Journal of Hydrology 376 (2009) 463–475

Contents lists available at ScienceDirect

Journal of Hydrology

journal homepage: www.elsevier.com/locate/jhydrol

Verification of ensemble flow forecasts for the River Rhine

M. Renner [a,*], M.G.F. Werner [b,c], S. Rademacher [d], E. Sprokkereef [e]

[a] Dresden Technical University, Faculty of Forestry, Geosciences and Hydrosciences, Institute of Hydrology and Meteorology, 01062 Dresden, Germany
[b] Deltares-Delft Hydraulics, P.O. Box 177, 2600 MH, Delft, The Netherlands
[c] UNESCO-IHE, P.O. Box 3015, 2601 DA, Delft, The Netherlands
[d] Federal Institute of Hydrology, Am Mainzer Tor 1, 56068 Koblenz, Germany
[e] Rijkswaterstaat, Centre for Water Management, Zuiderwagenplein 2, 8224 AD Lelystad, The Netherlands

ARTICLE INFO

Article history:
Received 21 March 2009
Received in revised form 5 June 2009
Accepted 18 July 2009

This manuscript was handled by K. Georgakakos, Editor-in-Chief, with the assistance of András Bárdossy, Associate Editor

Keywords:
Hydrologic ensemble forecasting
Probabilistic verification
River Rhine
ECMWF-EPS
COSMO-LEPS
HBV

SUMMARY

Ensemble stream flow predictions obtained by forcing rainfall–runoff models with probabilistic weather forecasting products are becoming more commonly used in operational flood forecasting applications. In this paper the performance of ensemble flow forecasts at various stations in the Rhine basin are studied by the means of probabilistic verification statistics. When compared to climatology positive skill scores are found at all river gauges for lead times of up to 9 days, thus proving the medium-range flow forecasts to be useful. A preliminary comparison between the low resolution ECMWF-EPS forecast and the high-resolution COSMO-LEPS forecast products shows that downscaling of global meteorological forecast products is recommended before use in forcing rainfall–runoff models in flow forecasting.

Introduction

Hydrological forecasts for the medium range have long been limited by the quality of quantitative precipitation forecasting, being the most challenging task in meteorological forecasting (Roulin, 2007). Meteorological forecasts are nowadays based upon the numerical solution of the atmospheric equations with Numerical Weather Prediction (NWP) models. These finite, nonlinear differential equations prove to have unstable solutions when initial conditions are slightly changed (Lorenz, 1963) and the recognition of this fact, as well as the increase in computing power has led to the development of meteorological ensemble forecasts, which simulate the evolution of the atmosphere due to perturbed initial conditions.

For hydrological forecasts it is a logical step to make use of these meteorological ensemble forecasts, as they provide additional and valuable information about meteorological forecast uncertainty, especially future precipitation amounts (Bartholmes and Todini, 2005; Roulin, 2007). One approach is to use ensemble forecasts as the input for hydrological rainfall–runoff models, to gain an ensemble of hydro-meteorological flow forecasts. This has been demonstrated for several historical flood events using the Ensemble Prediction System (EPS) of the European Center for Medium-range Weather Forecasts (ECWMF) by for example de Roo et al. (2003), Werner et al. (2004) and Gouweleeuw et al. (2005).

Although the running of hydrological ensemble forecasts is a challenge in its own right, there are several key scientific questions that need to be addressed. These focus on assessing the performance of the hydro-meteorological ensemble forecasts, the comparison of different ensemble products and understanding the relationship between the resolution of the meteorological forecasting models and the catchment scale. Furthermore, the question is how well the meteorological forecast uncertainty is reflected in the ensemble flow forecast. Forecast verification techniques may be applied to address these questions with several techniques developed within the atmospheric sciences being applicable to the hydrological sciences (Wilks, 2006).

Several authors have verified hydro-meteorological ensemble forecasts for single events, e.g. the Rhine and Meuse floods in 1993 and 1995 (de Roo et al., 2003) and the Oder flood in 1998 (Gouweleeuw et al., 2005), and although these provide valuable insight in the reliability of the forecast for the single event, the per-

* Corresponding author. Tel.: +49 351 463 31341; fax: +49 351 463 3130.
E-mail address: Maik.Renner@mailbox.tu-dresden.de (M. Renner).

0022-1694/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.jhydrol.2009.07.059

formance of ensemble flow forecasts may change from event to event. As a result forecasts should preferably be verified over a longer period. Gouweleeuw et al. (2005) suggest a period of at least one hydrological year is required to verify the performance of ensemble forecasts. Roulin (2007) verified ensemble flow forecasts made using the ECMWF Ensemble Predictions as an input for two Belgian catchments covering a period of six years, and Regimbeau et al. (2007) verified ensemble forecasts for all French basins for a period of one year. Both studies show that the ensemble forecasts offer skill. Additionally Roulin (2007) shows that an ensemble forecast has more relative economic value than a deterministic forecast, or even if only the ensemble mean is used.

In contrast, the verification of the global Ensemble Prediction Systems (EPS) by Buizza et al. (2005) led to the conclusion that EPS cannot account for all sources of meteorological uncertainty, as it typically does not display enough variability. As a result the output of the EPS cannot be considered as reliable and the empirical distribution derived from the ensemble members does not provide a useful probabilistic flow forecast (Fortin et al., 2006). High-resolution meteorological ensemble forecast systems, such as the COSMO-LEPS ensemble system showed an improvement in the prediction of heavy rainfall events for the 3–5 days lead time forecasts (Montani et al., 2003b; Marsigli et al., 2005). The application of COSMO-LEPS for hydrological ensemble forecasting has been presented for case studies of extreme events by Diomede et al. (2006), Walser (2006), Dietrich et al. (2008). Promising verification results of flow forecasts using COSMO-LEPS over a period of 2 years have recently been presented by Jaun and Ahrens (2009).

Hydro-meteorological ensemble forecasts have routinely been made for several years using an operational forecasting system for the Rhine basin by the forecasting centers of both the Rijkswaterstaat – Centre for Water Management in the Netherlands and the Federal Institute of Hydrology in Germany. In this paper ensemble forecasts made using the Rhine forecasting system are verified for the period over which ensemble meteorological input data are available. The ensembles used in the Rhine forecasting system include both the global ECMWF-EPS ensemble forecasts and the higher resolution COSMO-LEPS ensemble forecasts. And although the latter have only been available for a period of about nine months, the former have been available for a longer period of time, thus allowing comparison of the improvement of skill due to the use of a down-scaled local area ensemble system instead of the global ensemble input data. As the Rhine forecasting system provides forecasts for catchments that cover both a wide range of areas and types of hydrological response, the influence of these factors on the skill of the prediction at different lead times can be assessed.

Ensemble flow forecasting in the River Rhine

The River Rhine is the most important waterway in Europe. It is used intensively and the catchment is highly populated with about 58 million people living within the basin (IKSR, 2005). It is clear that reliable prediction of the behavior of the fluvial system is not only interesting for planning purposes, but also within the context of operational flood forecasting, warning and response to allow mitigation of risks when floods occur.

Here the performance of ensemble flow forecasts along the main Rhine stretch from the gauge at Maxau on the Upper Rhine (catchment area of 50,000 km^2) to the gauge at Lobith on the Dutch–German border (catchment area of 160,000 km^2) are studied. Additionally forecasts for river gauging stations on the main tributaries, such as the Moselle, Main and Ruhr have been analyzed. These gauging stations can be found on the map in Fig. 1. The map also shows isochrones of the estimated time of concentration of the catchment to the gauging station at Lobith. Because of the large concentration times, hydrodynamic models of the main Rhine, or even statistical multi-linear regression models using upstream water level observations can be expected to have a high forecast skill for the short range (1–4 days) for stations on the Middle and Lower Rhine.

However, hydrological and meteorological conditions may vary considerably in different parts of the basin. The basin can be divided into the southern, Alpine part which influences the High and the Upper Rhine, the hilly ranges in the middle influencing the Middle Rhine, and the Northern lowlands of the Lower Rhine. The hydrological regime in the Alpine areas is influenced by snowmelt and summer precipitation runoff, causing floods in the spring and summer months (e.g. flood events of May 1999, August 2005, and August 2007). In the Middle and Lower Rhine floods occur after heavy precipitation in the winter months, with the main flood genesis areas being the catchments of the Neckar, Main and Moselle (Fig. 1). It has been found that the coincidence of flood waves from these tributaries leads to the largest floods in the Middle and Lower Rhine, for example the events of 1993 and 1995 (Disse and Engel, 2001). The diversity of the hydrological regimes of the basin shows that the spatial and temporal patterns of future heavy precipitation are highly relevant for hydrological forecasting aiming at providing reliable forecasts for the medium range (4–10 days).

Forecasting suite

For this study the operational forecasting system was set-up using the flood forecasting shell Delft-FEWS (Flood Early Warning System, Werner et al. (2004)). This system serves to assist the forecasting departments of both the Federal Institute of Hydrology (Bundesanstalt für Gewässerkunde – BfG) in Koblenz, Germany (where the forecasting system is known under the acronym FEWS-DE) and the Centre for Water Management of Rijkswaterstaat (acronym: FEWS-NL).

Forecasts are made based on the response of the catchment to observed and future precipitation, which is simulated using a conceptual hydrological model. The HBV hydrological model is used, developed by the Swedish Meteorological and Hydrological Institute (SMHI). HBV belongs to the family of conceptual hydrological models with the response lumped into areas with the same hydrological properties, separated by elevation and vegetation zones. The routing from each sub-basin downstream is done using the Muskingum method or simple time lags, whereby each sub-basin has individual response functions. For more details see Bergström (2005) and Eberle (2001) for an application of the HBV model for the Rhine between Basel and Lobith. The model has been calibrated by the Federal Institute of Hydrology (Eberle et al., 2005) using observations in the period from 1990 to 1999. The Rhine basin has been divided into 134 sub-basins, with a river gauge at the outlet of each, thus allowing each sub-basin to be calibrated independently.

Within the forecasting suite the model is run in two modes; (i) historical mode and (ii) forecast mode. In historical mode, interpolated temperature and precipitation observations are used as dynamic input to the HBV model. These data are obtained from a network of meteorological stations across the basin, depicted in Fig. 1 as triangles. Data are delivered in *real time* to the operational forecasting centers and are stored in the data-management environment provided by the forecasting system. Prior to its use in the hydrological model, the data are validated against plausible ranges and then spatially and temporally interpolated to provide an input to the HBV model. The model time step was selected as 1 h and the precipitation and temperature data are spatially interpolated to provide mean areal values for every sub-basin through Kriging.

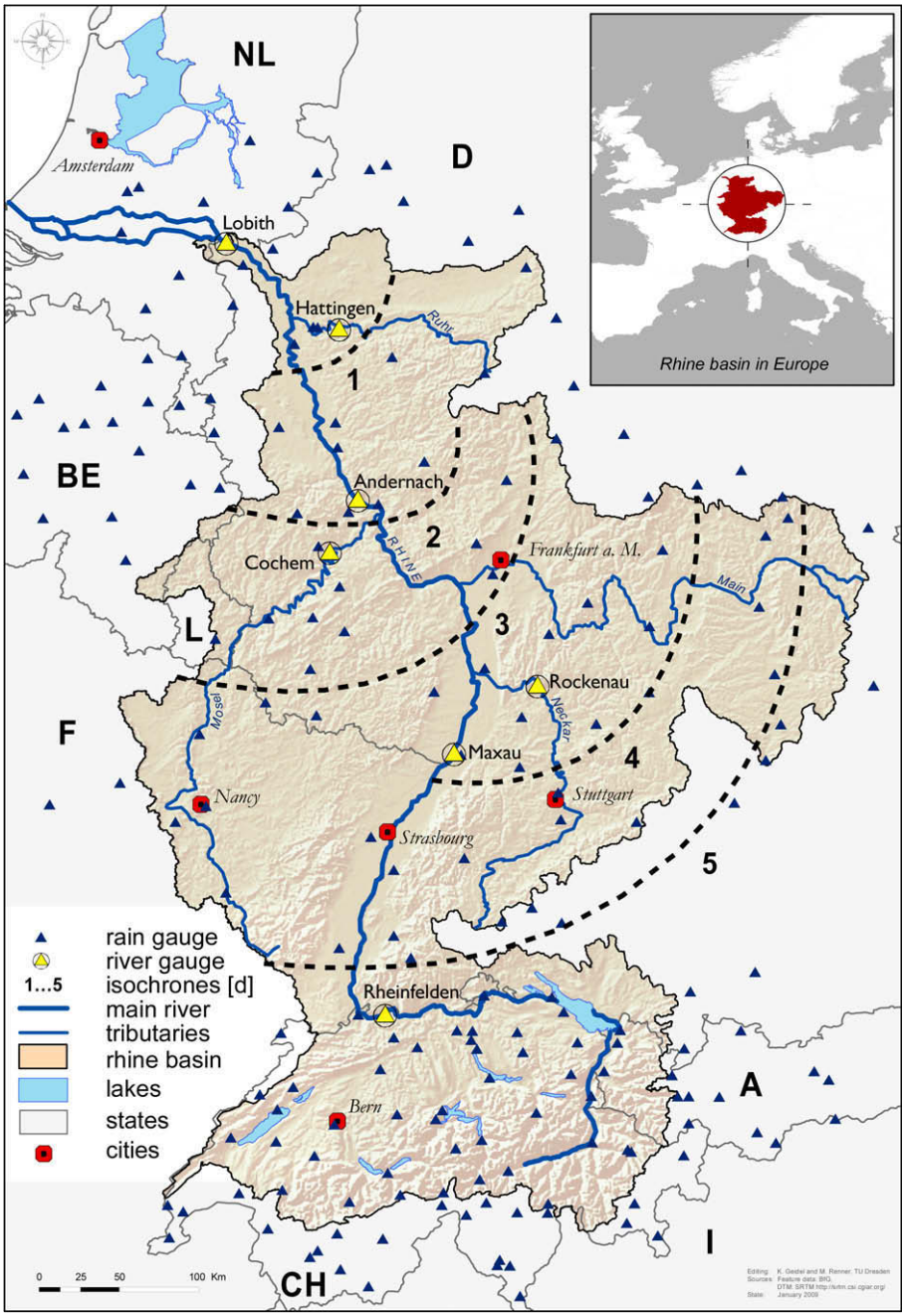

Fig. 1. Map of the Rhine basin. The dashed black lines indicate isochrones of the time of concentration in days to the gauge at Lobith on the German/Dutch border. Note, that only tributaries and river gauges relevant for the paper are shown (after Parmet and Sprokkereef (1997)).

In forecast mode, future mean areal precipitation and temperature are derived from the available probabilistic and deterministic meteorological forecast models, and are used as input into the HBV model. For meteorological ensemble forecasts the model is run sequentially for every ensemble member, thus producing an ensemble of flow forecasts.

To reduce bias in the flow forecasts thus obtained, an autoregressive (AR) error correction algorithm (Broersen and Weerts, 2005) is applied at river stations where discharge observations are available. The AR module establishes a forecast of the future error, based on the observed model error over a period (set at 192 h) prior to the start of the forecast. The model forecast is then corrected using this forecast of the future error. The correction is only effective for the respective forecast river station and does not transfer to stations downstream. This error correction method has proven to increase the performance of forecasts, especially in the short term. At longer lead times in relation to the hydrological response time of the catchment, the corrected forecast will converge to the uncorrected forecast.

Meteorological ensemble forecasts

Two meteorological ensemble forecasts were applied; (i) the ECMWF-EPS ensemble and (ii) the COSMO-LEPS ensemble. The European Centre for Medium-Range Weather Forecasts (ECMWF) provides the global circulation model ECMWF-EPS which is described e.g. in Molteni et al. (1996) or Buizza (2005). The regional scale COSMO-LEPS provided by the COnsortium for Small-scale

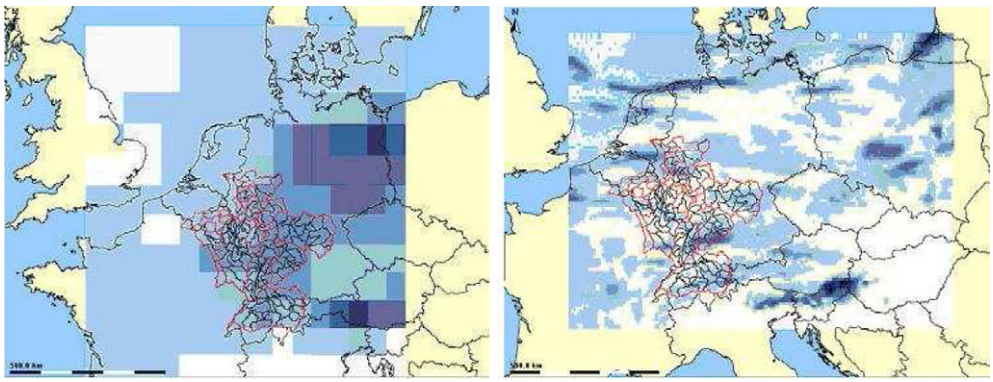

Fig. 2. Example of a rainfall prediction field for one time step and one member from ECMWF-EPS (left sub figure) and from COSMO-LEPS (on the right).

Table 1
Overview of the ensemble forecast models used.

	ECMWF-EPS	COSMO-LEPS
Ensemble size	50 + 1 control	16
Grid size	80 km, since 02/2006: 50 km	10 km
Time step	12 h	3 h
Forecast length	240 h	135 h

MOdeling (COSMO) is a dynamic downscaling approach based on a local area model (LAM). It uses initial and boundary conditions derived from the ECMWF-EPS over a domain covering a large part of Europe. Due to computational constraints, the COSMO-LEPS approach selects 16 members from the 51 member ECMWF-EPS forecast. These are so-called 'representative members', selected through a clustering technique to ensure most of the information of the global ensemble is preserved (Montani et al., 2003a; Marsigli et al., 2005). Fig. 2 gives an impression of the difference in scale of these two forecasts, showing precipitation fields over the Rhine basin for the same forecast time.

Note that for this study the meteorological forecast data has been used as it was available to the hydrologic forecasting agencies. These are previously resampled from the original forecast products to provide meteorological forecast inputs at regular time intervals (12 h and 3 h, respectively), on a regular longitude–latitude grid. Table 1 provides details of these two ensembles as used in this study. Also note that in February 2006 the resolution of the ECMWF-EPS forecast model has been increased. However, any effects related to this change are not discussed in this article.

Re-analysis of ensemble forecasts – hindcast approach

For the verification study it was necessary to re-run the hydrometeorological ensemble forecasts, as the ensemble flow forecasts made operationally have not been archived. The forecast system covering the Rhine basin was set to run in a batch mode for the full verification period, with forecasts run at defined intervals for the full period. Observed data after the forecast start time (t_0) were not used within these batch forecast runs, despite that data being available in the database of the system. Flow forecasts for the Rhine basin are routinely made once a day at 06*UTC*, whilst the meteorological forecasts used in this study are issued on a daily basis with the forecast start time at 12*UTC*. This means that at the t_0 of the hydrological forecast, the meteorological forecasts are already 18 h 'old'. The hindcast run was similarly set-up with a daily forecast start time at 06*UTC*, thus emulating the use of the system within the operational forecast environment.

The availability of archived meteorological ensemble forecasts formed an evident constraint on the length of the verification period. Table 2 shows the availability of these archived forecasts, and while the global ECMWF-EPS forecasts are available for over 3 years, the COSMO-LEPS forecasts span only about 9 months. Even though the original ECMWF-EPS and COSMO-LEPS forecasts have been available for quite a bit longer, the dates given in the table reflect their availability to the forecasting agencies mentioned. Observed data are available for a longer period, which were used to establish a stable HBV simulation to serve as a baseline for the hindcast run.

Table 2
Overview of the availability of the input data.

Data type	Time step (h)	Available from	Available to
Observed data			
Discharge (derived from water levels)	1	15/12/2000	10/10/2007
Precipitation	1–24	11/01/2001	01/10/2007
Temperature	1–24	11/01/2001	01/10/2007
Forecast data			
ECMWF-EPS	12	08/06/2004	10/10/2007
COSMO-LEPS	3	10/01/2007	10/10/2007

Verification measures for ensemble forecasts

Forecast verification allows the forecast to be monitored, thus helping to improve forecast quality by discovering the strength and deficiencies in a set of forecasts and allowing objective comparison of different forecasts. The aim of any forecasting activity is to support decision making, and the added value of a forecast therefore clearly depends on its error characteristics. These are established in an objective way through verification statistics, which delivers the information that is essential for the users to derive the full economic value from the forecasts (Wilks, 1995; WMO, 2007). However, the economic value can not be estimated using the verification statistics alone, but requires the potential value of the response to the forecast to be known, as well as the potential loss should a forecast not be available or not be acted on. The establishing of these relative economic values for the Rhine basin is beyond the scope of this paper and is therefore not addressed.

Scores suitable to the verification of ensemble and probabilistic forecasts are different to common verification scores used for deterministic forecasts. It is not useful to compare single ensemble forecast members against the reference, and probabilistic verification statistics are needed instead. WMO (2007) defined three properties of an accurate probabilistic forecast:

Reliability: the agreement between forecast probability of an event and the mean observed frequency of that event.

Sharpness: the tendency to forecast probabilities of an event occuring being near 0 or 1, as opposed to values clustered around the mean.
Resolution: the ability of the forecast to resolve the set of sample events into subsets with characteristically different outcomes.

To assess these properties, several statistical measures should be considered concurrently (Cloke and Pappenberger, 2008). In this paper several statistical measures are considered, including the Brier (Skill) Score and the Ranked Probability (Skill) Score, as well as the reliability diagram and the rank histogram.

Verification statistics considered

Brier (Skill) Score

The Brier Score is the most common scalar accuracy measure for dichotomous predictands (Wilks, 1995). To apply the score to continuous probability forecasts of precipitation or flow, these forecasts have to be translated into a binary event, e.g. using a threshold which can either be exceeded or not. Considering an ensemble forecast of continuous discharges, the forecast probability y_i is derived by the relative frequency of ensemble members exceeding the chosen threshold at the desired lead time. The observations are translated similar to the forecasts, if the threshold is exceeded, i.e. there is an event, then the observation $o_i = 1$ or if the event does not occur $o_i = 0$. The Brier Score BS can then be calculated from a set of m pairs of forecast probabilities y_i and corresponding observations o_i:

$$BS = \frac{1}{m}\sum_{i=1}^{m}(y_i - o_i)^2 \tag{1}$$

Essentially the Brier Score is the mean-square error of probability forecasts. It is negatively orientated, with a perfect score of $BS = 0$. As observations and probability forecasts are bounded by 0 and 1, the Brier Score equally ranges between 0 and 1. The Brier Score is the most important score to verify prediction models, because it accounts both for reliability and sharpness of the forecast. However, as the score depends on the verification dataset, a model comparison should be based on that same dataset (Kirk and Fraedrich, 1990). To compare with other stations or datasets it is recommended to use the Brier Skill Score $(BSS), BSS = 1 - BS_{forecast}/BS_{reference}$. In this way the BS of the forecast is compared with a reference forecast e.g. climatology or persistence. The skill score ranges from $-\infty$ to 1, with a perfect skill of 1. A value of zero indicates no skill when compared to the reference forecast.

The reliability diagram

Forecast verification methods typically compare corresponding forecast–observation pairs. At a more fundamental level forecast verification involves the investigation of the joint distribution between forecast and observation $p(y_i, o_j)$. Applying the definition of conditional probabilities, the joint distribution can be factored in two ways. The calibration-refinement (CR) factorization is conditional on the forecast (Wilks, 1995):

$$p(y_i, o_j) = p(o_j|y_i)p(y_i), \quad \text{with } i = 1,\ldots,I;\ j = 1,\ldots,J. \tag{2}$$

A conditional probability $p(o_i|y_j)$ is derived, which specifies how often an event occurred, when a forecast y_i of that event has been issued. The reliability diagram plots this conditional probability, which is also referred to as observed relative frequency $\bar{o}_i$, against discrete values of forecast probabilities. The forecast probabilities are put into I discrete categories. The observed relative frequency $\bar{o}_i$ is derived by:

$$\bar{o}_i = p(o_i|y_i) = \frac{1}{D_i}\sum_{k\in D_i} o_k, \quad \text{with } i = 1,\ldots,I, \tag{3}$$

where D_i is the number of forecasts in forecast value category i (conditional sample size) and o_k are the corresponding observations, which can take the value $o_k = 1$ if the event occurred and $o_k = 0$ if the event did not occur.

Usually a histogram of the forecast probabilities is shown together with a reliability diagram. This histogram is an immediate graphical indication of the sharpness of the probabilistic forecast. High frequencies for the forecast probabilities 0 and 1 and low frequencies in between indicate a 'sharp' forecast. The sharper a forecast, the better it can distinguish between an event and a non-event. However, sharpness alone does not determine if the prediction was right. Points along the diagonal line of the reliability diagram show that the observed relative frequency matches with predicted probabilities and therefore indicate a reliable or well calibrated probabilistic forecast.

Ranked Probability (Skill) Score

In order to verify the whole range of possible outcomes, the Ranked Probability Score (RPS) can be used (Wilks, 1995). For verifying flow rates, M categories are defined, which cover all possible outcomes. For all categories the squared differences between the cumulative forecast probability and the corresponding cumulative observation of each category are averaged to gain the RPS:

$$\mathrm{RPS} = \frac{1}{M-1}\sum_{m=1}^{M}\left[\left(\sum_{i=1}^{m} y_i\right) - \left(\sum_{i=1}^{m} o_i\right)\right]^2. \tag{4}$$

The RPS is sensitive to distance, e.g. if a forecast falls into a more distant category than the observation, it will be penalized more. The Ranked Probability Score is negatively orientated and its skill score the Ranked Probability Skill Score $(\mathrm{RPSS}), \mathrm{RPSS} = 1 - \mathrm{RPS}_{forecast}/\mathrm{RPS}_{reference}$ is computed using a reference forecast.

Verification rank histogram

Verification rank histograms or simply rank histograms are used to check if an ensemble forecast is well calibrated (reliable) and consistent. The ensemble is called consistent if the actual future state of the predictand is drawn from the same distribution as the ensemble, i.e. the observed value can not be statistically distinguished from the ensemble members (Wilks, 2006).

The rank histograms can be established for a set of N forecasts at a specific lead time and an ensemble with n_{ens} members. Then for each forecast N the rank of the observation, within the ensemble is determined, e.g. if the observed value is smaller than all ensemble members the rank i is 1 or if the observation is larger then the forecast, the rank $i = n_{ens} + 1$ is assigned. Finally a histogram of the assigned ranks can be drawn as graphical verification means. A good discussion on the interpretation of rank histograms can be found in Wilks (2006).

Practical aspects of forecast evaluation

Verification categories and restricted sample size

Within the verification analysis the problem of the restricted sample size needs to be addressed. Ideally verification would measure the performance of the forecasting system at important warning thresholds. However, to establish meaningful verification statistics, a sufficiently large number of observed events (an event = flow/level exceeding a certain threshold) is needed. Typically thresholds that are meaningful within the context of operational flow forecasting are relatively extreme. As the verification period considered here is relatively short, these may not have occurred

Table 3
Percentiles of observed 24 h precipitation totals for a sample sub-basin.

Percentile	10%	40%	50%	60%	70%	80%	90%	100%
Threshold (mm)	0	0	0.07	0.60	1.54	2.94	5.11	9.26

at all or only so infrequently that the number of events is not large enough to give a meaningful statistic. To resolve this issue, the method proposed by Roulin (2007) was adopted here, where the thresholds are defined as the percentiles of the observed sample to guarantee that there is a fixed and large enough number of events to verify. An added advantage of this method is that it helps to compare verification statistics at different locations, which might have true flood warning levels at different return period flows.

Thresholds for precipitation forecasts

While thresholds and the associated categories for the RPS of flow forecasts were simply derived by deciles of observed flows, the usage of deciles for verification of precipitation forecasts is less meaningful as there is a high frequency of zero precipitation. Table 3 shows percentiles of the observed 24 h rainfall totals of an example sub-basin. It can be seen that dry days occur about 40% of the time. As a result thresholds, chosen for the RPS are a combination of fixed precipitation values (0.01 and 0.1 mm) for dry conditions, and the percentiles of 50%, 60%, 70%, 80%, 90% and 100% of the observed basin precipitation for wet conditions. The use of fixed values for wet conditions was not practical because the range of sub-basin precipitation amounts can differ significantly across the basin.

Multiple and single category measures

From the definition of the RPS (cf. Eq. (4)) it follows that it is less sensitive to the choice of thresholds than for the dichotomous score methods such as the reliability diagram or the commonly used Brier score, cf. Wilks (2006). Because multiple categories are used, the RPS and its associated skill score are more versatile to short verification periods. Single category statistics are more sensitive and may prove unreliable due to under-sampling. Dichotomous event scores have therefore been applied only for the full ECMWF-EPS verification period to look at forecast attributes such as bias, reliability, resolution and sharpness at low and high thresholds. In contrast the RPS is used as the primary verification statistic for the COSMO-LEPS forecast which is available for a comparatively short period.

Table 4
HBV model performance scores at the main river stations on the Rhine and its tributaries calculated for the verification period June 2004–October 2007. With area = catchment area, $\overline{Q}_{obs}$ = observed mean discharge, $\overline{Q}_{sim}$ = simulated mean discharge, MAE = mean absolute error, MARE = mean absolute relative error, R = correlation coefficient, NSE = Nash Sutcliffe efficiency.

Station	River	Area (km^2)	$\overline{Q}_{obs}$ (m^3/s)	$\overline{Q}_{sim}$ (m^3/s)	MAE (m^3/s)	MARE (%)	R	NSE
Lobith	Rhine	160,800	2052	1865	254	12	0.94	0.84
Andernach	Rhine	139,549	1859	1600	287	15	0.94	0.77
Maxau	Rhine	50,196	1137	968	199	16	0.91	0.70
Rheinfelden	Rhine	34,550	1097	940	198	17	0.90	0.71
Cochem	Moselle	27,262	268	226	72	35	0.92	0.79
Rockenau	Neckar	12,616	141	113	42	30	0.78	0.50
Hattingen	Ruhr	4124	78	74	20	26	0.90	0.80

Results

Performance of the HBV baseline simulation

The HBV model of the Rhine basin was first applied in historical (simulation) mode over the period from 01/11/2001 to 01/10/2007 to obtain a baseline simulation using the observed precipitation and temperature. Overall, the model performance of the HBV model within the verification period from 01/06/2004 to 01/10/2007 was found to be reasonable. Table 4 provides an overview of the model performance summarized using various various statistics for selected gauging stations on the Rhine and its major tributaries. It can be seen that there is a steady underestimation of the flow for the Rhine river stations. As an example, the hydrograph at Lobith is shown in Fig. 3. The trend of the accumulated differences seen at Lobith is repeated at most of the stations and seems to originate from Alpine catchments, which exhibit an underestimation (cf. Ta-

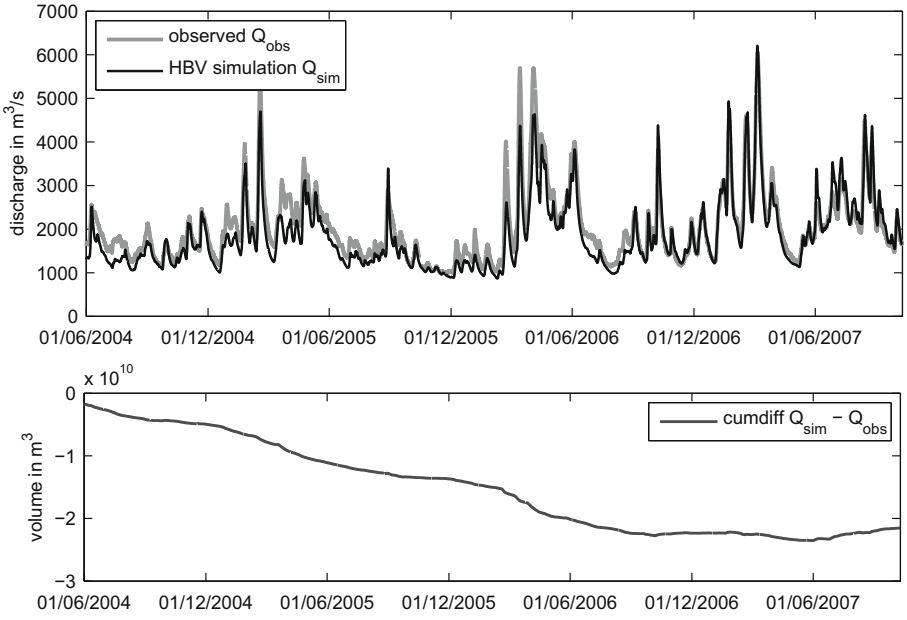

Fig. 3. Baseline simulation of the HBV model (thin black line) compared with observed discharges (bold grey line) at the river gauge at Lobith from June 2004 to October 2007. The lower subplot depicts the accumulated volume difference.

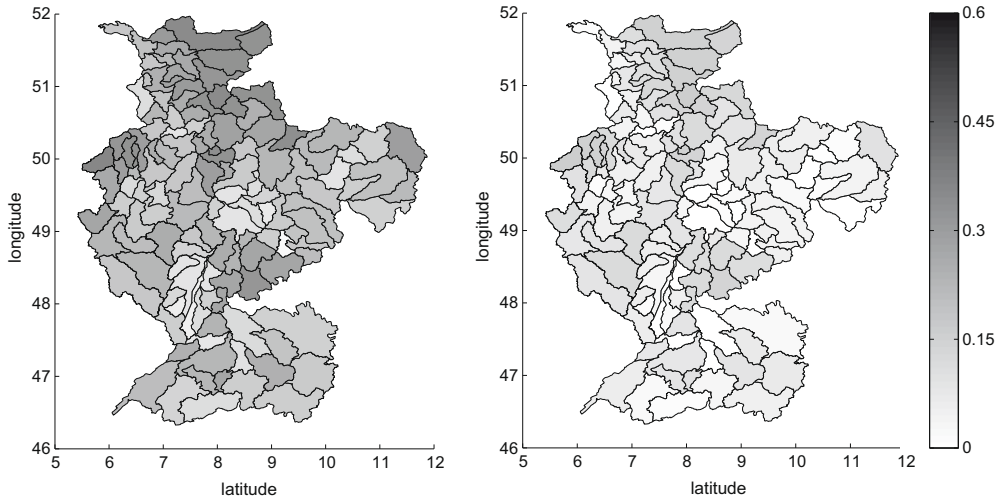

Fig. 4. Maps of Ranked Probability Skill Scores of ECMWF-EPS daily precipitation totals of the HBV sub-basins. The left subplot shows the scores for a lead time of one day, the right one for 5 days. Higher skills are indicated with dark grey, as displayed by the colorbar on the right.

ble 4, columns observed and simulated mean discharges at Rheinfelden and Maxau).

Due to river regulation on the larger tributaries (Moselle, Main and Neckar), the water level–discharge relations are not reliable during low flows, which results in bad performance scores during low flows. However, the relatively high correlation showed that during higher flow regimes the HBV model simulations compare well with the observed discharges. Despite this poor performance during low flows as a consequence of regulation, and a generally poor performance for some of the more minor stations (not included in the table), the baseline simulation obtained was considered acceptable as a basis for the hindcast.

Verification of ECMWF-EPS precipitation forecasts

As mentioned previously, ECMWF-EPS forecasts were available for the period from 08/06/2004 to 01/10/2007, and verification of precipitation forecasts was first done for all forecasts within this period.

For the full range of precipitation categories, the RPSS was computed on the basis of the 24 h precipitation totals for every sub-basin and is displayed for lead times of one and five days in the maps in Fig. 4. These indicate regional patterns in forecast skill, where some regions show a relatively high skill of around 0.3, while sub-basins with very low or no skill are isolated. Generally the basins in the North show higher skill scores than those in the South, most likely as a result of the increasing heterogeneity of rainfall in the mountainous areas. These regional patterns are preserved at higher lead times.

To summarize the behavior of the RPSS with lead time, Fig. 5 presents the median, the first, and the ninth decile of all 134 basins. Within the first two days, the skill is at a constant high level and then deteriorates with increasing lead time. After 5–6 days there is no skill in the precipitation forecasts. The regional patterns and the decline observed from the RPSS maps is similar for the BSS for high precipitation amounts, verified at a level of the 80th percentile threshold.

Verification of ensemble flow forecasts using ECMWF-EPS

Verification of the full flow domain

The RPSS was computed for the river gauges Hattingen, Rockenau, Cochem, Rheinfelden, Maxau, Andernach and Lobith and is presented for different lead times of 1, 3, 5, 7 and 9 days in Fig. 6. The gauging stations are plotted on the *x*-axis, ranked in order of ascending catchment area. The skill generally increases with catchment area, with the exception of Rockenau on the Neckar. This gauging station is influenced most by regulation at low flows, thus explaining the low skill. A very high skill is observed at a lead time of one day, which is primarily due to the effects of the AR error correction. With increasing lead time the skill deteriorates faster in smaller basins, such as at the gauge at Hattingen on the Ruhr than in larger basins, which show a relatively high skill for the longer lead times.

Reliability of ensemble flow forecasts

Although reliability diagrams were established for all the major gauging stations in the Rhine and its tributaries, for brevity only two reliability diagrams are presented here; one for the gauging station at Maxau (Fig. 7) in the upper Rhine and one for the gauging station at Andernach (Fig. 8) in the Middle Rhine. Both diagrams were prepared for a threshold exceedance of the 8th decile of the observed flows within the verification period, and at a lead time of 8 days. Two verification pairs are shown. The first evaluates the reliability of the error corrected forecast (Q_{corr}) against the ob-

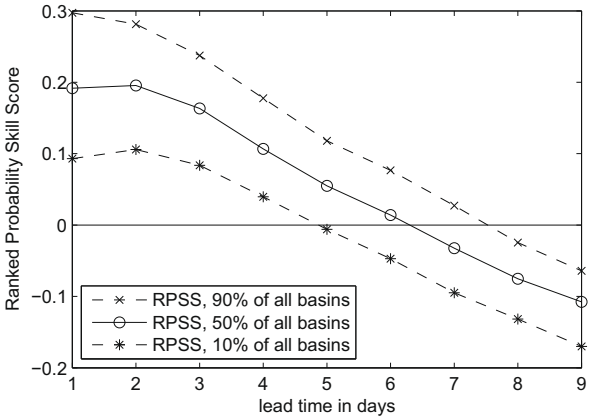

Fig. 5. RPSS of forecast precipitation at sub-basin scale using the ECMWF-EPS forecasts against sample climatology. The upper line denotes that 90% of all sub-basins have a RPSS at or below this line. The middle line denotes the median and the lower one the 10% percentile.

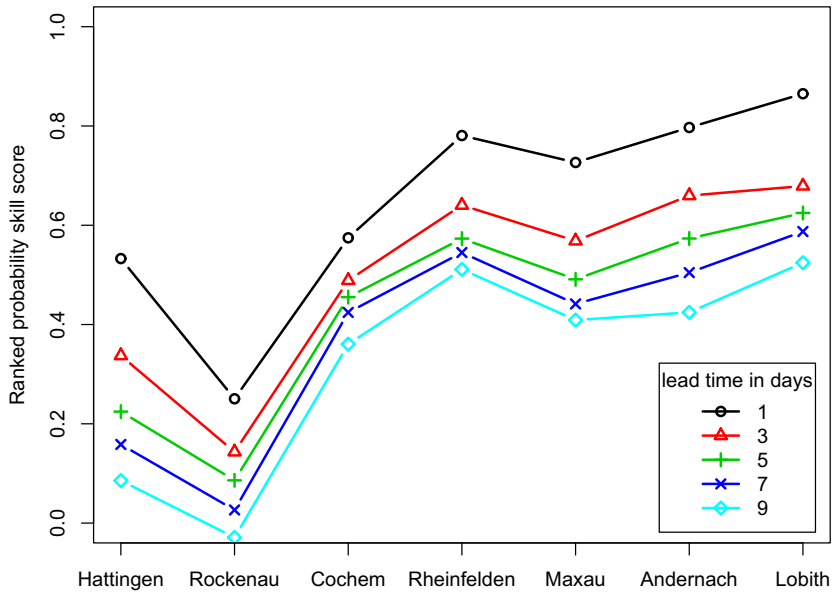

Fig. 6. RPSS of error corrected flow forecasts using ECWMF-EPS forecasts as input for several gauges, sorted by catchment size. The sample climatology of the observed flows has been used as reference forecast.

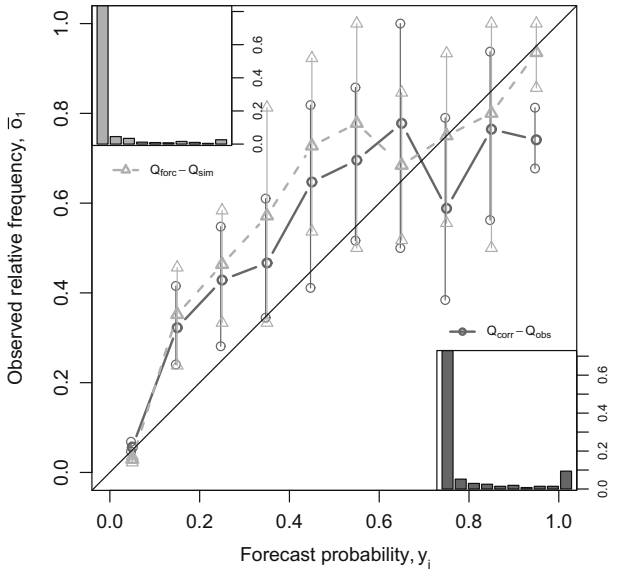

Fig. 7. Reliability diagram for Maxau at a lead time of 8 days exceeding 1460 m^3/s. The dark grey line marked with circles shows the reliability line for the verification of the error-corrected flow forecast against observations ($Q_{corr}-Q_{obs}$). The light grey dashed line marked with triangles shows the reliability of the uncorrected flow forecast against the baseline simulation ($Q_{forc}-Q_{sim}$). Confidence intervals are given for $\alpha = 0.05$. In the upper left and lower right corners respective histograms of forecast probabilities y_i are shown.

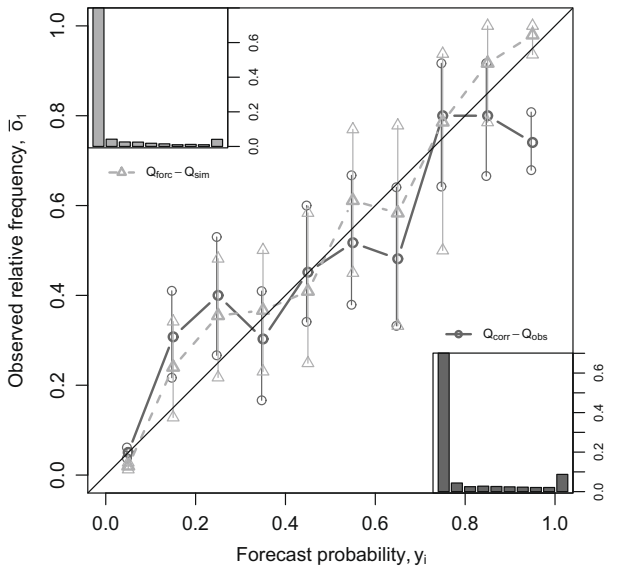

Fig. 8. Reliability diagram for Andernach at a lead time of 8 days exceeding 2370 m^3/s. See Fig. 7 for explanation.

served discharge (Q_{obs}), while the other evaluates the uncorrected forecast (Q_{forc}) against the baseline simulation (Q_{sim}). The first of these pairs is used to evaluate the reliability of the ensemble forecast, including all meteorological, hydrological and observational uncertainties, while the second is used to evaluate reliability due to the meteorological model only. Plotting both in one graph allows the contribution of the different sources of uncertainty to be compared simultaneously.

To address the effect of the small sample size, confidence intervals ($\alpha = 0.05$) for the reliability lines are shown, using the same marker symbol, which was used for the reliability line. The confidence intervals were derived by applying a bootstrap with a sample size of 100. The bootstrap sampling draws forecast event pairs randomly with replacement from the verification set and then for each sample reliability lines are calculated. The wide intervals for forecast probabilities between 0.1 and 0.9 underline the effect of sample size.

At the gauging station at Andernach, the reliability diagram in Fig. 8 indicates that the meteorological verification pair (Q_{forc} against Q_{sim}) is reliable, as the dashed line marked with triangles follows the diagonal. This is similar for the verification pair Q_{corr} against Q_{obs}, although the corrected forecast probabilities y_i larger than 90% are over-predicted, which can be seen in the solid line marked with circles below the diagonal. Over-prediction in this case means that for all cases when a forecast probability y_i indicated the threshold being exceeded with a chance of more than

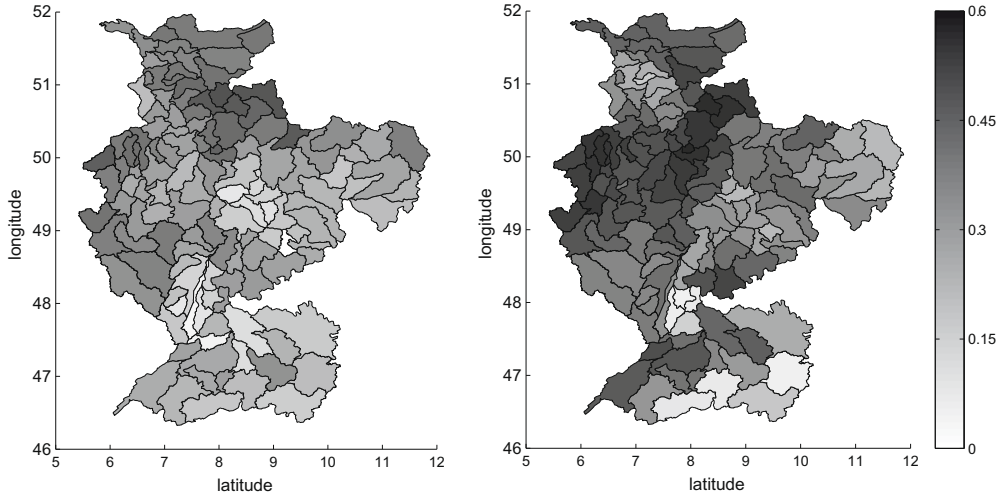

Fig. 9. Maps of Ranked Probability Skill Scores of ECMWF-EPS and COSMO-LEPS precipitation forecasts at sub-basin scale. The sub plot on the left shows the RPSS for ECMWF-EPS and the right one of COSMO-LEPS forecasts. The comparison is based on daily precipitation totals, shown here for the first 24 h after t_0. Higher skills are indicated with dark grey, as displayed by the colorbar on the right.

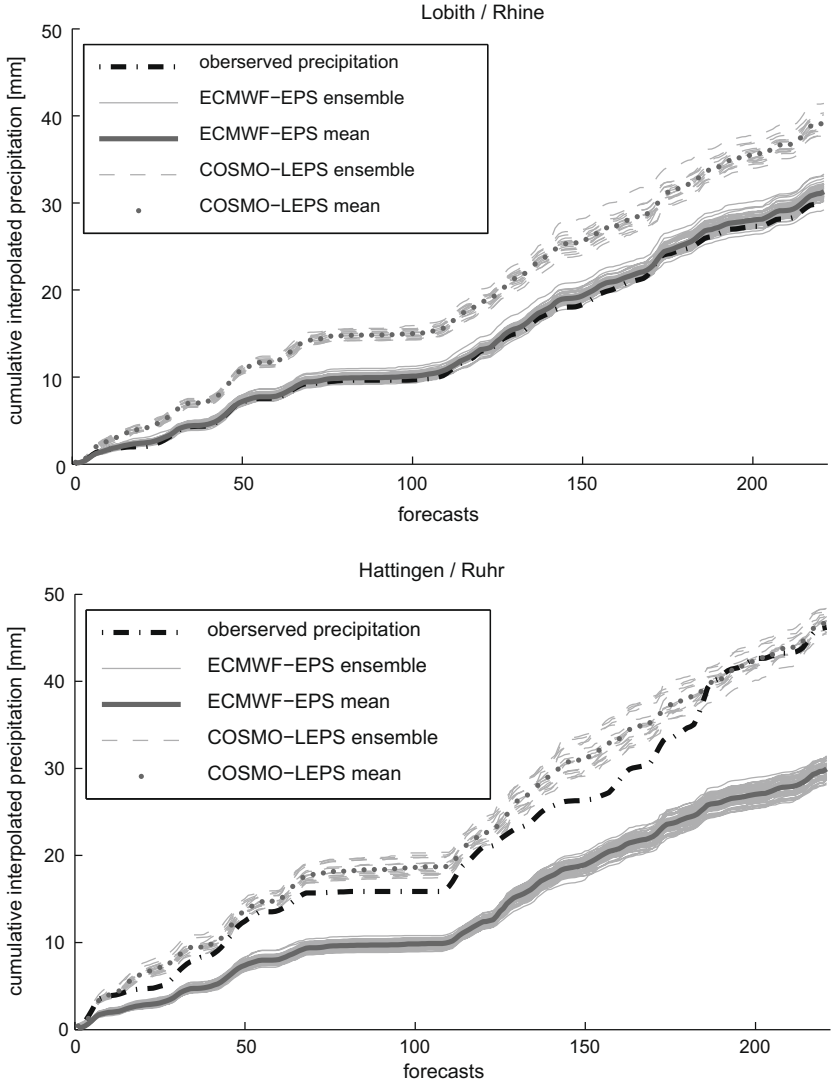

Fig. 10. Cumulative curves of precipitation forecasts of COSMO-LEPS (dashed lines) and ECMWF-EPS (grey lines) are compared to the observations (dot-dashed line). The bold grey line depicts the mean of ECMWF-EPS and the dotted curve the mean of COSMO-LEPS. The upper plot shows the resulting curves for Lobith/Rhine and the lower one for Hattingen/Ruhr.

90%, only 80% of these cases resulted in an observed exceedance. This phenomenon is apparent for most stations, when looking at the exceedance of the higher thresholds (see also Fig. 7).

At Maxau (Fig. 7), both curves are above the diagonal for low forecast probabilities $y_i < 70\%$. This is a clear indication that the meteorological input of ECMWF-EPS is biased, thus resulting in an under-prediction of the observed flows. Similar behavior was also found at basins of different size, such as Hattingen (Ruhr) and Rheinfelden (Rhine).

Comparison between COSMO-LEPS and ECMWF-EPS

To be able to compare the verification results of both meteorological ensemble forecasts, all results presented in this section are based on the shorter period for which the COSMO-LEPS forecasts were available (10/01/2007–01/10/2007). This period is only about 9 months in length, with verification being difficult due to under-sampling. All scores which are based on dichotomous events are sensitive to under-sampling, and are therefore not presented. Only the Ranked Probability Score and the Rank Histogram have been used for verification, as these measures are less affected by under-sampling.

Comparison of precipitation forecasts

Precipitation at the sub-basin scale. The RPSS of precipitation forecasts at the sub-basin scale were established analogous to those of the whole ECMWF-EPS verification period, except that the reference forecast is based on the sample climatology for the shorter period. A map of RPSS values for all sub-basins is presented in Fig. 9, comparing the RPSS of ECMWF-EPS on the left and COSMO-LEPS on the right. It can be immediately seen that the COSMO-LEPS forecasts show higher skill scores for most basins, particularly in the Western and Northern Alpine basins. Although the ECMWF-EPS forecasts show slightly higher RPSS values for this shorter verification period, the regional pattern does not change, when compared to Fig. 4. As with the ECMWF-EPS forecasts, the regional skill pattern of the COSMO-LEPS forecasts is preserved over lead time, with the skill decreasing. At a lead time of one day the skill of COSMO-LEPS is generally higher in the extent of 10%. The superiority of COSMO-LEPS is slowly decreasing with lead time (not shown).

Precipitation forecasts at the aggregated basin scale. The skill of the cumulative precipitation forecasts were additionally compared at the aggregated basin scale. The upper subplot in Fig. 10 shows the accumulated forecast precipitation for the entire Rhine basin to the gauge at Lobith, while the lower subplot presents the accumulated forecasts for the Ruhr basin to the gauge at Hattingen. COSMO-LEPS (dashed lines) and ECMWF-EPS (grey lines) are compared with the respective cumulative observed precipitation (dot-dashed line). The accumulation is based on precipitations sums of 108 h and is shown for aggregated basins, which represent the catchments of river stations used in the text. The dotted curve represents the ensemble mean of COSMO-LEPS and the bold grey line the mean of ECMWF-EPS.

It can be seen that the slope of the cumulative curve of COSMO-LEPS precipitation forecasts is steeper than that of the ECMWF-EPS forecasts for all aggregated basins. In most basins ECMWF-EPS forecast show a similar curve when compared to that of the observed catchment precipitation, while in the Ruhr basin at Hattingen, COSMO-LEPS forecasts are closer to the observation and the cumulative curve of ECMWF-EPS is below the observation curve. However, there are indications that there is an underestimation of observed catchment precipitation from rain gauge data, like the underestimation of the average flow of the baseline simulation (see Table 4).

Flow forecasts

To measure the skill of the COSMO-LEPS forecasts the RPSS was established. However, rather than using the climatology of the (shorter) verification period, the ECMWF-EPS flow forecasts made for the same period were used as a reference when computing the RPSS. Fig. 11 presents the derived RPSS at lead times of 1, 3, 5, 7 and 9 days for seven river stations on the main tributaries and the main Rhine itself. Values greater 0 indicate superiority of flow forecasts using COSMO-LEPS when compared with ECMWF-EPS.

It can be seen that COSMO-LEPS driven flow forecasts gain skill in the range of 1–5 days lead time, with the exception of the gauge

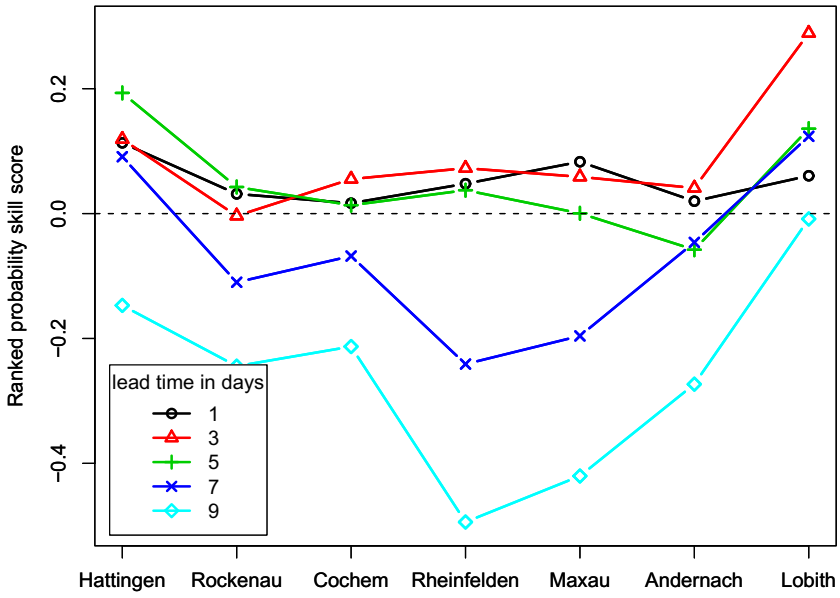

Fig. 11. RPSS of error corrected flow forecasts forced by COSMO-LEPS, using ECMWF-EPS flow forecasts as reference. The RPSS is given for lead times 1, 3, 5, 7 and 9 days, shown as lines connecting the stations, which are sorted according to catchment size.

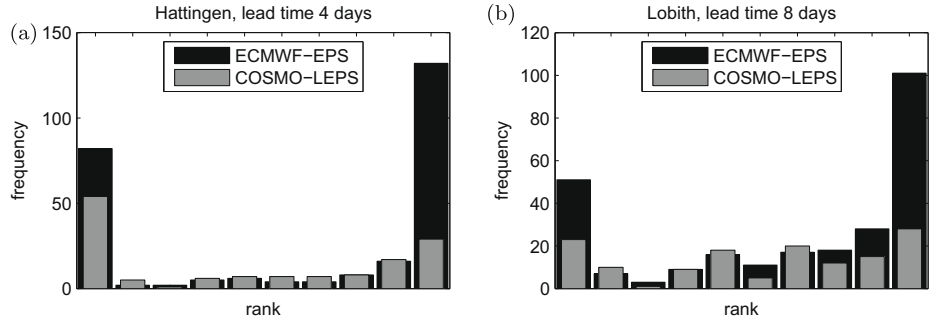

Fig. 12. Rank histograms for Hattingen (a) and Lobith (b). The ranks are 1–52 for ECMWF-EPS and 1–17 for COSMO-LEPS. For display reasons the ranks were put into 10 bins. Both diagrams comprise daily forecasts of the period for which COSMO-LEPS forecasts were available.

at Andernach, which does not show improvement beyond a lead time of 4 days. For lead times longer than 5 days a negative skill was found for most stations. However, Hattingen (Ruhr) and Lobith (Rhine) are positive exceptions, with high skill scores for COSMO-LEPS. That the increase in skill is only established after 1–2 days is not surprising as at the shorter lead times the meteorological forecasts will have little influence, and the skill is dominated by uncertainties in the hydrological models. At lead times in excess of 5–7 days the skill of the COSMO-LEPS forecast would again be expected to deteriorate as this is beyond the lead time of the meteorological forecast.

To compare reliability and consistency of the ensemble flow forecasts rank histograms are provided in Fig. 12. For display reasons, the ranks have been categorized into 10 bins. Both forecast models have high frequencies of very low and high ranks, which is a sign of underdispersion, i.e. showing too little uncertainty. However, COSMO-LEPS forced flow forecasts tend to be more equally distributed than the ECMWF-EPS ones and there is a clear lower frequency of very high ranks, i.e. events when the ensemble is below the observation. This can be seen at all locations considered, but to show the effect on small and large catchments Hattingen/Ruhr and Lobith/Rhine have been selected.

Discussion

The performance of hydro-meteorological ensemble forecasts for the Rhine forecasting system has been evaluated using probabilistic verification scores. These allowed to assess probabilistic accuracy properties such as reliability, resolution and sharpness of the ensemble forecasts. Beside the performance itself, forecasts made with the meteorological ensemble forecast ECMWF-EPS are compared with the downscaled COSMO-LEPS forecasts. The influence of catchment and hydrological response time are addressed by evaluating forecasts at several river gauges, with catchment areas between 4000 and 160,000 km^2.

Performance of ensemble flow forecasts across the Rhine basin

Generally a positive forecast skill for the medium range is found, i.e. 3–9 days lead time, depending on the catchment area. As expected, skill decreases with decreasing catchment area, and the smaller the respective catchment, the faster the forecast skill deteriorates with lead time. This is because the hydrological response time determines the lead time at which most of the input variability from the meteorological ensemble forecast reaches the basin outlet. Furthermore it was found that there is no skill in the ECMWF-EPS precipitation forecasts at the sub-basin scale after 5–6 days. This time plus the average hydrological response time gives a good estimator for the maximal lead time of a flow forecast with positive skill at the respective basin outlet.

Forecast reliability and resolution of ECMWF-EPS driven flow forecasts was assessed by means of reliability diagrams. Two main features were found, which are apparent at most river gauges. First there is a steady underprediction of lower forecast probabilities y_i exceeding a certain threshold, i.e. when there are just a few ensemble members over the threshold. As this feature is visible for both verification pairs, forecast (Q_{forc}) – simulation (Q_{sim}) and error corrected forecast (Q_{corr}) – observation (Q_{obs}), it is clear that this bias originates from the meteorological forecast. The global ECWMF-EPS does not produce enough variability as its spatial resolution is very coarse when compared to the model input scale. This confirms through verification of the flow forecast the results of Buizza et al. (2005) when verifying the precipitation forecast, who found the ECMWF-EPS forecast does not account for sufficient variability. The need of a better representation of dispersion through downscaling the global ensemble forecasts, for example through a high-resolution ensemble meteorological model such as COSMO-LEPS, is thus clear.

The second feature is apparent when most ensemble members are over the respective threshold, i.e. a high exceedance probability y_i has been predicted. In these cases the Q_{forc}–Q_{sim} pair is quite reliable, but the Q_{corr}–Q_{obs} pair tends to over-prediction. In this case the bias originates from a combination of errors in the forecast meteorological inputs as well as the hydrological model. Additionally errors in the precipitation observations and the process used to obtain a mean areal precipitation based on gauge data for the subbasins contributes to the error in the forecast. When looking at the resulting forecasts (Q_{corr}) a lack of forecast resolution must be attributed. In other words the ability of the forecast to resolve between an event and a non-event is relatively low.

Effects of increased meteorological forecast model resolution

The comparison of the flow forecasts forced using ECMWF-EPS or COSMO-LEPS shows that the higher resolution in time and space provided by the COSMO-LEPS forecast clearly improves forecast performance. Firstly there is a significantly higher skill (one day lead time) in precipitation forecasts at the sub-basin scale, secondly COSMO-LEPS produces higher precipitation amounts which consequently transfers to flow forecasts with higher skill and three times larger forecast variability.

Although Fig. 6 shows that the skill of the forecast generally increases with catchment size as would be expected, a clear relation between the catchment size and the resolution of the meteorological forecast was not found. The flow forecasts forced with COSMO-LEPS improved forecast skill at all catchment scales considered, with the rank histograms comparing both meteorological forecast models showing that the frequency of observations larger than the ensemble decreases for both small and large catchments.

Instead it was found that the resolution of the meteorological forecast output and the input of hydrological model in charge should be commensurate. In this case the global ECMWF-EPS model, with an approximate raster size of 10,000 km^2, is far to coarse to be used as an input to a sub-basin of the size of about 400 km^2. The downscaled COSMO-LEPS model with an approximate raster size of 100 km^2 is more appropriate in this case. This confirms the results of Clark and Hay (2004) and Regimbeau et al. (2007), and once more underlines the argument for downscaling of global weather prediction products before using in a regional hydrological model.

Sample size and other limitations

The verification statistics have been limited by the small sample size, i.e. the relatively short re-analysis period. Instead of using warning thresholds, which are most interesting for decision makers, the sample size has been increased by using percentiles of observed flows as thresholds. This limits the interest in the study as it may not be a clear representation of forecast performance for the more extreme events. However, the approach does allow flow forecasts at several gauging stations to be compared. To assess random effects of single forecast events in the re-analysis record, the bootstrap method has been applied to gain confidence intervals for reliability diagrams. Another way to manage low sample size, was to use multicategory verification measures, which make use of the whole verification data set, but consequently only provides a general measure of forecast skill and again not for the interesting extremes.

Uncertainties in the measurements and in the hydrological model are important in hydrological forecasting, and depending on the lead time will have significant effect on forecast skill. Although these have not been considered in the scope of this study, the comparison of the two pairs (Q_{forc} against Q_{sim}) and (Q_{corr} against Q_{obs}) allows the contribution to the error from the meteorological model to be separated from that of the hydrological model and input data. It should be kept in mind though that when comparing forecasts to observations, that the observations are also uncertain. Of particular note is the uncertainty in the "observed" areal precipitation, and there is some evidence that the areal precipitation is systematically underestimated. For example the average flow of the baseline simulation is significantly smaller than the observed average flow at most stations. This explains the steeper slope of COSMO-LEPS accumulated basin precipitation forecasts, compared to the observed and spatially interpolated data. To constrain the uncertainties in the inputs, as well as model uncertainties, data-assimilation methods can be applied, such as for example the (Extended) Kalman filter, see Weerts and El Serafy (2006).

Conclusions and recommendations

In this paper the performance of ensemble flow forecasts for the Rhine basin is assessed using four different verification measures. Ensemble inputs were available from both the low resolution ECMWF-EPS global ensemble and the high-resolution COSMO-LEPS local area ensemble system.

Flow forecasts derived using the forecasts provided by the ECMWF-EPS ensemble show positive skill over climatology for lead times of up to nine days, proving the value of these medium-range ensemble forecasts. As expected, skill was found to deteriorate with lead time and a general increase of skill with catchment size was found for all lead times. Despite the value of the forecast, it was found that the resolution of the global ECMWF-EPS forecast was insufficient to properly represent the variability of precipitation and consequently flow forecasts. This was found for catchments of all sizes, even for the entire Rhine basin which has an area an order larger than that of the grid-cell size of the ECMWF forecast. While the basin itself is an order larger than the grid-cell size, this is not the case for the much smaller sub-basins that form part of the hydrological model used.

Although the available ensemble forecasts from COSMO-LEPS covered a much shorter period of time than the ECMWF-EPS forecasts, the increased resolution of the COSMO-LEPS local area model was found to provide a better representation of the variability, with higher skills across all catchment sizes, particularly in the forecast of short term precipitation. This confirms the need for downscaling of the ensemble forecasts to a more representative scale for the sub-basins in the hydrological model.

The accuracy of the ensembles of flow forecasts has been assessed using probabilistic verification measures that are common in the field of meteorology. With the help of skill scores and e.g. the reliability diagram different ensembles can be compared and improvements in the forecasting systems can be identified and measured. However, the verification information itself also provides useful information to the user of the forecasts as it gives that forecaster an expectation of model bias, and the forecaster can use this information effectively in establishing the confidence in a forecast. Particularly the reliability diagram displays most forecast properties and it can be prepared for any threshold and lead time of interest. Therefore it should be a tool for the decision maker, who would like the decision to be made partly with previous forecast performance in mind.

Acknowledgements

The project behind this paper has been initialized and financially supported by the Federal Institute of Hydrology, Germany and the Centre for Water Management of Rijkswaterstaat, The Netherlands. We thank A. Weerts (Deltares) for his assistence with this work. We further would like to acknowledge the work of the groups creating and running continuous meteorological forecasts products such as ECMWF-EPS and COSMO-LEPS. The paper has been improved by comments of two anonymous reviewers.

References

Bartholmes, J., Todini, E., 2005. Coupling meteorological and hydrological models for flood forecasting. Hydrology and Earth System Sciences 9 (4), 333–346.

Bergström, S., 2005. The HBV model. In: Singh, V. (Ed.), Computer Models of Watershed Hydrology. Water Resources Publications, Colorado, USA, pp. 443–476.

Broersen, P., Weerts, A., 2005. Automatic error correction of rainfall–runoff models in flood forecasting systems. In: Instrumentation and Measurement Technology Conference, 2005. IMTC 2005. Proceedings of the IEEE vol. 2, 963–968.

Buizza, R., 2005. EPS skill improvements between 1994 and 2005. ECMWF Newsletter 104, 10–14. <http://www.ecmwf.int/publications/newsletters/pdf/104.pdf>.

Buizza, R., Houtekamer, P., Toth, Z., Pellerin, G., Wei, M., Zhu, Y., 2005. A comparison of the ECMWF, MSC, and NCEP global ensemble prediction systems. Monthly Weather Review 133 (5), 1076–1097.

Clark, M., Hay, L., 2004. Use of medium-range numerical weather prediction model output to produce forecasts of streamflow. Journal of Hydrometeorology 5 (1), 15–32.

Cloke, H., Pappenberger, F., 2008. Evaluating forecasts of extreme events for hydrological applications: an approach for screening unfamiliar performance measures. Meteorological Applications 15 (1), 181.

de Roo, A., Gouweleeuw, B., Thielen, J., Bartholmes, J., Bongioannini-Cerlini, P., Todini, E., Bates, P., Horritt, M., Hunter, N., Beven, K., Pappenberger, F., Heise, E., Rivin, G., Hills, M., Hollingsworth, A., Holst, B., Kwadijk, J., Reggiani, P., van Dijk, M., Sattler, K., Sprokkereef, E., 2003. Development of a European flood early warning system. International Journal of River Basin Management 1, 49–59.

Dietrich, J., Trepte, S., Wang, Y., Schumann, A., Voß, F., Hesser, F., Denhard, M., 2008. Combination of different types of ensembles for the adaptive simulation of probabilistic flood forecasts: hindcasts for the Mulde 2002 extreme event. Nonlinear Processes in Geophysics 15 (2), 275.

Diomede, T., Marsigli, C., Nerozzi, F., Paccagnella, T., Montani, A., 2006. Quantifying the discharge forecast uncertainty by different approaches to probabilistic quantitative precipitation forecast. Advances in Geosciences 7, 189–191.

Disse, M., Engel, H., 2001. Flood events in the Rhine basin: genesis, influences and mitigation. Natural Hazards 23 (2), 271–290.
Eberle, M., 2001. Hydrological Modelling in the River Rhine basin. Part II Report on hourly modelling. Federal Institute for Hydrology, Koblenz, Germany.
Eberle, M., Buiteveld, H., Wilke, K., Krahe, P., 2005. Hydrological Modelling in the River Rhine Basin Part III – Daily HBV Model for the Rhine Basin. Tech. Rep. Federal Institute for Hydrology, Koblenz, Germany.
Fortin, V., Favre, A., Saïd, M., 2006. Probabilistic forecasting from ensemble prediction systems: improving upon the best-member method by using a different weight and dressing kernel for each member. Quarterly Journal of the Royal Meteorological Society 132 (617), 1349–1369.
Gouweleeuw, B.T., Thielen, J., Franchello, G., de Roo, A.P.J., Buizza, R., 2005. Flood forecasting using medium-range probabilistic weather prediction. Hydrology and Earth System Sciences 9, 365–380.
IKSR, 2005. Internationale Flussgebietseinheit Rhein, Merkmale, Überprüfung der Umweltauswirkungen menschlicher Tätigkeiten und wirtschaftliche Analyse der Wassernutzung. Internationale Kommission zum Schutz des Rheins (IKSR) Bericht Teil A.
Jaun, S., Ahrens, B., 2009. Evaluation of a probabilistic hydrometeorological forecast system. Hydrology and Earth System Sciences Discussions 6 (2), 1843–1877.
Kirk, E., Fraedrich, K., 1990. Prognose der Niederschlagswahrscheinlichkeit – Modelle und Verifikation. Meteorologisches Institut Universität Hamburg. <http://www.mi.uni-hamburg.de/Kurzfrist-Vorhersagen.213.0.html>.
Lorenz, E.N., 1963. Deterministic nonperiodic flow. Journal of Atmospheric Sciences 20, 130–141.
Marsigli, C., Boccanera, F., Montani, A., Paccagnella, T., 2005. The COSMO-LEPS mesoscale ensemble system: validation of the methodology and verification. Nonlinear Processes in Geophysics 12, 527–536.
Molteni, F., Buizza, R., Palmer, T.N., Petroliagis, T., 1996. The ECMWF ensemble prediction system: methodology and validation. Quarterly Journal of the Royal Meteorological Society 122, 73–119.
Montani, A., Capaldo, M., Cesari, D., Marsigli, C., Modigliani, U., Nerozzi, F., Paccagnella, T., Patruno, P., Tibaldi, S., 2003a. Operational limited-area ensemble forecasts based on the Lokal Modell. ECMWF Newsletter 98, 2–7.
Montani, A., Marsigli, C., Nerozzi, F., Paccagnella, T., Tibaldi, S., Buizza, R., 2003b. The Soverato flood in Southern Italy: performance of global and limited-area ensemble forecasts. Nonlinear Processes in Geophysics 10, 261–274.
Parmet, B., Sprokkereef, E., 1997. Hercalibratie Model Lobith – Onderzoek naar mogelijke verbeteringen van de voorspellingen met het Meervoudig Lineaire Regressie Model Lobith na de hoogwaters van 1993 en 1995. RIZA rapport 97.061, Lelystad.
Regimbeau, F., Habets, F., Martin, E., Noilhan, J., 2007. Ensemble streamflow forecasts over France. ECMWF Newsletter 111, 21–27.
Roulin, E., 2007. Skill and relative economic value of medium-range hydrological ensemble predictions. Hydrol. Earth Syst. Sci. 11 (2), 725–737.
Walser, A., 2006. COSMO-LEPS forecasts for the August 2005 floods in Switzerland. COSMO-LEPS Newsletter 2006, 142–145.
Weerts, A., El Serafy, G., 2006. Particle filtering and ensemble Kalman filtering for state updating with hydrological conceptual rainfall–runoff models. Water Resources Research 42 (9), W09403.
Werner, M., Reggiani, P., Roo, A.D., Bates, P., Sprokkereef, E., 2004. Flood forecasting and warning at the river basin and at the European scale. Natural Hazards 36 (1), 25–42.
Wilks, D.S., 1995. Statistical methods in the atmospheric sciences : an introduction. International Geophysics Series, vol. 59. Academic Press, San Diego.
Wilks, D.S., 2006. Statistical methods in the atmospheric sciences: an introduction, second ed. International Geophysics Series, vol. 91, pp. 648.
WMO, 2007. Forecast verification – issues, methods and faq. WWRP/WGNE Joint Working Group on Verification. <http://www.bom.gov.au/bmrc/wefor/staff/eee/verif/verif_web_page.html>.