.	-248.90	-268.00	-299.07	-341.92	-379.17	-385.42	-383.59
	EPS-med.	-260.00	-288.65	-329.65	-381.35	-423.06	-436.61	-437.75
Hofkirchen	DWD-GME	6.11	2.56	-3.14	-16.45	-33.45	-49.04	-64.58
	ECMWF-det.	6.02	8.86	9.32	0.68	-15.10	-30.65	-45.63
	EPS-med.	3.76	-3.16	-12.08	-26.09	-41.49	-56.21	-69.26
Wiblingen	DWD-GME	2.01	1.16	-1.87	-5.51	-8.96	-11.48	-12.60
	ECMWF-det.	2.70	2.36	1.54	-3.62	-6.94	-8.55	-9.77
	EPS-med.	1.40	0.19	-3.35	-7.44	-10.88	-13.85	-16.06
Schaerding	DWD-GME	-47.13	-49.34	-60.23	-70.22	-74.66	-74.28	-75.89
	ECMWF-det.	-45.35	-31.06	-31.56	-35.74	-32.96	-24.55	-25.10
	EPS-med.	-50.29	-48.69	-40.91	-33.04	-26.39	-17.04	-18.03

Table 4 Persistence efficiency measure

Catchment	Forecast	Lead time (days)						
		1	2	3	4	5	6	7
Bratislava	DWD-GME	-1.50	-0.40	-0.12	-0.05	-0.01	0.04	0.09
	ECMWF-det.	-1.08	0.12	0.29	0.29	0.30	0.28	0.29
	EPS-med.	-1.28	-0.09	0.10	0.12	0.14	0.18	0.22
Hofkirchen	DWD-GME	0.33	0.60	0.62	0.52	0.45	0.41	0.39
	ECMWF-det.	0.22	0.39	0.56	0.66	0.59	0.56	0.52
	EPS-med.	0.12	0.49	0.62	0.56	0.51	0.47	0.46
Wiblingen	DWD-GME	-0.24	-0.09	-0.13	-0.01	0.08	0.12	0.13
	ECMWF-det.	-0.35	-0.22	0.13	0.58	0.45	0.42	0.35
	EPS-med.	-0.22	0.12	0.39	0.36	0.33	0.31	0.31
Schaerding	DWD-GME	-0.63	0.10	0.05	0.02	0.02	0.01	-0.01
	ECMWF-det.	0.11	0.51	0.44	0.36	0.32	0.30	0.31
	EPS-med.	-0.21	0.38	0.29	0.21	0.17	0.17	0.19

Table 5 Performance measures for Hofkirchen

Forecast	Measure	Lead time (days)						
		1	2	3	4	5	6	7
DWD-GME	MAE	31.935	43.554	63.109	88.980	116.328	139.269	156.117
	RMSE	7.920	10.686	13.468	15.231	16.774	17.873	18.687
	ME	6.113	2.559	-3.145	-16.449	-33.452	-49.040	-64.585
	Nash	0.974	0.939	0.867	0.720	0.575	0.459	0.369
	PME	0.326	0.596	0.618	0.523	0.449	0.410	0.390
ECMWF-det.	MAE	31.975	46.808	59.847	80.128	101.944	121.046	139.530
	RMSE	8.128	11.635	12.544	13.452	15.415	16.585	17.604
	ME	6.018	8.860	9.322	0.683	-15.105	-30.647	-45.634
	Nash	0.970	0.907	0.847	0.800	0.685	0.599	0.504
	PME	0.219	0.387	0.561	0.658	0.592	0.563	0.520
EPS-med.	MAE	33.851	43.925	58.480	79.221	100.820	120.991	136.912
	RMSE	8.392	10.393	12.307	14.662	16.291	17.459	18.192
	ME	3.758	-3.161	-12.082	-26.094	-41.487	-56.207	-69.261
	Nash	0.966	0.922	0.869	0.742	0.622	0.515	0.438
	PME	0.120	0.486	0.624	0.560	0.510	0.471	0.457

when the forecast probability is low. That means that for the example shown in Figure 6 (at the station Hofkirchen), in order to achieve maximum economic gain from the forecast the effects of autocorrelation are important to consider when setting reaction thresholds. A decision-maker should therefore react to lower probabilities of exceedance than calculated (discounting the computed ‘optimum’) if autocorrelation effects are neglected. However, determination of the optimal discount is more difficult than taking account of the autocorrelation a priori.

### 4.3 Cost loss

The concept of cost loss is for many hydrological forecasters very novel, thus we would like to give some additional guidance in this paper. Assuming that the cost-loss ratio is 0.1 indicates it would cost 1M to prevent a 10M loss (a ratio which is quite reasonable for flooding situations). Looking at the different expected cost values  $EC(\xi)$  for this ratio of  $\xi = 0.1$  (shown as dashed vertical lines in Figure 7) the following quantities can be evaluated:

#### 4.3.1 Probabilistic versus deterministic forecast quality

In all four graphs, the probabilistic forecast outperforms the deterministic forecast having lower expected costs. For the flood event (Figure 7(a) and (c)) the COSMO-LEPS forecast would always be the most valuable one having by far the lowest expected costs, whereas for the continuous year 2002 (Figure 7(b) and (d)) the EPS forecast gives the lowest expected costs. Only for the 3 days ahead forecast the deterministic forecast from ECMWF is better than the VAREPS forecast for the flood event (Figure 7(b)).

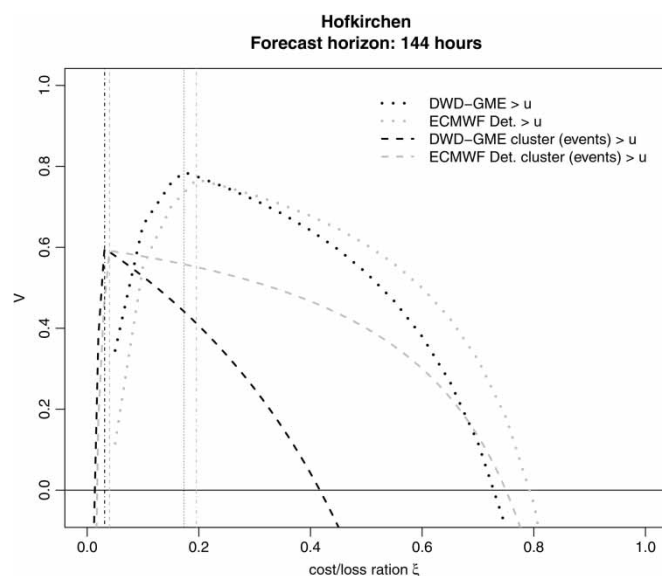


Figure 6 Economic value for the DWD-GME and for the deterministic ECMWF forecast based on (1) all values above the threshold  $u$  and (2) on clusters of events only

#### 4.3.2 Properties of deterministic forecasts

From the comparison of the different deterministic forecast systems for the flood event, one can see how the positive effect of higher *spatial resolution* vanishes with an increase of lead time. For the 2 days ahead predictions the COSMO-EU and IMK forecast have lower expected costs than the deterministic ECMWF one, and the reverse is true for the 3 days ahead predictions.

It is also interesting to look at the *slope of the deterministic forecasts* (i.e. the bias of the prediction). Although all deterministic systems underestimate the flood event indicated by the negative slope (Figure 7(a) and (c)), the slope changes between the different systems and lead times for the continuous year. For instance, the DWD-GME forecast at a lead time of 2 days is slightly overestimating, but it is slightly underestimating in the 3 days ahead predictions, in comparison to the deterministic ECMWF forecast, which is overestimating both lead times significantly. The MAE, given by the intercept with the  $\xi = 0.5$  vertical line, also changes between the two forecast horizons. For the lead time of 2 days the MAE is smaller for the DWD-GME forecast, whereas the MAE for the deterministic ECMWF forecast is smaller for 3 days lead time.

Besides the comparison of multi-model deterministic and probabilistic forecast systems, different catchments have been analysed and the possible impact of re-scaling has been considered.

#### 4.3.3 Impact of catchment characteristics and re-scaling

In Figures 8–10 some results for the remaining catchments are shown. The forecast horizons less than the time of concentration of the catchment have been excluded, thus the results with rescaled  $EC(\xi)$  values greater than 1 will not be shown in the next diagrams. In the left part of the diagrams ( $\xi < 0.5$ ) the probabilistic methods outperform the deterministic ones in almost all catchments and forecast horizons, demonstrating the higher operational value of the probabilistic forecast.

The August 2002 flood peak took more than 6 days to reach the outlet at Bratislava, therefore all  $EC(\xi)$  values for this event were greater than one and are not shown. The IMK forecasts are difficult to verify, because of the very limited amount of forecast data available for just a few days in August 2002. Furthermore the effects of the mismatch of temporal and spatial scales between the meteorological and hydrological model are hard to evaluate and longer time series would be necessary to show possible improvements in the initial forecast time steps taking such high resolution models as input.

Re-scaling of the cost-loss function to represent skill with respect to a more naive forecast is important to evaluate the value of an individual analysis. The approach of Laio and Tamea (2007) would result in rescaled  $EC(\xi)$  values, which are lower than 1 for all catchments and forecast horizons, indicating the improved quality of the forecast discharge with respect to the climatological (respectively mean observed) prediction. Whereas the rescaling with respect to zero precipitation

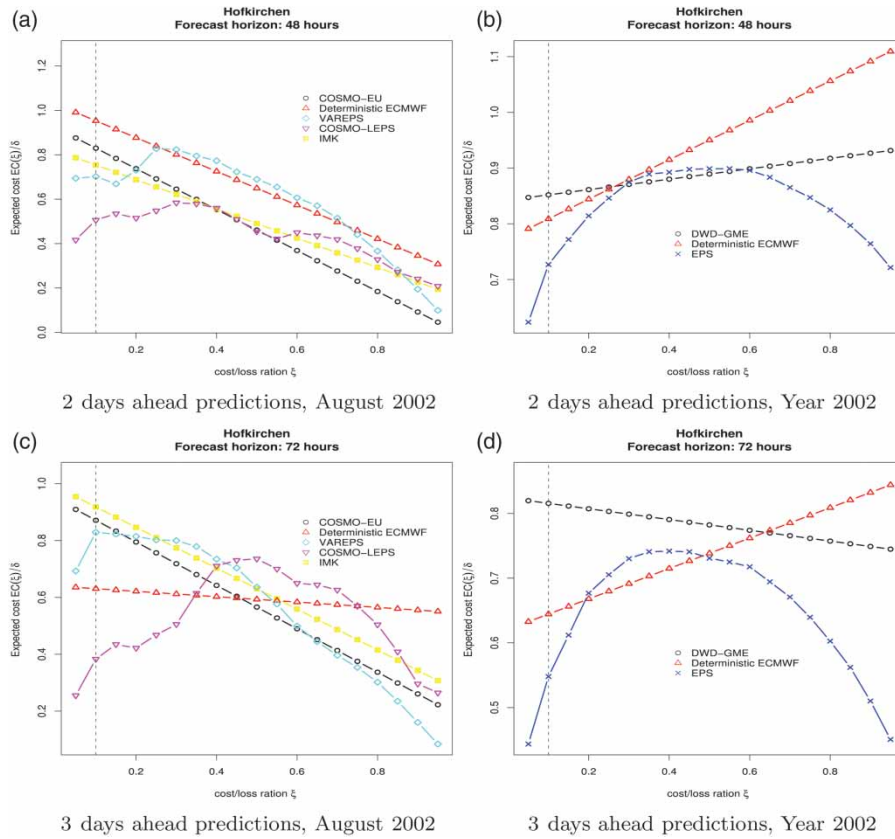


Figure 7 Expected cost rescaled by  $\delta =$  zero precipitation prediction for the 2 days ahead predictions (a) for the August 2002 event and (b) for the hydrological year 2002; 3 days ahead predictions (c) for the August 2002 event and (d) for the hydrological year 2002 at the station Hofkirchen

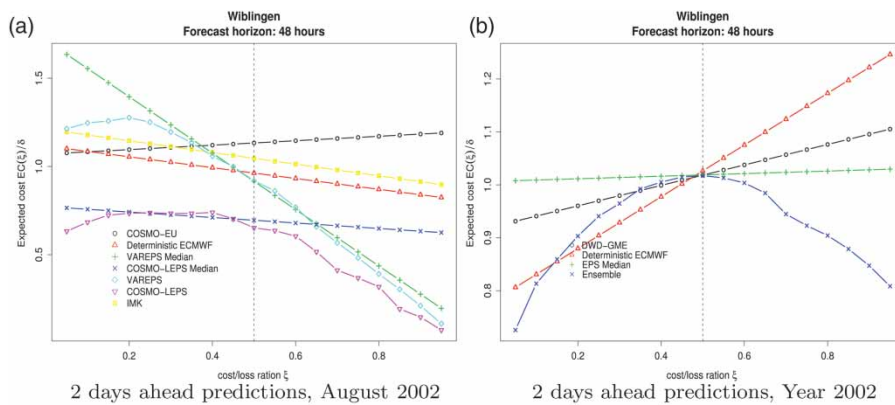


Figure 8 Expected cost rescaled by  $\delta =$  zero precipitation prediction for the 2 days ahead predictions for August 2002 event (a) and for the hydrological year 2002 (b) for Wiblingen

predictions, results in  $EC(\xi)$  values, which must be lower than 1, only if the forecast horizon is greater than the time of concentration of the catchment. This allows for a more hydrologically relevant evaluation. The results demonstrating this impact of re-scaling are not shown for reasons of brevity.

In summary, this prediction experiment has demonstrated the power of the novel cost-loss function approach as a versatile analysis tool which is appropriate for the evaluation and comparison of multi-model deterministic and probabilistic forecast systems. Through this simple method the higher operational value of probabilistic forecast has been shown.

### 5 Discussion and conclusions

Cloke and Pappenberger (2009) highlighted several weaknesses in current practice of NWP driven flood forecasting, in particular those relevant to the evaluation of forecasts and suitable performance measures. The methods presented in this paper seek to address some of these weaknesses. In this paper the evaluation of the hydrological forecast system has been divided into two parts according to the difference in the length of the provided data sets. Classical performance measures for continuous variables have been calculated for the continuous hydrological

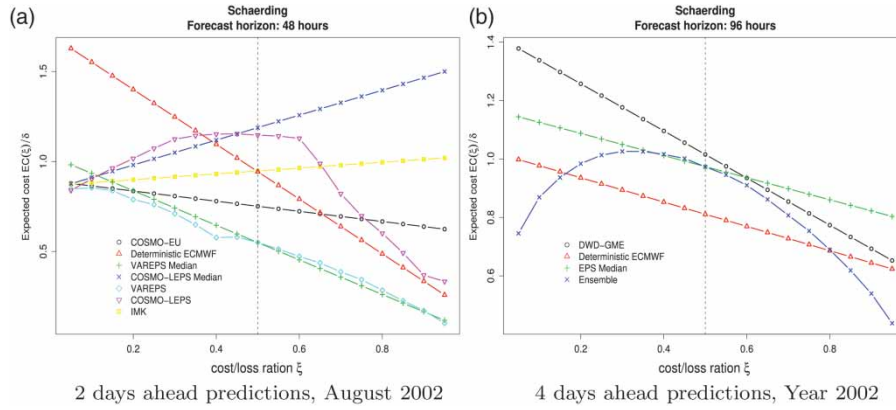


Figure 9 Expected cost rescaled by  $\delta =$  zero precipitation prediction for the 2 days ahead predictions for August 2002 event (a) and for 4 days ahead prediction for the hydrological year 2002 (b) for Schaerding

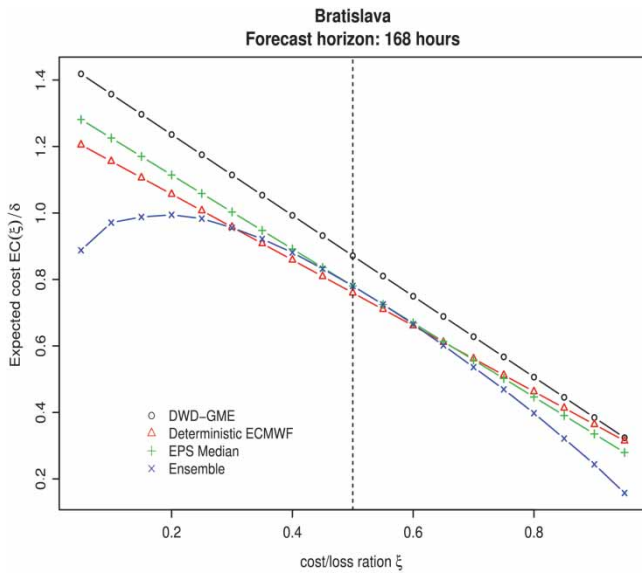


Figure 10 Expected cost rescaled by  $\delta =$  zero precipitation prediction for the 7 days ahead predictions for the hydrological year 2002 for Bratislava

year, like the ME, RMSE and a modified Nash–Sutcliffe coefficient. Depending on the size and the hydrological characteristics of the catchments, at each forecast horizon a different forecast product shows better results. For example, at the station Hofkirchen the DWD-GME is slightly better regarding the ME for the forecast horizon of 2 and 3 days, whereas from day 4 onward the deterministic ECMWF show better results. For the evaluation of the flood event of August 2002 (and also the continuous year) a simple cost-loss function has been tested in order to compare the different forecast products (deterministic and probabilistic) and to estimate the economic value of the forecast. The results of this performance method show that the probabilistic methods outperform the deterministic ones in almost all catchments and forecast horizons, demonstrating the higher operational value of the probabilistic forecast. Cost-loss inherently includes an indication of false alarm rate and provides a quantitative assessment of the benefits of probabilistic forecasts. In particular the COSMO-LEPS forecast clearly indicates this

added value of using probabilistic forecasts. To date, this is some of the most robust evidence presented for the benefits of ensemble forecasting, and acts to support the current uptake of hydrological ensemble prediction systems in flood forecasting (citetcloketal09). Evaluating the quality of a flood forecasting system is always difficult as usually event frequency is very low and results suffer from a lack statistical significance. The evaluation of longer time periods allows for statistically more robust results, but with less significance for the performance of a system with respect to extreme events. This is also valid in this paper. Therefore, we contrast a longer evaluation period with a shorter flood specific one. Nevertheless, all the calculated measures of performance that we have presented should be seen indicative and cannot be taken as absolute truth for these reasons.

### Acknowledgements

This work was funded by the EC PREVIEW (FP6 – Work Package: Plain Floods) programme (<http://www.preview-risk.com>). The authors wish to thank the Deutsche Wetterdienst, the ECMWF, the Bavarian Environment Agency, and the JRC’s Institute for Protection and Security of the Citizen (IPSC) for data and information. Finally the authors gratefully acknowledge the support of all staff of the JRC’s Institute for Environment and Sustainability (IES) – Land Management and Natural Hazards Unit – FLOODS Action.

### Nomenclature

- $Y_i^{\text{bench}}$   $i$ th value of the benchmark model
- $Y_i^{\text{sim}}$   $i$ th forecast
- $Y_i^{\text{obs}}$   $i$ th observation
- $V$  relative economic value
- $E_{\text{climate}}$  expected expense of a user of climatological information
- $E_{\text{forecast}}$  expected expense of a user of a forecast system
- $E_{\text{perfect}}$  expected expense of a user of a perfect forecast system
- $u$  threshold for separating between events and non-events

$x$	observed event
$\tilde{x}$	point forecast
$\chi$	hypothetical event (design value)
$\chi^*$	optimal design value, which minimizes the expected expenses
$C(\chi)$	cost as a function of the hypothetical event $\chi$
$L(x, \chi)$	loss as a function of the observed event $x$ and the design value $\chi$
$\overline{CL}(\chi)$	expected expenses
$p(\tilde{x})$	probabilistic forecast
$c$	constant in linear cost function
$l$	constant in stepwise linear loss function
$H(x - \chi)$	heavy-side function
$\xi$	cost-loss ratio ( $c/l < 1$ )
$\rho_{\xi}(x, \chi)$	cost-loss function
$P(\cdot)$	cumulative distribution function of the forecasts
$EC(\xi)$	expected cost

## References

- Bartholmes, J.C., *et al.*, 2009. The European Flood Alert System EFAS part 2: statistical skill assessment of probabilistic and deterministic operational forecasts. *Hydrology and Earth System Sciences*, 130 (2), 141–153.
- Beven, K.J., 2001. *Rainfall-runoff modelling: the primer*. New York: John Wiley & Sons.
- Bogner, K. and Pappenberger, F., 2011. Multiscale error analysis, correction, and predictive uncertainty estimation in a flood forecasting system. *Water Resources Research*, 470 (7), W07524. doi:10.1029/2010WR009137.
- Buizza, R., 2008. The value of probabilistic prediction. *Atmospheric Science Letters*, 9, 36–42. doi:10.1002/asl.170.
- Cloke, H.L. and Pappenberger, F., 2008. Evaluating forecasts of extreme events for hydrological applications: an approach for screening unfamiliar performance measures. *Meteorological Applications*, 150 (1), 181–197.
- Cloke, H.L. and Pappenberger, F., 2009. Ensemble flood forecasting: a review. *Journal of Hydrology*, 3750 (3–4), 613–626.
- Criss, R.E. and Winston, W.E., 2008. Do Nash values have value? Discussion and alternate proposals. *Hydrological Processes*, 220 (14), 2723–2725.
- Dawdy, D.R. and Matalas, N.C., 1964. Statistical and probability analysis of hydrologic data, part III: analysis of variance, covariance and time series. In: Ven Te Chow, ed. *Handbook of applied hydrology*. New York: McGraw-Hill, 8.68–8.90.
- Duan, Q. and Sorooshian, S., 1992. Effective and efficient global optimization for conceptual rainfall-runoff models. *Water Resources Research*, 280 (4), 1015–1031.
- Ehret, U. and Zehe, E., 2011. Series distance – an intuitive metric to quantify hydrograph similarity in terms of occurrence, amplitude and timing of hydrological events. *Hydrology and Earth System Sciences*, 150 (3), 877–896. doi:10.5194/hess-15-877-2011.
- Feyen, L., Kalas, M., and Vrugt, J.A., 2008. Semi-distributed parameter optimization and uncertainty assessment for large-scale streamflow simulation using global optimization. *Hydrological Sciences Journal*, 530 (2), 293–308.
- Feyen, L., *et al.*, 2007. Parameter optimisation and uncertainty assessment for large-scale streamflow simulation with the LIS-FLOOD model. *Journal of Hydrology*, 3320 (3–4), 276–289.
- Gupta, H.V., Sorooshian, S., and Yapo, P.O., 1999. Status of automatic calibration for hydrologic models: comparison with multilevel expert calibration. *Journal of Hydrologic Engineering*, 40 (2), 135–143.
- Gupta, H.V., *et al.*, 2009. Decomposition of the mean squared error and nse performance criteria: implications for improving hydrological modelling. *Journal of Hydrology*, 3770 (1–2), 80–91.
- He, Y., *et al.*, 2009. Tracking the uncertainty in flood alerts driven by grand ensemble weather predictions. *Meteorological Applications*, 160 (1), 91–101.
- Hersbach, H., 2000. Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting*, 150 (5), 559–570.
- Jolliffe, I.T. and Stephenson, D.B., 2003. *Forecast verification – a practitioner's guide in atmospheric science*. New York: John Wiley & Sons.
- Laio, F. and Tamea, S., 2007. Verification tools for probabilistic forecasts of continuous hydrological variables. *Hydrology and Earth System Sciences*, 110 (4), 1267–1277.
- Liu, Y., *et al.*, 2011. A wavelet-based approach to assessing timing errors in hydrologic predictions. *Journal of Hydrology*, 3970 (3–4), 210–224.
- Mantovan, P. and Todini, E., 2006. Hydrological forecasting uncertainty assessment: incoherence of the GLUE methodology. *Journal of Hydrology*, 3300 (1–2), 368–381.
- Mathevet, T., *et al.*, 2006. A bounded version of the Nash–Sutcliffe criterion for better model assessment on large sets of basins. In: V. Andréassian, A. Hall, N. Chahinian and J. Schaake, eds. *Large sample basin experiments for hydrological model parameterisation: Results of the Model Parameter Experiment - MOPEX*. IAHS Red Books Series n° 307, 211–219.
- McCuen, R.H., Knight, Z., and Cutter, A.G., 2006. Evaluation of the Nash–Sutcliffe efficiency index. *Journal of Hydrologic Engineering*, 110 (6), 597–602.
- McMillan, H., *et al.*, 2010. Impacts of uncertain river flow data on rainfall-runoff model calibration and discharge predictions. *Hydrological Processes*, 240 (10), 1270–1284. doi:10.1002/hyp.7587.
- Mo, X., *et al.*, 2006. Parameter conditioning and prediction uncertainties of the LISFLOOD-WB distributed hydrological model. *Hydrological Sciences Journal*, 510 (1), 45–65.
- Nash, J.E. and Sutcliffe, J.V., 1970. River flow forecasting through conceptual models part I – a discussion of principles. *Journal of Hydrology*, 100 (3), 282–290.
- Richardson, D.S., 2000. Skill and relative economic value of the ECMWF ensemble prediction system. *Quarterly Journal of the Royal Meteorological Society*, 1260 (563), 649–667.

- Roulin, E., 2007. Skill and relative economic value of medium-range hydrological ensemble predictions. *Hydrology and Earth System Sciences*, 110 (2), 725–737.
- Schaefli, B. and Gupta, H.V., 2007. Do nash values have value? *Hydrological Processes*, 210 (15), 2075–2080.
- Seibert, J., 2001. On the need for benchmarks in hydrological modelling. *Hydrological Processes*, 150 (6), 1063–1064.
- Spellman, F.R. and Whiting, N.E., 2004. *Environmental engineer's mathematics handbook*. Abingdon: CRC Press.
- Thiébaux, H.J. and Zwiers, F.W., 1984. Interpretation and estimation of effective sample size. *Journal of Climate and Applied Meteorology*, 230 (5), 800–811.
- Thielen, J., et al., 2009a. The European Flood Alert System part 1: concept and development. *Hydrology and Earth System Sciences*, 130 (2), 125–140.
- Thielen, J., et al., 2009b. Monthly-, medium-, and short-range flood warning: testing the limits of predictability. *Meteorological Applications*, 160 (Special Issue), 77–90.
- Thiemig, V., et al., 2010. Ensemble flood forecasting in Africa: a feasibility study in the JubaShabelle river basin. *Atmospheric Science Letters*, 110 (2), 123–131.
- Van Der Knijff, J.M., Younis, J., and De Roo, A.P.J., 2010. LISFLOOD: a GIS-based distributed model for river basin scale water balance and flood simulation. *International Journal of Geographical Information Science*, 240 (2), 189–212.
- Weijs, S.V., Schoups, G., and van de Giesen, N., 2010. Why hydrological predictions should be evaluated using information theory. *Hydrology and Earth System Sciences*, 140 (12), 2545–2558. doi:10.5194/hess-14-2545-2010.
- WMO, 1966. *Climatic change*. Technical note no. 79. Geneva: World Meteorological Organization
- Yang, J., et al., 2007. Hydrological modelling of the Chaohe Basin in China: statistical model formulation and Bayesian inference. *Journal of Hydrology*, 3400 (3–4), 167–182.