Hydrol. Earth Syst. Sci., 13, 1031–1043, 2009
www.hydrol-earth-syst-sci.net/13/1031/2009/

Hydrology and
Earth System
Sciences

Evaluation of a probabilistic hydrometeorological forecast system

S. Jaun[1,2] **and B. Ahrens**[3]

[1]Institute for Atmospheric and Climate Science, ETH Zurich, Switzerland
[2]Swiss Federal Research Institute WSL, Birmensdorf, Switzerland
[3]Institute for Atmosphere and Environment, Goethe-University Frankfurt a. M., Germany

Received: 11 February 2009 – Published in Hydrol. Earth Syst. Sci. Discuss.: 6 March 2009
Revised: 30 June 2009 – Accepted: 30 June 2009 – Published: 8 July 2009

Abstract. Medium range hydrological forecasts in mesoscale catchments are only possible with the use of hydrological models driven by meteorological forecasts, which in particular contribute quantitative precipitation forecasts (QPF). QPFs are accompanied by large uncertainties, especially for longer lead times, which are propagated within the hydrometeorological model system. To deal with this limitation of predictability, a probabilistic forecasting system is tested, which is based on a hydrological-meteorological ensemble prediction system. The meteorological component of the system is the operational limited-area ensemble prediction system COSMO-LEPS that downscales the global ECMWF ensemble to a horizontal resolution of 10 km, while the hydrological component is based on the semi-distributed hydrological model PREVAH with a spatial resolution of 500 m.

Earlier studies have mostly addressed the potential benefits of hydrometeorological ensemble systems in short case studies. Here we present an analysis of hydrological ensemble hindcasts for two years (2005 and 2006). It is shown that the ensemble covers the uncertainty during different weather situations with appropriate spread. The ensemble also shows advantages over a corresponding deterministic forecast, even under consideration of an artificial spread.

1 Introduction

Recent flood events (e.g., the Alpine flood of August 2005, see Bezzola and Hegg, 2007; Jaun et al., 2008; Hohenegger et al., 2008) showed the vulnerability of the infrastructure we depend on. To reduce flood damages by taking appropriate precautions, long lead times (several days) in hydrological forecasting are needed, which is only possible with the use of medium range weather forecasts in a coupled hydrometeorological model chain. Especially the large uncertainties in precipitation forecasting affect the accuracy and reliability of the resulting hydrological forecast. As it would be imprudent to simply ignore these uncertainties (Pappenberger and Beven, 2006), they have to be forecasted too. For this purpose, probabilistic forecasts can be applied (Ehrendorfer, 1997).

In meteorology, probabilistic ensemble forecasts have been established for operational forecasts some time ago. Meteorological ensemble prediction systems (EPSs) are operationally available at the global scale from, for e.g., the US National Center for Environmental Predictions (NCEP, Toth and Kalnay, 1997), the European Centre for Medium Range Weather Forecasts (ECMWF, Molteni et al., 1996) and the Meteorological Center of Canada (MSC, Houtekamer et al., 1996). From these ensemble forecasts, a measure of the forecast uncertainty can be gained in terms of the ensemble spread. The spread of the ensemble members represents mainly the initialization uncertainty of the meteorological model, the main source of uncertainty for large-scale atmospheric circulation patterns in forecasts up to about five days (Buizza, 2003).

A number of case studies were conducted, which directly use the output from a global scale EPS to drive hydrological models (e.g., Bartholmes and Todini, 2005; Pappenberger et al., 2005; Roulin and Vannitsem, 2005; Siccardi et al., 2005; Rousset et al., 2007; Komma et al., 2007). While demonstrating promising results, some of the case studies suffer from biases related to the coarse resolution of the meteorological model (and depend in turn on the scale the hydrological model applied). The large scale meteorological

Correspondence to: S. Jaun
(simon.jaun@wsl.ch)

Published by Copernicus Publications on behalf of the European Geosciences Union.

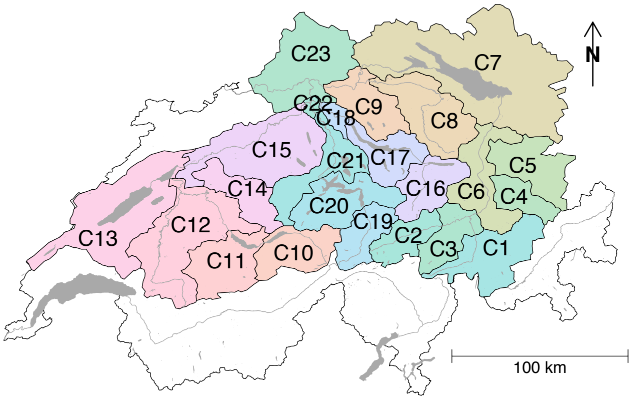

Fig. 1. Catchment overview, showing the defined catchments with respective identifier (C1,..., C23, cf. Table 1) upstream of the Rheinfelden gauge (published in Jaun et al., 2008).

models are not accurate at modeling local weather, because local sub-grid scale features and dynamics are not resolved, especially in regions with complex topography. To overcome this limitation, global-scale EPS forecasts can be dynamically downscaled by use of a limited area numerical weather model (e.g., the COSMO model, nested into the ECMWF ensemble as described in the following section). Mass et al. (2002) showed that such a refinement from a grid spacing of 36 km to a grid spacing of 12 km results in better forecasts, as it allows the definition of the major topographical features of the region and their corresponding atmospheric circulation. A dynamical downscaling of the global meteorological forecasts is expensive in terms of computing resources, and thus it is not feasible to downscale the full ensemble for everyday operational applications. Therefore, the ensemble size is normally reduced and only a subset of the ensemble members is used (Molteni et al., 2001). This approach has successfully been used for several hydrological case studies (e.g., Verbunt et al., 2007; Jaun et al., 2008). In contrast, statistical downscaling approaches, like the use of meteorological analogues, rely heavily on the availability of long historical data sets and do not appear to be suitable to provide useful information about the future small-scale streamflow by itself (Diomede et al., 2008), especially in the case of extreme events.

The aforementioned publications on evaluation of hydrological ensemble forecast systems have been limited to flood case studies and/or single catchments. Only recently have larger data sets become available. Olsson and Lindström (2007) and Bartholmes et al. (2009) provide analysis of extended time series over large areas (Sweden and Europe, respectively), both of which use direct output from the global scale ECWMF EPS to drive the hydrological model. An extensive list of recent studies applying ensemble approaches for runoff forecasts can be found in Cloke and Pappenberger (2009), together with a review of ensemble techniques.

Table 1. Catchment identifiers with names of the respective rivers and gauges as well as the size of the catchments.

identifier	river	gauge	size [km^2]
C1	Hinterrhein	Furstenau	1575
C2	Vorderrhein	Ilanz	776
C3	Rhine	Domat-Ems	3229
C4	Landquart	Felsenbach	616
C5	Ill	Gisingen (A)	1281
C6	Rhine	Diepoldsau	6119
C7	Rhine	Neuhausen	11 887
C8	Thur	Andelfingen	1696
C9	Rhine	Rekingen	14 718
C10	Aare	Ringgenberg	1129
C11	Aare	Thun	2490
C12	Aare	Hagneck	5128
C13	Aare	Brugg-Agerten	8217
C14	Emme	Wiler	939
C15	Aare	Brugg	11 750
C16	Linth	Weesen	1061
C17	Limmat	Zurich	2176
C18	Limmat	Baden	2396
C19	Reuss	Seedorf	832
C20	Reuss	Luzern	2251
C21	Reuss	Mellingen	3382
C22	Aare	Untersiggenthal	17 625
C23	Rhine	Rheinfelden	34 550

This paper investigates the applicability of a high-resolution meteorological-hydrological ensemble system, using the dynamical downscaling approach for two continuous years (2005 and 2006). The study area consists of the upper Rhine basin, encompassing an overall area of 34 550 km^2. To account for inhomogeneities in topography, atmospheric processes and runoff regimes, the domain is divided into 23 subcatchments with a typical size of 900 to 1600 km^2 (cf. Fig. 1 and Table 1), based on the setup described in Verbunt et al. (2006). In addition to the analysis of selected catchments, the full extent of the study area is considered.

Besides the input uncertainty (uncertainty from the meteorological data used to drive the hydrological model), which is addressed by the use of the meteorological ensemble, two additional components affect the output uncertainty of a hydrological model: the initialization uncertainty (i.e., the initial state of the model) and the model uncertainty itself (uncertainty from parameters and the conceptualization, Vrugt et al., 2005). In this work, the main focus remains on the input uncertainty, as forecasted meteorological data is regarded as the most uncertain component (Todini, 2004).

2 Methods

The meteorologic-hydrologic model chain used is the same as described in Jaun et al. (2008), where it was used for a case study of the extreme event in August 2005. For the meteorological component, either an ensemble forecast system or a deterministic forecast system, providing a single model realization, is applied. Forecast setups and strategies were adopted and applied to the operationally available meteorological forecasts for the years 2005 and 2006.

2.1 Deterministic and probabilistic meteorological forecast systems

The deterministic meteorological forecasts are provided by the operational weather forecast model COSMO-7. This model is the MeteoSwiss implementation of the COSMO model (COnsortium for Small-scale Modeling, Steppeler et al., 2003), which is nested in the global deterministic forecast model from ECMWF. COSMO-7 uses a horizontal grid-spacing of 0.0625 degrees (7 km) and 45 model levels. Six meteorological variables (temperature, precipitation, humidity, wind, sunshine duration derived from cloud cover, global radiation) are further downscaled to 500 m grid-spacing (using bilinear interpolation, temperature adjusted according to elevation by adopting a constant lapse rate of 0.65°C/100 m), to meet the grid size requirements of the hydrological model.

The global meteorological ensemble is provided by the operational global atmospheric EPS of ECMWF and consists of 51 members. The generation of this ensemble is based on singular vectors to create optimally perturbed initial states (Buizza and Palmer, 1995). This global ensemble is downscaled by the limited-area EPS COSMO-LEPS (Marsigli et al., 2005; Montani et al., 2003). Due to computational constraints, the operational COSMO-LEPS refines a subsample of 10 (16 from February 2006) representative global ensemble members only, selected by a cluster analysis (Molteni et al., 2001). Prior to the clustering analysis, the preceding global EPS simulation from the previous day is combined with the actual forecast. Hence the clustering is applied to a recombined ensemble consisting of 102 members. This procedure, using "old" forecast information, generally results in a widening of the spread of the reduced ensemble. The clustering identifies similar circulation patterns based on the analysis of wind, geopotential height and humidity on three pressure levels (500 hPa, 700 hPa, 850 hPa) for two lead times (96 h, 120 h). From the resulting 10 (16) clusters, the respective representative cluster members (RMs) are selected and dynamically downscaled over a domain covering central and southern Europe. These ensemble members are run on a rotated spherical grid with a horizontal grid-spacing of $0.09° \times 0.09°$, equivalent to about $10 \times 10\,\mathrm{km}^2$, and with 32 (40 from February 2006) model levels. The meteorological variables of the resulting high-resolution meteorological ensemble are treated analogous to the COSMO-7 variables. The cluster sizes can optionally be used to weight the representative members of COSMO-LEPS.

2.2 The hydrological model

The semi-distributed hydrological model PREVAH (Viviroli et al., 2009) is then driven by COSMO-LEPS with hourly time steps. PREVAH (Preciptation Runoff EVApotranspiration Hydrotope) uses hydrologic response units (HRUs, Flügel, 1997) and the runoff generation module is based on the conception of the HBV-model (Bergström and Forsman, 1973; Lindström et al., 1997), adapted to a spatially distributed application. Further information on the model physics, structure, interpolation methods and parameterisations can be found in Gurtz et al. (1999), Gurtz et al. (2003) and Zappa et al. (2003). The initial conditions of the hydrological model are obtained from a continuous reference simulation driven by meteorological observations, subsequently referred to as HREF. No additional perturbations were realised at the level of the hydrological model, e.g., consideration of initialization uncertainties.

The use of the deterministic meteorological forecast variables as input to PREVAH results in a deterministic hydrological forecast subsequently referred to as HDET. The coupling of PREVAH with COSMO-LEPS provides probabilistic hydrological forecasts in terms of a hydrological EPS (HEPS).

2.3 Set-up of simulations

Hindcasts were conducted daily for both years, 2005 and 2006, and for the deterministic forecasts as well as the ensemble set-up. The deterministic forecasts from COSMO-7 provide a forecast range of 72 h (3 days) and are initialized at 00:00 UTC.

The meteorological EPS forecasts are initialized at 12:00 UTC and span 132 h (120 h until June 2005). The first 12 h are not considered for the hydrological coupling, which is initialized at 00:00 UTC, resulting in a forecast range of 120 h (108 h) for HEPS. This cutoff considers the temporal availability of the operational ensemble forecasts and eases comparison to the deterministic forecast. To ensure consistency for the differing HEPS forecast ranges over the considered time period, the analysis of HEPS was restricted to 96 h (4 days).

For the quantitative analysis we focus on daily runoff values. The hindcasts are chained for the respective forecast ranges (0–24 h, 24–48 h, 48–72 h, 72–96 h), resulting in four (three for HDET) daily time series, which are accounted for separately and compared to each other. In the case of HEPS, we therefore get four time series consisting of daily ensembles derived from the summed up hourly values of the respective individual ensemble member within the forecast ranges. All calculations, e.g., the estimation of the ensemble interquartile range (IQR) are based on these daily values. Examples of chained daily runoff hindecasts are shown in Fig. 2a to Fig. 3.

2.4 Validation methodology

To evaluate the skill of the forecasts, score measures are applied. These are complemented by the evaluation of general ensemble properties to verify the statistical appropriateness of the probabilistic forecast (Laio and Tamea, 2007).

Yearly discharges were estimated for all catchments, in order to assess the representation of runoff volumes by the model chain. To test the general performance of the ensemble, a method used by the ECMWF for meteorological EPS verification was adopted (Lalaurette et al., 2005). Assuming a perfect probabilistic forecast with symmetric error quantiles, the following spread skill relation should be found: the absolute difference between the ensemble median and the verifying simulation should exceed half the interquartile range (referred to as spread) in exactly 50% of the cases. Therefore, for a theoretical perfect probabilistic forecast, averaging over spread categories should result in a diagonal relationship. Evaluating HEPS, deviations from this diagonal relationship will show whether the ensemble produces too high/low spread to cover the associated ensemble median error. As the assumption that error quantiles are symmetrical is not met in this application, positive and negative errors are accounted separately (cf. Lalaurette et al., 2005). Other methods for spread-skill evaluations, e.g. conducted by Scherrer et al. (2004), are based on the use of a skill score, e.g. the ensemble RMSE or the Brier skill score, which is compared to a measure of spread. The resulting relationship is then interpreted with respect to the relationship which results from a ”perfect” forecast (e.g. from a toy model).

In addition to the spread-skill evaluation, the rank histogram (Anderson, 1996) of the probabilistic runoff forecast is evaluated, to check whether the ensembles include the observations being predicted as equiprobable members (consistency condition). If rank uniformity is not met, this can reveal deficiencies in ensemble calibration, or reliability (Wilks, 2006). In difference to the spread-skill relation, the rank histogram allows a distinction between bias and under-/overdispersion, but does not account for relative ensemble error.

To perform a probabilistic verification of the time series within the time window considered, we use the ranked probability skill score (RPSS) described in Wilks (2006). This score is widely used for the evaluation of probabilistic forecasts in meteorological sciences (e.g., Weigel et al., 2007a; Ahrens and Walser, 2008).The RPSS is based on the ranked probability score (RPS). The RPS is a squared measure that compares the cumulative density function of a probabilistic forecast with that of the corresponding observation over a given number of discrete probability categories. Thus the RPS measures how well the probabilistic forecast predicts the category in which the observation is found. For a given forecast-observation pair, the RPS is defined as

$$\mathrm{RPS} = \sum_{k=1}^{K} \left[\sum_{j=1}^{k} py_j - \sum_{j=1}^{k} po_j \right]^2, \tag{1}$$

where K is the number of forecast categories, py_j is the predicted probability in forecast category j, and po_j the observation in category j (0=no, 1=yes). The RPS is bounded by zero and $K-1$. While a perfect forecast would result in RPS=0, less accurate forecasts receive higher sores. By averaging the RPS over a number of forecast–observation pairs, these can be jointly evaluated, resulting in the mean $\langle\mathrm{RPS}\rangle$.

The RPSS is finally obtained by relating the $\langle\mathrm{RPS}\rangle$ of the forecast to the $\langle\mathrm{RPS}_{\mathrm{ref}}\rangle$ of a reference forecast according to

$$\mathrm{RPSS} = 1 - \frac{\langle\mathrm{RPS}\rangle}{\langle\mathrm{RPS}_{\mathrm{ref}}\rangle}. \tag{2}$$

The RPSS can take values in the range $-\infty \leq \mathrm{RPSS} \leq 1$. Whereas RPS$>$0 indicates an improvement over the reference forecast, a forecast with RPSS$\leq$0 lacks skill with respect to the reference forecast.

In this paper we chose climatological quantiles derived from 10 years of runoff data as catchment specific thresholds for the RPSS categories. Apart from the quartiles (0.25, 0.5, 0.75) we additionally selected the 0.95 quantile to better resolve higher runoff occurrences. For $\langle\mathrm{RPS}_{\mathrm{ref}}\rangle$ we use the climatological probabilities of the mentioned quantiles.

For the two years of forecasts considered (2005 and 2006), we are faced with different ensemble sizes (10 and 16, respectively). From Müller et al. (2005), it is known that the RPSS is negatively biased for ensemble prediction systems with small ensemble sizes. The influence of the differing ensemble sizes is assessed by the additional use of a debiased version of the RPSS (RPSSd, after Weigel et al., 2007a).

Other probabilistic evaluation methods such as the Brier skill score (BSS, the probabilistic equivalent to the mean squared error), the reliability diagram or the relative operating characteristic (ROC), as described in Wilks (2006), are not considered as they require a single evaluation threshold, whereas the categorical evaluation by the RPSS allows a better judgement of the evolution of the hydrograph over an extended time period.

For evaluation of the deterministic hydrological forecast HDET, the Nash-Sutcliffe coefficient (E) (Nash and Sutcliffe, 1970) is applied, which is widely used for hydrological verification purposes (Legates and McCabe, 1999). The usual formulation of E is given by

$$\mathrm{E} = 1 - \frac{\sum_{t=1}^{n} (o_t - y_t)^2}{\sum_{t=1}^{n} (o_t - \bar{o})^2}, \tag{3}$$

where y_t and o_t denote the forecasted and observed time series, respectively, and $\bar{o}$ the mean of the observations over the forecast period. E can take values in the range $-\infty \leq \mathrm{E} \leq 1$, with E$>$0 indicating an improvement over a forecast with the observed mean discharge, while E$\leq$0 shows no additional skill. E can also be interpreted as the coefficient of determination, representing the fraction of variability in o_t that is contained in y_t.

Direct comparison of the performance of HDET and HEPS is difficult to achieve. An evaluation of HDET by means of the RPSSd is not carried out, as the RPSSd does not directly quantify whether a specific forecast is more skillful, but rather is a measure for the gain in potentially usable ensemble information (Weigel et al., 2007b). The RPSS in turn suffers most from its negative bias in the deterministic case ("one member ensemble"). An evaluation based on a deterministic skill score like E implies the conversion of the probabilistic forecast into a deterministic one, e.g., by use of the ensemble median. As such a conversion implies a loss of valuable forecast information and can bias the ensemble performance in specific cases, it should not be applied for the sake of a simplified forecast interpretation (Wilks, 2006). Nevertheless we carry out a comparative deterministic evaluation of the HEPS median against HDET. If this evaluation reveals that the HEPS median performs equal or better than HDET over an extended time period, a first indication of an added value of the ensemble forecast system is given.

To further challenge the ensemble forecast system, it was tested against an artificial ensemble (HART) by means of the RPSS. HART is based on the climatological properties of HEPS, assuming a linear correlation between the ensemble median and the individually sorted ensemble members (separately for the different catchments and lead times). That means, for a specific catchment and lead time, as an example, the daily forecasts are sorted by runoff values in ascending order. Correlating the lowest member (second lowest,..., highest) of all daily forecasts with the according HEPS median results in a linear relationship. Applying these linear correlations, the daily artificial members are then constructed by use of the HEPS median. Consequently, spread and range of the artificial ensemble mainly depend on the actual runoff quantity of the HEPS median.

This evaluation reveals whether the ensemble forecast performance is better than a deterministic forecast with climatological ensemble spread. If HEPS shows no advantage over HART, the value of the ensemble forecast is at least questionable with regard to a deterministic forecast system that considers some sort of uncertainty information. Please note that this evaluation only reveals the minimal added value, as the median of the ensemble is used as base for HART. Consequently, HART contains ensemble information that is not available in the case of a deterministic forecast.

If the HEPS median outperforms HDET (in terms of E) and HEPS outperforms HART (in terms of RPSS), we can confidently state an added value of the ensemble forecast system, provided that the probabilistic evaluation of HEPS, including the general ensemble performance, shows positive results.

Apart from direct evaluations against runoff observations, we substitute the runoff observations by the reference simulation HREF to eliminate the additional uncertainties introduced by the hydrological model.

3 Results and discussion

3.1 Analysis of a selected catchment

Figures 2a to c allow us to discuss important features of probabilistic hydrologic forecasts. Daily hindcasts (using HEPS) were conducted for the years 2005 and 2006, which were then chained for selected forecast ranges. As an example, graphs of chained daily hindcasts for the ranges 72–96 (Fig. 2a), 48–72 (Fig. 2b) and 24–48 (Fig. 2c) hours are shown at the Brugg (Aare) gauge for 2005. The ensemble IQR generally encompasses the reference simulation driven by observed meteorological inputs and also the measured runoff. Also, the ensemble range is much larger during flood peaks, representing the additional uncertainties during unstable weather situations.

Comparisons against the area-mean precipitation of the ensemble members (used as input for the hydrologic model, not shown) show the expected reduction in variability and amplitude due to hydrological processes. Note that there does not appear to be a problem with an overprediction of flood events (e.g., an event with a return period of two years is forecasted with some probability a few times in 2005) or a constantly large spread. For decreasing lead times, the HEPS full range and HEPS interquartile range decrease gradually and constrict around the reference simulation as expected.

Figure 3 shows the chained daily hindcasts with HART for the same catchment as in Fig. 2a for the forecast lead time of 72–96 h. Compared to the corresponding Fig. 2a, peaks in spread and range are less distinctive. While the uncertainty seems to be well covered during runoff peaks, it remains constantly high during recession periods, as the simple synthetic ensemble construction cannot distinguish between inclining and declining phases of the runoff peaks.

3.2 Evaluation of yearly discharge

Figure 4 shows the yearly discharges for two example catchments (C12, C21) and catchment C23, which captures the out-flow from all catchments and can therefore be used as an indicator for the entire study area. Yearly discharge sums of daily range, IQR, and median for different lead times (0–24 h, 24–48 h, 48–72 h, 72–96 h) are compared to the respective values of HDET, HREF and measured runoff values. Figure 4 summarizes Fig. 2a to Fig. 2c and allows for simple and straightforward comparison between catchments and lead times.

While HEPS IQR nicely encompasses HREF and HDET, the HEPS IQR does not contain the observations ideally. The yearly bias in volume from HREF to OBS visible in Fig. 4 is +5% for catchment C12 and −10% for catchment C21. The

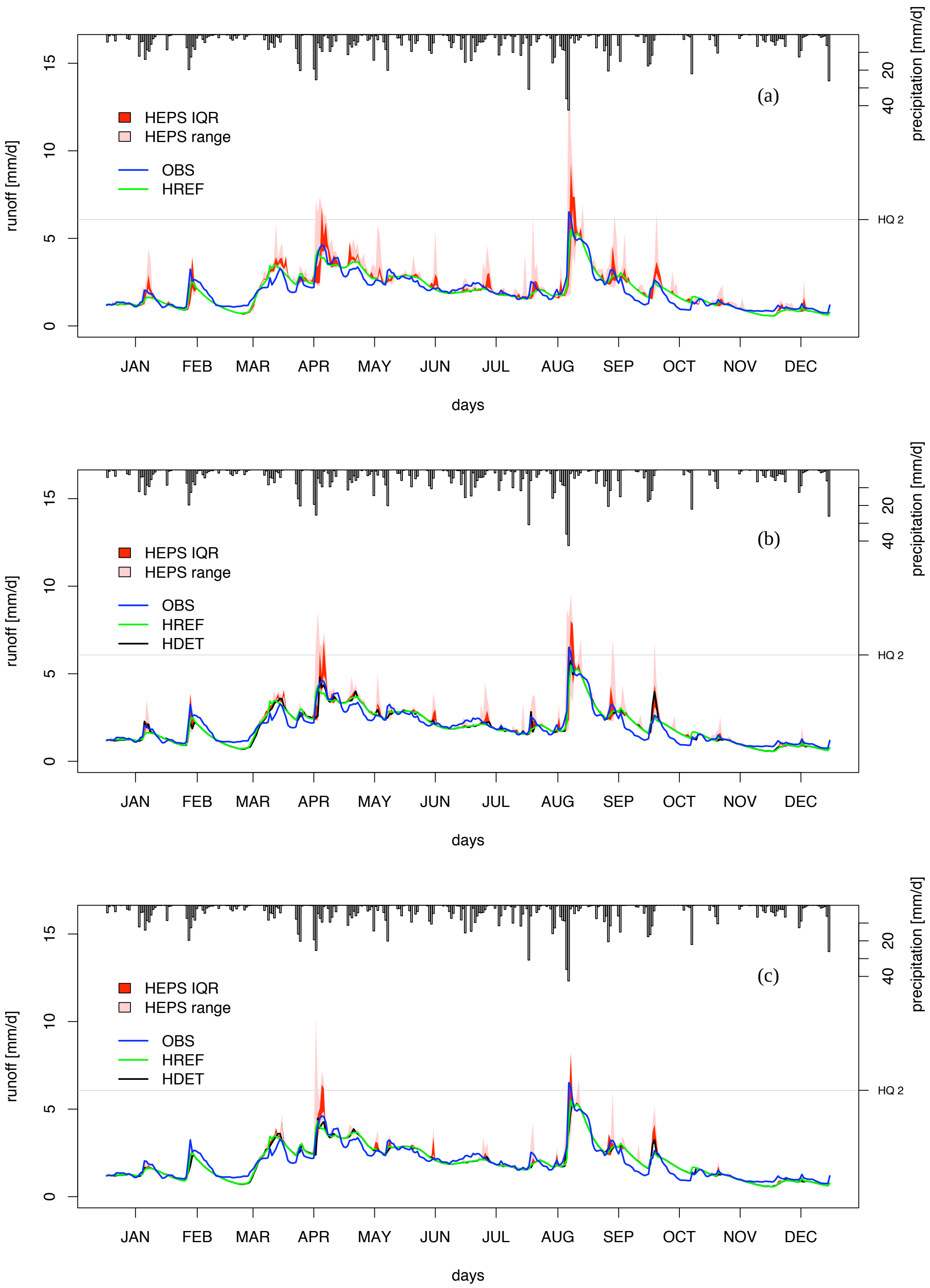

Fig. 2. Chained daily runoff hindcast at the gauge Brugg (Aare, 11 538 km^2), with a lead time of **(a)** 72–96 h, **(b)** 48–72 h and **(c)** 24–48 h in 2005. Measured runoff is plotted in blue. The light red area shows the full range of the HEPS simulation (HEPS range) and the red area represents the IQR of the same simulation (HEPS IQR). Spatially interpolated observed precipitation is plotted from the top. In (b) and (c) HDET is additionally marked in black. HQ 2 marks the two year recurrence period.

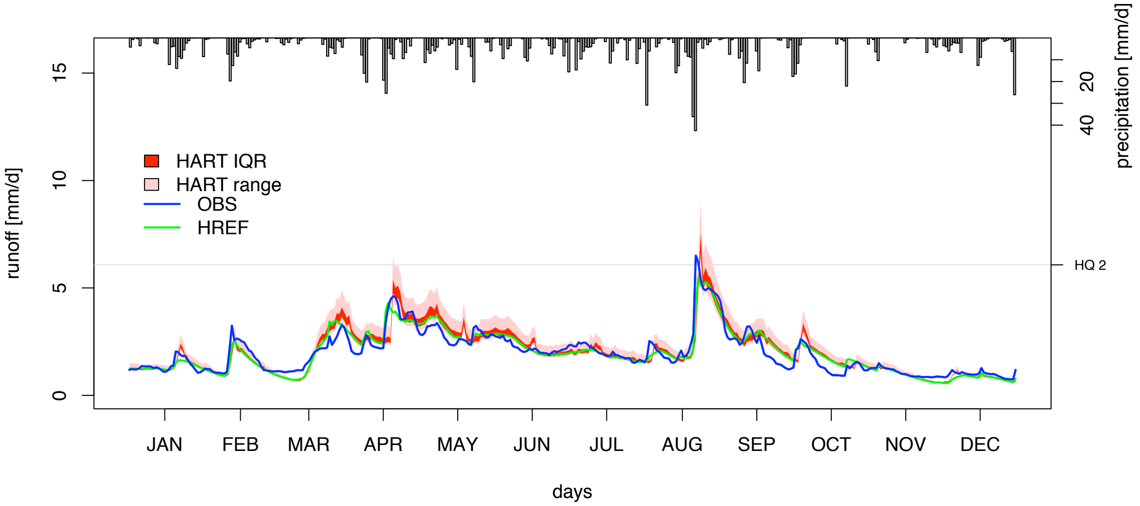

Fig. 3. Same as Fig. 2a, but for HART instead of HEPS.

overall bias (median over catchments C1 to C23) is −6%, which is reflected in the bias of catchment C23 (−6%).

HDET and the HEPS median show very similar performance when compared to HREF, with a slightly pronounced tendency of HDET to overforecast for the lead time 48–72 h.

Comparing 2005 and 2006 gives qualitatively very similar results, differing mainly in the observed yearly runoff sum (median increase of 3.5% from 2005 to 2006), changes in the bias of HREF to OBS (reflecting the skill of the hydrological model itself), a wider total range of the 16 member HEPS compared to the 10 member HEPS, but very similar IQR. As the relative positions of HREF, HDET and HEPS do not change between 2005 and 2006, the change of the HREF bias should not be neglected for inter-annual comparison of the forecast performance (compensation/amplification through overforecasting when compared to measured runoff).

For catchment C23, spread of HEPS shows a distinct reduction in ensemble spread and error for all lead times (2005 and 2006), which is visible in Fig. 4 as well as after normalization by measured runoff (not shown). On the one hand this indicates the overall decrease in uncertainty for forecasts over larger areas (i.e., differences in forecasts for small catchments even out over larger areas). On the other hand this is a result of the increase in the time of concentration. In the case of a forecasted large scale event (or a local event in the northern part of Switzerland), the contributing catchment area for C23 grows quickly and shows up in the (small) ensemble spread. A forecast of a local event in an alpine catchment will not be reflected in the ensemble spread of C23 (for short lead times due to the time of concentration, for longer lead times due to averaging). However, this is the real forecast situation and as we treat all forecasts the same way, none of them should benefit.

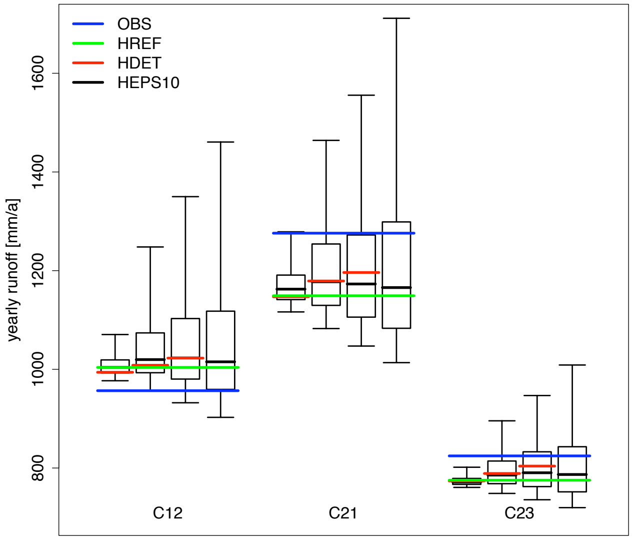

Fig. 4. Total discharge for 2005. The observed runoff (blue), HREF (green), HDET (red), and ensemble forecasts (black) are shown for three catchments (C12, C21 and C23). The ensemble forecasts are illustrated by box-whisker-plots. All displayed forecasts are shown with resp. lead times (1, 2, 3 and 4 days from left to right).

3.3 Verification of general ensemble properties

The scatter diagram for runoff comparing the ensemble spread and absolute error of the HEPS median is given in Fig. 5a for observed runoff and in Fig. 5b for HREF. Both figures show daily values for the year 2005 (72–96 h lead time). All catchments (cf. Table 1) are included. Analogous to the results from the EPS verification (Lalaurette et al., 2005), the coherence between ensemble spread and error exists for both choices of runoff reference (HREF and OBS). Large day-to-

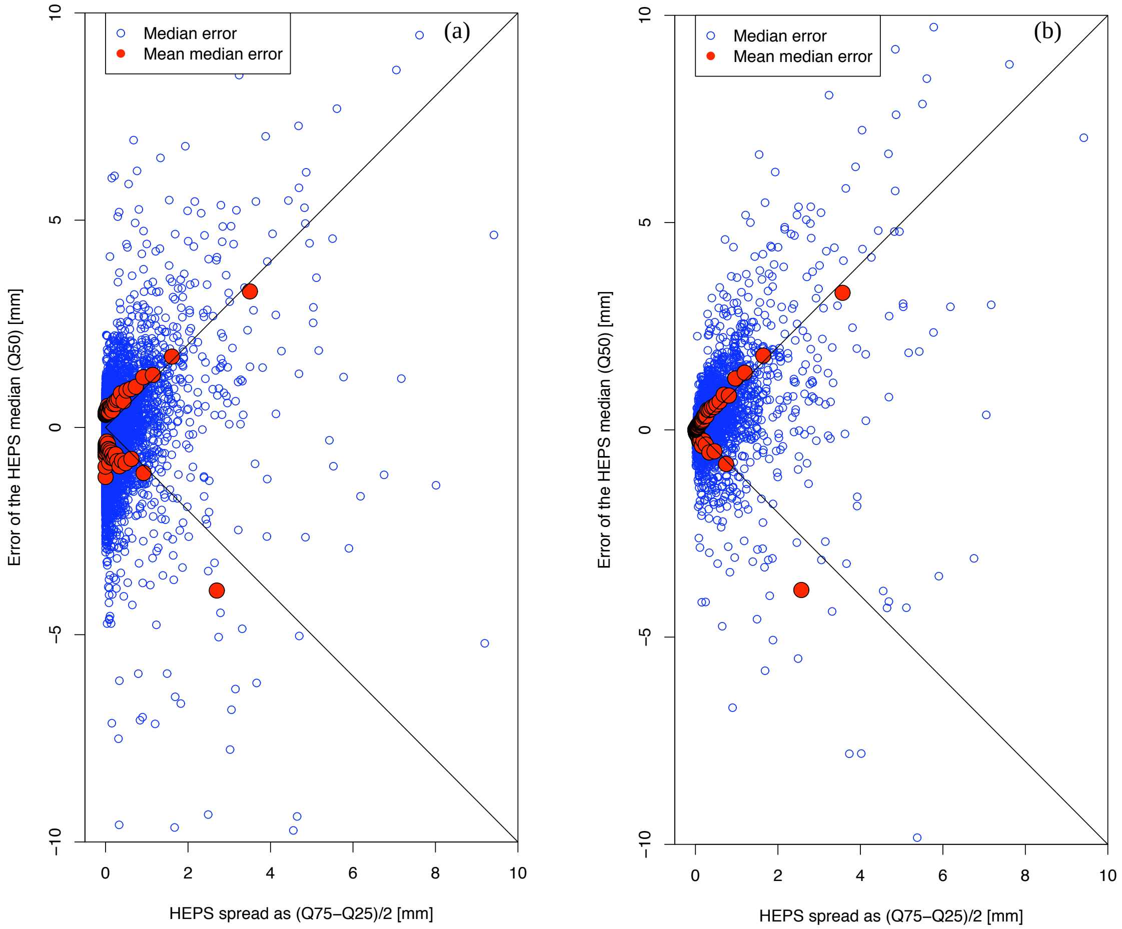

Fig. 5. The HEPS median error from observed runoff is compared to the half interquartile HEPS range for daily runoff (72–96 h hindcasts) for catchments C1 to C23 in 2005. **(a)** shows the evaluation against observed runoff, **(b)** against HREF. The empty blue circles represent the daily values, while the filled red circles show the means of the spread categories, averaged over 100 daily values. Positive and negative errors are considered separately.

day variations occur within the shown relation, but the statistical relationship that should exist, when gathering a large sample of cases with similar spread (81 spread categories, each containing 100 daily values from the 23 considered catchments), holds and the distribution of errors within each spread category is centered around the diagonal. This evaluation shows that additional uncertainty is reasonably represented by an increase in spread (also cf. Fig. 2a).

A closer examination reveals that Fig. 5a shows a tendency to underestimate spread in the forecasts (especially for small and negative median errors). This is clearly reduced in Fig. 5b, reflecting the bias of HREF with regard to observed runoff (cf. Fig. 4). However, even with HREF, large negative errors are still not met by a sufficiently wide ensemble spread for longer lead times. This underestimation of spread disappears with shorter lead times and results from a single event (August 2005). Excluding the period of this event also removes the underestimation of spread for large negative errors. The analysis for the year 2006 yields very similar results for the lead time of 72–96 h, although spread is overestimated for large negative errors with HREF. The overestimation in spread remains with shorter lead times, which might be a result of the changed ensemble configuration.

Note that the considered period for evaluation of general ensemble properties is to short to allow for robust statistics. As shown above, features of the relation can be dominated by single events or catchments. Therefore the spread-skill relation should not be interpreted on its own.

Figure 6a shows that the ensemble forecasts for the year 2005 do not satisfy the consistency condition (i.e., the ensembles do not reflect equiprobability of observations within their distributions). This is also the case for the year 2006 and the following remarks apply to both years considered. The U-shaped rank histogram indicates an under-dispersion (over-confidence) of the ensembles, as the observations are too frequently falling into the low and high ranks, resulting

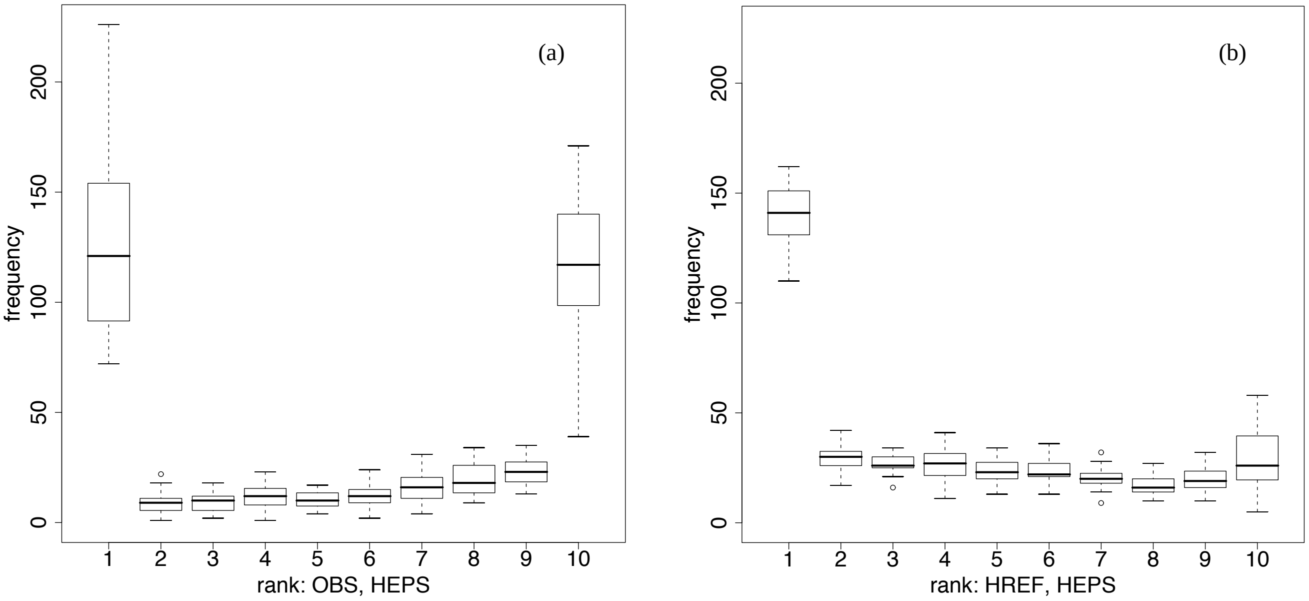

Fig. 6. Modified rank histogram showing combined results for all catchments. The rank of the observed runoff within HEPS is shown for the leadtime of 72–96 h in 2005. **(a)** shows the evaluation against observed runoff, **(b)** against HREF.

in an overpopulation of the extreme ranks. Evaluation of the rank histogram of HREF and HEPS (Fig. 6b) reveals that the overpopulation of high ranks is mostly a result of the uncertainty introduced by the hydrological model. This bias is also visible in Fig. 4 for C21 and C23. Still, the overpopulation of low ranks remains. We argue, that this is not only due to the slight overestimation of HEPS over HREF (Fig. 4), but also a result of the characteristics of the verified variable. We are considering a variable that is non-normally distributed, as the hydrograph and its evolution is bound by the baseflow. Runoff forecasts that lead to an increase in runoff are therefore less constrained, which explains the overforecasting bias in Fig. 6b. This hypothesis is supported by the skewness of HEPS towards lower runoff values in Fig. 4.

For shorter lead times than those shown (72–96 h), the rank histograms for HREF and observations show an increase in frequencies at the highest rank. Indeed, the spread of the ensemble narrows with reduced lead times. The narrowing in spread mostly represents the increase in predictability. The indicated overconfidence of the ensemble for shorter lead times is probably an effect of the ensemble generation method, focusing on optimal spread for midrange forecasts. It should be noted that the relative error of the ensemble is not accounted for by the rank histogram. While the ensemble with a lead time of 0–24 h is actually overconfident, this has little effect for practical application, since the forecasted ensemble runoff for all members is almost identical to HREF and shows small relative errors. Indeed the associated scatter diagram (not shown) for error and spread shows that values group towards the left center for the shorter lead times.

The rank histogram for HART with HREF (not shown) features two superimposed characteristics: on the one hand, we see the same overpopulation of low ranks as for HEPS, on the other hand the central ranks are overpopulated too. This indicates an additional overestimation in spread for certain classes of runoff occurrences. Reviewing the chained plots for HART (cf. Fig. 3), this can be traced back to constantly high spread for median runoff occurrences as already mentioned in section 3.1. While spread for HART in Fig. 3 seems sufficiently wide to cover the peaks, although significantly narrower than HEPS spread, the combined spread-skill evaluation for all catchments indicates an underestimation of HART spread for high runoff occurrences.

3.4 Verification of time series

For further performance evaluation, the ranked probability skill scores (RPSS, 1: perfect skill, 0: no skill) of the ensemble hindcasts against the reference simulation and observed runoff were calculated, as described in Wilks (2006), separately for different lead times. This allows the temporal evolution of the hydrograph to be considered.

As the ensemble size differs for the two years considered (2005 and 2006), the influence of the change in ensemble size is assessed by an inter-comparison of the two years. This inter-comparison is restricted to the evaluation against HREF in order to exclude the varying performance of the hydrological model itself (changing biases of HREF from OBS as stated in Sect. 3.2). Furthermore the debiased ranked probability skill score (RPSSd) is used in addition to the RPSS.

Table 2. Skill scores for different lead times. On the left, the combined median HEPS RPSS for all catchments for the full evaluation period relative to HREF, OBS and HART. On the right, the combined median Nash-Sutcliffe coefficient (E) for all catchments for the full evaluation period relative to HREF. For HEPS, the ensemble median was used to calculate E.

	RPSS			E	
lead time	HREF	OBS	HART	HDET	HEPS
0-24	0.969	0.601	0.175	0.996	0.997
24–48	0.902	0.607	0.223	0.898	0.966
48–72	0.870	0.598	0.188	0.770	0.910
72–96	0.829	0.582	0.143	*na*	0.904

Comparison of RPSSd for the years 2005 and 2006 reveals that other modifications introduced into the atmospheric model with the increase in ensemble size (mainly the increase in vertical resolution from 32 to 40 levels) do not result in a noticeable change in skill (e.g., with a lead time of 48–72 h: 0.850 and 0.853 for 2005 and 2006, respectively). To test the direct effect of the change in ensemble size, the RPSS was calculated without the debiasing separately for the two years considered. Again, resulting score differences are minor and cannot be clearly associated to the increase in ensemble size for the year 2006. Considering the minor score differences and the fact that a separate evaluation of the two years yields the same conclusions, we assume that it is valid to evaluate the two years jointly.

The debiasing of the RPSS is only used for the inter-annual comparison but not for the evaluation of the full time period, as in the latter case we are primarily interested in the actual skill of the model system and not the theoretically obtainable skill score with a perfectly calibrated ensemble (represented by the RPSSd, Weigel et al., 2007b). In Table 2 we show results for the forecast system over the period 2005–2006. The skill scores show differing results depending on the catchments, but in general, the RPSS is decreasing with increasing lead time (cf. Table 2). The decrease of RPSS is consistent with the results of the yearly analysis regarding HEPS range and HEPS IQR for different lead times and quantifies the additional uncertainty that is associated with longer lead times.

In contrast to, e.g., Olsson and Lindström (2007), where the use of the global EPS necessitates an additional bias correction, we find high score values using HREF as reference. This shows the suitability of the substitution of observed meteorological variables with forecasted ones and therefore the applicability of the coupled forecast system. Evaluation against observed runoff (OBS) results in RPSS values which clearly indicate improved skill of the forecast system over a climatological forecast. Nevertheless, the difference in skill between the evaluation against HREF and OBS leaves room for further improvements of the forecast system. While the uncertainty introduced with the use of meteorological forecasts is well covered by the ensemble approach, the score differences shown represent the bias of HREF against OBS, i.e., the model uncertainty of the hydrological model and the uncertainty in its initial state, which arises from uncertainties in the observed meteorological input (e.g., from interpolation uncertainties, cf. Ahrens and Jaun, 2007). Note that HREF evolves freely throughout the evaluation period and is not nudged against measured runoff. As HREF is used to generate the initial conditions for the forecast runs, a nudging against measured runoff would lead to a substantial increase in skill score values for HEPS relative to OBS.

While the median of an ensemble should not be applied for evaluation of single events, it was used for a deterministic evaluation of two continuous years (2005 and 2006). The comparison of the HEPS median to HDET by means of the deterministic skill score E is performed with regard to HREF. It reveals almost identical numbers for the shortest lead time (cf. Table 2). For longer lead times, the skill of the HEPS median decreases less rapidly than that of HDET. While the result of this evaluation should not be interpreted on its own (as the probabilistic information of the ensemble is lost), it gives a clear indication that the ensemble does not fall behind HDET in performance, even though the underlying meteorological model features a coarser numerical grid. This may not remain true for evaluations based on a shorter (hourly) timescale within the first 24 h.

The results of the evaluation of HEPS against its climatological correspondent HART are shown in Table 2. It reveals that HEPS performs better in terms of RPSS. For the longest lead time (72–96 h), HEPS shows a slight decrease in advantage over HART. This is consistent with the expectation that the relative performance of a climatological forecast should increase with longer lead times.

Taking into account the positive results of the probabilistic evaluation and the verification of the general ensemble performance, we can confidently state an added value of HEPS with regard to HDET: apart from better "deterministic results" for longer lead times, the ensemble is better than its own median forecast with climatological spread information and therefore shows the importance of temporal variability in the ensemble range and spread.

Weighting of the ensembles with the cluster sizes shows only marginal effects for all applied evaluation methods and can be neglected for the analysis of the probabilistic hydrological forecast series. This is consistent with the findings for precipitation verification by Marsigli et al. (2001), but it should be noted that weighting can improve the skill of the hydrological ensemble for specific cases and higher temporal resolution as showed in Jaun et al. (2008).

4 Conclusions

Using two years (2005 and 2006) of continuous daily hindcasts for the upper Rhine catchment, we find a good hindcast performance of the applied hydrometeorological forecast system. Statistical analysis shows that general ensemble requirements are reasonably met. The high RPSS values resulting from the evaluation against HREF demonstrate the applicability of the proposed coupled forecast system. It was shown that the chosen approach works for a wide band of weather conditions and that the ensemble spread represents the additional uncertainties during weather situations with low predictability. As expected, the IQR and the full spread of the ensemble increase systematically with lead time. Compared to the deterministic forecast HDET, the HEPS median shows higher skill for longer lead times. In addition, the evaluation of HEPS against its climatological correspondent HART shows the importance of temporal variability in the ensemble range and spread. As the ensemble forecast is better in both aspects, it is safe to assume that the use of the ensembles is superior to the deterministic alternative, especially with regard to the additional provision of probabilistic forecast information.

Although the evaluation against HREF reveals that the input uncertainty, introduced by the use of meteorological weather predictions, is well covered by the ensemble approach, the forecast system would probably profit from an additional ensemble calibration (Hamill et al., 2004). The reforecast series required for such a calibration recently became available for COSMO-LEPS (Fundel et al., 2009).

The evaluation against observed runoff shows further potential for improvements of the model system. Consequently, future work is planned to include the remaining uncertainties as adopted by, e.g., Pappenberger et al. (2005). Special attention will be payed to the initialization uncertainty of the hydrological component of the forecast system. Efforts towards an operational application of probabilistic forecasts, using similar setups as the described forecast system, are ongoing and were first demonstrated quasi-operationally within the framework of MAP D-PHASE (Zappa et al., 2008).

Acknowledgements. We would like to thank the Swiss Federal Office for the Environment (FOEN), the Landeswasserbauamt Voralberg, the German Weather Service (DWD) and MeteoSwiss for provision of meteorological and hydrological data. Financial support is acknowledged to the Swiss National Centre for Competence in Research (NCCR-Climate) and the State Secretariat for Education and Research SER (COST 731). The COSMO-7 weather forecasts were done at the Swiss National Supercomputing Centre (CSCS) in Manno, Switzerland. André Walser from MeteoSwiss kindly provided the used atmospheric model data. Data analysis was carried out by use of R (R Development Core Team, 2008). Christoph Schär is thanked for his support and scientific advice. Finally, we would like to acknowledge Joachim Gurtz and Massimiliano Zappa for support with hydrological questions, Tracy Ewen for proofreading and the anonymous reviewers for constructive and valuable comments.

Edited by: F. Pappenberger

References

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

Ahrens, B. and Walser, A.: Information-based skill scores for probabilistic forecasts, Mon. Weather Rev., 136, 352–363, doi:10.1175/2007MWR1931.1, 2008.

Anderson, J.: A method for producing and evaluating probabilistic forecasts from ensemble model integration, J. Climate, 9, 1518–1530, 1996.

Bartholmes, J. C. and Todini, E.: Coupling meteorological and hydrological models for flood forecasting, Hydrol. Earth Syst. Sci., 9, 333–346, 2005,
http://www.hydrol-earth-syst-sci.net/9/333/2005/.

Bartholmes, J. C., Thielen, J., Ramos, M. H., and Gentilini, S.: The european flood alert system EFAS – Part 2: Statistical skill assessment of probabilistic and deterministic operational forecasts, Hydrol. Earth Syst. Sci., 13, 141–153, 2009,
http://www.hydrol-earth-syst-sci.net/13/141/2009/.

Bergström, S. and Forsman, A.: Development of a conceptual deterministic rainfall-runoff model, Nord. Hydrol., 4, 147–170, 1973.

Bezzola, G. R. and Hegg, C. (Eds.): Ereignisanalyse Hochwasser 2005, Teil 1 – Prozesse, Schäden und erste Einordnung, no. 0707 in Umweltwissen, Bundesamt für Umwelt BAFU, Bern, Eidg. Forschungsanstalt WSL, Birmensdorf, online available at: http://www.bafu.admin.ch/publikationen/publikation/00044/, 2007.

Buizza, R.: Encyclopaedia of Atmospheric Sciences, chap. Weather Prediction: Ensemble Prediction, Academic Press, London, 2546–2557, 2003.

Buizza, R. and Palmer, T.: The singular-vector structure of the atmospheric global circulation, J. Atmos. Sci., 52, 1434–1456, doi:10.1175/1520-0469(1995)052⟨1434:TSVSOT⟩2.0.CO;2, 1995.

Cloke, H. and Pappenberger, F.: Ensemble flood forecasting: a review, J. Hydrol., doi:10.1016/j.jhydrol.2009.06.005, in press, 2009.

Diomede, T., Nerozzi, F., Paccagnella, T., and Todini, E.: The use of meteorological analogues to account for LAM QPF uncertainty, Hydrol. Earth Syst. Sci., 12, 141–157, 2008,
http://www.hydrol-earth-syst-sci.net/12/141/2008/.

Ehrendorfer, M.: Predicting the uncertainty of numerical weather forecasts: a review, Meteorol. Z., 6, 147–183, 1997.

Flügel, W.-A.: Combining GIS with regional hydrological modelling using hydrological response units (HRUs): An application from Germany, Math. Comput. Simulat., 43, 297–304, doi:10.1016/S0378-4754(97)00013-X, 1997.

Fundel, F., Liniger, M., Walser, A., Frei, C., and Appenzeller, C.: Reliable precipitation forecasts for a limited area ensemble forecast system using reforecasts, Mon. Weather Rev., in review, 2009.

Gurtz, J., Baltensweiler, A., and Lang, H.: Spatially distributed hydrotope-based modelling of evapotranspiration and runoff in mountainous basins, Hydrol. Process., 13, 2751–2768, doi:10.1002/(SICI)1099-1085(19991215)13:17⟨2751::AID-HYP897⟩3.0.CO;2-O, 1999.

Gurtz, J., Zappa, M., Jasper, K., Lang, H., Verbunt, M., Badoux, A., and Vitvar, T.: A comparative study in modelling runoff and its components in two mountainous catchments, Hydrol. Process., 17, 297–311, doi:10.1002/hyp.1125, 2003.

Hamill, T., Whitaker, J., and Wei, X.: Ensemble reforecasting: improving medium-range forecast skill using retrospective forecasts, Mon. Weather Rev., 132, 1434–1447, doi:10.1175/1520-0493(2004)132⟨1434:ERIMFS⟩2.0.CO;2, 2004.

Hohenegger, C., Walser, A., Langhans, W., and Schär, C.: Cloud-resolving ensemble simulations of the August 2005 Alpine flood, Q. J. Roy. Meteorol. Soc., 134, 889–904, doi:10.1002/qj.252, http://dx.doi.org/10.1002/qj.252, 2008.

Houtekamer, P. L., Lefaivre, L., Derome, J., Ritchie, H., and Mitchell, H. L.: A system simulation approach to ensemble prediction, Mon. Weather Rev., 124, 1225–1242, doi:10.1175/1520-0493(1996)124⟨1225:ASSATE⟩2.0.CO;2, 1996.

Jaun, S., Ahrens, B., Walser, A., Ewen, T., and Schär, C.: A probabilistic view on the August 2005 floods in the upper Rhine catchment, Nat. Hazards Earth Syst. Sci., 8, 281–291, 2008, http://www.nat-hazards-earth-syst-sci.net/8/281/2008/.

Komma, J., Reszler, C., Blöschl, G., and Haiden, T.: Ensemble prediction of floods – catchment non-linearity and forecast probabilities, Nat. Hazards Earth Syst. Sci., 7, 431–444, 2007, http://www.nat-hazards-earth-syst-sci.net/7/431/2007/.

Laio, F. and Tamea, S.: Verification tools for probabilistic forecasts of continuous hydrological variables, Hydrol. Earth Syst. Sci., 11, 1267–1277, 2007, http://www.hydrol-earth-syst-sci.net/11/1267/2007/.

Lalaurette, F., Bidlot, J., Ferranti, L., Ghelli, A., Grazzini, F., Leutbecher, M., Paulsen, J.-E., and Viterbo, P.: Verification statistics and evaluations of ECMWF forecasts in 2003–2004, Tech. Rep. 463, ECMWF, Shinfield Park Reading, Berks RG2 9AX, online available at: http://www.ecmwf.int/publications/library/ecpublications/_pdf/tm/401-500/tm463.pdf, 2005.

Legates, D. R. and McCabe, G. J.: Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation, Water Resour. Res., 35, 233–242, doi:10.1029/1998WR900018, 1999.

Lindström, G., Johansson, B., Persson, M., Gardelin, M., and Bergström, S.: Development and test of the distributed HBV-96 hydrological model, J. Hydrol., 201, 272–288, doi:10.1016/S0022-1694(97)00041-3, 1997.

Marsigli, C., Montani, A., Nerozzi, F., Paccagnella, T., Tibaldi, S., Molteni, F., and Buizza, R.: A Strategy for high-resolution ensemble prediction. II: Limited-area experiments in four Alpine flood events, Q. J. Roy. Meteorol. Soc., 127, 2095–2115, doi:10.1256/smsqj.57612, 2001.

Marsigli, C., Boccanera, F., Montani, A., and Paccagnella, T.: The COSMO-LEPS mesoscale ensemble system: validation of the methodology and verification, Nonlin. Processes Geophys., 12, 527–536, 2005, http://www.nonlin-processes-geophys.net/12/527/2005/.

Mass, C., Ovens, D., Westrick, K., and Colle, B.: Does Increasing Horizontal Resolution Produce More Skillful Forecasts?, B. Am. Meteorol. Soc., 83, 407–430, doi:10.1175%2F1520-0477%282002%29083%3C0407%3ADIHRPM%3E2.3.CO%3B2, 2002.

Molteni, F., Buizza, R., Palmer, T., and Petroliagis, T.: The ECMWF Ensemble Prediction System: Methodology and validation, Q. J. Roy. Meteorol. Soc., 122, 73–119, doi:10.1002/qj.49712252905, 1996.

Molteni, F., Buizza, R., Marsigli, C., Montani, A., Nerozzi, F., and Paccagnelli, T.: A Strategy for high-resolution ensemble prediction. I: Definition of representative members and global-model experiments., Q. J. Roy. Meteorol. Soc., 127, 2069–2094, doi:10.1256/smsqj.57611, 2001.

Montani, A., Capaldo, M., Cesari, D., Marsigli, C., Modigliani, U., Nerozzi, F., Paccagnella, T., Patruno, P., and Tibaldi, S.: Operational limited-area ensemble forecasts based on the Lokal Modell, ECMWF Newsletter, 98, 2–7, online available at: http://www.ecmwf.int/publications/newsletters/pdf/98.pdf, 2003.

Müller, W. A., Appenzeller, C., Doblas-Reyes, F. J., and Liniger, M. A.: A debiased ranked probability skill score to evaluate probabilistic ensemble forecasts with small ensemble sizes, J. Climate, 18, 1513–1523, doi:10.1175/JCLI3361.1, 2005.

Nash, J. E. and Sutcliffe, J. V.: River flow forecasting through conceptual models: Part 1 - A discussion of principles, J. Hydrol., 10, 282–290, doi:10.1016/0022-1694(70)90255-6, 1970.

Olsson, J. and Lindström, G.: Evaluation and calibration of operational hydrological ensemble forecasts in Sweden, J. Hydrol., 350, 14–24, doi:10.1016/j.jhydrol.2007.11.010, 2007.

Pappenberger, F. and Beven, K. J.: Ignorance is bliss: Or seven reasons not to use uncertainty analysis, Water Resour. Res., 42, 1–8, doi:10.1029/2005WR004820, 2006.

Pappenberger, F., Beven, K. J., Hunter, N. M., Bates, P. D., Gouweleeuw, B. T., Thielen, J., and de Roo, A. P. J.: Cascading model uncertainty from medium range weather forecasts (10 days) through a rainfall-runoff model to flood inundation predictions within the European Flood Forecasting System (EFFS), Hydrol. Earth Syst. Sci., 9, 381–393, 2005, http://www.hydrol-earth-syst-sci.net/9/381/2005/.

R Development Core Team: R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project.org, ISBN 3-900051-07-0, 2008.

Roulin, E. and Vannitsem, S.: Skill of Medium-Range Hydrological Ensemble Predictions, J. Hydrometeorol., 6, 729–744, doi:10.1175/JHM436.1, 2005.

Rousset, F., Habets, F., Martin, E., and Noilhan, J.: Ensemble streamflow forecasts over France, ECMWF Newsletter, 111, 21–27, http://www.ecmwf.int/publications/newsletters/pdf/111.pdf, 2007.

Scherrer, S. C., Appenzeller, C., Eckert, P., and Cattani, D.: Analysis of the spread-skill Relations using the ECMWF ensemble prediction system over Europe, Weather Forecast., 19, 522–565, 2004.

Siccardi, F., Boni, G., Ferraris, L., and Rudari, R.: A hydrometeorological approach for probabilistic flood forecast, J. Geophys. Res., 110, 1–9, doi:10.1029/2004JD005314, 2005.

Steppeler, J., Doms, G., Schättler, U., Bitzer, H.-W., Gassmann, A., Damrath, U., and Gregoric, G.: Meso-gamma scale forecasts using the nonhydrostatic model LM, Meteorol. Atmos. Phys., 82, 75–96, doi:10.1007/s00703-001-0592-9, 2003.

Todini, E.: Role and treatment of uncertainty in real-time flood forecasting, Hydrol. Process., 18, 2743–2746, doi:10.1002/hyp.5687, 2004.

Toth, Z. and Kalnay, E.: Ensemble forecasting at NCEP and the breeding method, Mon. Weather Rev., 125, 3297–3319, doi:10.

1175/1520-0493(1997)125⟨3297:EFANAT⟩2.0.CO;2, 1997.

Verbunt, M., Zappa, M., Gurtz, J., and Kaufmann, P.: Verification of a coupled hydrometeorological modelling approach for alpine tributaries in the Rhine basin, J. Hydrol., 324, 224–238, doi:10.1016/j.jhydrol.2005.09.036, 2006.

Verbunt, M., Walser, A., Gurtz, J., Montani, A., and Schär, C.: Probabilistic Flood Forecasting with a Limited-Area Ensemble Prediction System: Selected Case Studies, J. Hydrometeorol., 8, 897–909, doi:10.1175/JHM594.1, 2007.

Viviroli, D., Zappa, M., Gurtz, J., and Weingartner, R.: An introduction to the hydrological modelling system PREVAH and its pre- and post-processing-tools, Environ. Model. Softw., 24, 1209–1222, doi:10.1016/j.envsoft.2009.04.001, 2009.

Vrugt, J. A., Diks, C. G. H., Gupta, H. V., Bouten, W., and Verstraten, J. M.: Improved treatment of uncertainty in hydrologic modeling: Combining the strengths of global optimization and data assimilation, Water Resour. Res., 41, W01017, doi:10.1029/2004WR003059, 2005.

Weigel, A. P., Liniger, M. A., and Appenzeller, C.: The Discrete Brier and Ranked Probability Skill Scores, Mon. Weather Rev., 135, 118–124, doi:10.1175/MWR3280.1, 2007a.

Weigel, A. P., Liniger, M. A., and Appenzeller, C.: Generalization of the Discrete Brier and Ranked Probability Skill Scores for Weighted Multi-model Ensemble Forecasts, Mon. Weather Rev., 135, 2778–2785, doi:10.1175/MWR3428.1, 2007b.

Wilks, D.: Statistical methods in the atmospheric sciences, vol. 91 of International geophysics series, Elsevier, Amsterdam, 2nd edn., 2006.

Zappa, M., Pos, F., Strasser, U., Warmerdam, P., and Gurtz, J.: Seasonal water balance of an Alpine catchment as evaluated by different methods for spatially distributed snowmelt modelling, Nord. Hydrol., 34, 179–202, 2003.

Zappa, M., Rotach, M. W., Arpagaus, M., Dorninger, M., Hegg, C., Montani, A., Ranzi, R., Ament, F., Germann, U., Grossi, G., Jaun, S., Rossa, A., Vogt, S., Walser, A., Wehrhan, J., and Wunram, C.: MAP D-PHASE: Real-time demonstration of hydrological ensemble prediction systems, Atmos. Sci. Lett., 9, 80–87, doi:10.1002/asl.183, 2008.