Flood forecast errors and ensemble spread—A case study

T. Nester,[1] J. Komma,[1] A. Viglione,[1] and G. Blöschl[1]

Received 18 November 2011; revised 17 August 2012; accepted 23 August 2012; published 2 October 2012.

[1] Flood forecasts are generally associated with errors, which can be attributed to uncertainties in the meteorological forecasts and the hydrologic simulations, and ensemble spreads are usually considered capable of representing them. To quantify these two components of the total forecast errors and to compare these to ensemble spreads, an extended data set is used. Four years of operational flood forecasts at hourly time step with lead times up to 48 h are evaluated for 43 catchments in Austria and Germany. Catchment sizes range from 70 to 25,600 km^2, elevations from 200 to 3800 m, and mean annual precipitation from 700 to 2000 mm. A combination of ECMWF and ALADIN ensemble forecasts are used as input in a semidistributed conceptual water balance model on an hourly time step. The results indicate that, for short lead times, the ratio of hydrological simulation error to precipitation forecast error is 1.2 to 2.7 with increasing catchment size from 100 to 10,000 km^2. For long lead times the ratio of hydrological simulation error to precipitation forecast error decreases from 1.1 to 0.9 with increasing catchment size. Clear scaling relationships of the forecast error components with catchment area are found. A similar scaling is also found for ensemble spreads, which are shown to represent quantitatively the total forecast error when forecasting floods.

Citation: Nester, T., J. Komma, A. Viglione, and G. Blöschl (2012), Flood forecast errors and ensemble spread—A case study, *Water Resour. Res.*, *48*, W10502, doi:10.1029/2011WR011649.

1. Introduction

[2] One of the main challenges in flood forecasting and warning is to extend forecast lead times beyond the catchment response time as the forecasts will then rely on rainfall predictions [*Blöschl*, 2008]. This is of particular concern in small and medium sized catchments where the catchment response times are short and more time is required for flood response actions.

[3] As the forecast lead time increases the forecast errors tend to increase. For flood response actions it is therefore essential to get an indication of the magnitudes of the forecast errors to be expected at any point in time [*Montanari*, 2007]. There are two main sources of uncertainty that contribute to the flood forecast errors: precipitation forecast errors and hydrological simulation errors [*Krzysztofowicz*, 2001] with many independent sources contributing to both errors (see, e.g., Table 1). The precipitation forecast errors represent the differences between predicted and observed (and interpolated) precipitation. Because of the nonlinearity of the atmospheric system the errors tend to increase drastically with the forecast lead time. The common approach to quantifying the precipitation forecast uncertainty are ensemble simulations where a numerical weather prediction (NWP) model is run for a number of cases with slightly different initial conditions. The cases evolve along different trajectories which produce a range of precipitation forecasts [*Buizza*, 2003; *Grimit and Mass*, 2007]. The different cases or ensemble members of precipitation are used as inputs into a hydrological model to produce a range of flood forecasts [*Demeritt et al.*, 2007]. The spread of the ensemble members in terms of flood discharge is then used as a measure of forecast uncertainty due to uncertain precipitation forecasts [e.g., *Pappenberger et al.*, 2005]. These types of ensemble forecasts have been performed for short range (around 48 h) [e.g., *Komma et al.*, 2007, *Thirel et al.*, 2008], and medium-range (up to 15 days) forecasts [e.g., *Gouweleeuw et al.*, 2005; *Roulin and Vannitsem*, 2005; *Roulin*, 2007; *Verbunt et al.*, 2007; *Thielen et al.*, 2009; *Hopson and Webster*, 2010]. A review of ensemble flood forecasting systems and the additional value of using ensembles is given by *Cloke and Pappenberger* [2009]. Most of these studies assume that the precipitation forecast uncertainty is the main source of uncertainty impacting on the flood forecasts.

[4] Hydrological simulation errors have been the subject of numerous studies in hydrology [see, e.g., *Montanari and Brath*, 2004; *Montanari et al.*, 2009; *Weerts et al.*, 2011; *Coccia and Todini*, 2011]. The hydrological simulation errors represent the differences between predicted and observed runoff using observed (and interpolated) precipitation. The errors are usually classified into input errors, model parameter errors, and model structure errors. There are a range of methods of quantifying the first two types of errors including Monte Carlo simulations and analytical approaches

[1] Institute of Hydraulic Engineering and Water Resources Management, Vienna University of Technology, Vienna, Austria.

Corresponding author: T. Nester, Institute for Hydraulic and Water Resources Engineering, Vienna University of Technology, Austria, Karlsplatz 13/222, 1040 Vienna. (nester@hydro.tuwien.ac.at)

0043-1397/12/2011WR011649

Table 1. Contributions to the Hydrological Simulation Error and Precipitation Forecast Error

Hydrological Simulation Errors	Precipitation Forecast Errors
Parameters of runoff model	Parameters of atmospheric model
Structure of runoff model	Structure of atmospheric model
Precipitation measurements	Initial conditions
Precipitation interpolation	

[*Montanari et al.*, 2009]. A typical representative of a method of estimating the simulation error is *Montanari and Grossi* [2008] who infer the probability distribution of the error through a multiple regression with current forecasted discharges, past forecast error, and past rainfall. *Weerts et al.* [2011] proposed to use quantile regressions for the assessment of the relationship between the hydrological forecast and the associated forecast error. *Krzysztofowicz and Kelly* [2000] presented Bayesian theory and a meta-Gaussian model for estimating the hydrological simulation error. *Coccia and Todini* [2011] adapted the model conditional processor to be used with several joint truncated normal distributions to reduce the predictive uncertainty. A few studies have combined the precipitation forecast uncertainty and hydrological simulation uncertainty. *Krzysztofowicz* [2001], for example, presented an analytic numerical of combining the two uncertainties resulting in a probability of forecast river stages that is a mixture of two distributions related to occurrence and nonoccurrence of precipitation. However, most ensemble flood forecast systems focus on the precipitation forecast uncertainty alone.

[5] In practice, the ensemble spread of the flood forecasts is often interpreted as an index of forecast errors rather than as a quantitative estimate of the errors. However, it is also of interest to understand how well the ensemble spread matches the actual forecast errors. Ensemble forecasts of precipitation are generally used based on two assumptions: (1) the members of the ensemble are equally likely, which is a work hypothesis needed to perform the calculations, and (2) the ensemble spread captures the precipitation forecast uncertainty, which is a hypothesis that can be checked a posteriori. However, this is often not the case. For example, *Schaake et al.* [2004] analyzed precipitation ensemble forecasts of the US National Centers for Environmental Prediction (NCEP) over the period 1997–1999 over the US. They found that the ensembles were biased and the spread was insufficient to capture measured precipitation. They proposed different methods for preprocessing the precipitation forecasts in order to remove bias and adjust the ensemble spread. Similarly, *Scherrer et al.* [2004] analyzed ensemble predictions of European Centre for Medium-Range Weather Forecasts (ECMWF) against precipitation observed at a rain gauge in Switzerland and found significant bias which they corrected with a neural network method. *Buizza et al.* [2005] compared the ensemble spreads of the methodologies used at the European Centre for Medium-Range Weather Forecasts (ECMWF), the Meteorological Service of Canada (MSC), and the US National Centers for Environmental Prediction (NCEP) for a 3 month period in 2002. Again, for all systems, the spread of ensemble forecasts was insufficient to capture reality, suggesting that preprocessing of the precipitation ensemble estimates is needed. Similar conclusions were arrived at by *Hamill et al.* [2008] for the ECMWF and the Global Forecast System (GFS) analyzing a longer time period but postprocessing methods do not always improve the ensemble forecasts of precipitation [*Schmeits and Kok*, 2010].

[6] The biases and uncertainties of precipitation forecasts may amplify when cascaded through the hydrological system. *Komma et al.* [2007] showed that for their flood forecasts the variability of the precipitation ensemble was amplified for lead times longer than the response time of the catchment. They used a combination of ECMWF and ALADIN ensemble forecasts as input into a distributed hydrologic model in a 620 km^2 catchment in Austria. They showed that small errors in rainfall may translate into larger errors in runoff. As an example *Komma et al.* [2007] showed that an uncertainty range of 70% in terms of precipitation translated into an uncertainty range of 200% in terms of runoff for a lead time of 48 h. They related this to the nonlinearity of the catchment response, however, uncertainties such as precipitation measurements and runoff model parameters may contribute to the amplification of the uncertainty in terms of runoff. In the context of flood forecasts it is therefore important to assess the precipitation uncertainty in terms of the effect on runoff rather than in terms of comparing forecast precipitation against observed precipitation. *Johnell et al.* [2007] used ECMWF ensemble forecasts to estimate runoff ensemble forecasts using the HBV model for 45 catchments in Sweden with areas ranging from 6 to 6110 km^2 (mean catchment size 647 km^2). They defined the ensemble spread as the range between the upper and the lower quartile of the runoff ensemble forecasts and compared it to the mean absolute error of the median runoff ensemble forecast. They classified the forecasts into five classes representing "very small" to "very large" ensemble spread with each class containing 20% of the forecasts. The mean absolute error, defined as the absolute difference between forecasted and observed discharges averaged over a number of days and divided by the observed discharge, increased from 2% for the class of very small ensembles to 18% for the class of very large ensembles on forecast day 1, and from 10% for very small ensembles to 75% for very large ensembles on forecast day 9. Errors for the forecast days 5 to 7 were similar, indicating that the EPS forecast has its main strength in the second part of the forecast period. *Jaun and Ahrens* [2009] estimated the runoff ensembles using downscaled ECMWF ensemble forecasts as input into the PREVAH model for 23 Swiss catchments with areas ranging from 610 to 34,550 km^2 (mean 6000 km^2). They defined the ensemble spread as the half interquartile runoff ensemble range and compared it to the forecast error obtained by comparing the median runoff ensemble forecast error with observed discharge and a reference forecast obtained by using meteorological observations, respectively. They found a tendency toward underestimation in the forecast spread when evaluating against the observed discharge with positive forecast errors 1.5 to 5 times and negative forecast errors 1.5 to 100 times larger than the ensemble spread. When comparing against the reference forecast, the underestimation in the ensemble spread was smaller with factors around 1.5 to 2 for both positive and negative forecast errors. However, additional uncertainties during unstable weather situations were found to be captured by larger ensemble spreads during flood peaks.

[7] A comparison with observed discharge and the reference forecast also allows separating the contributions of the precipitation forecast errors and the hydrological simulation errors to the total forecast errors. Understanding the relative contributions can assist in the future development of ensemble flood forecasting systems. *Olsson and Lindström* [2008] compared ensemble runoff forecasts in 45 Swedish catchments using forecasted ECMWF precipitation with observed runoff and a reference runoff simulation using observed precipitation. They found that 26% of the runoff simulations were within the interquartile range of the ensemble spread when comparing the forecasts to the simulations. This means that the precipitation ensembles are too narrow as the figure should be 50%. However, only 14% of the observed runoff were within the interquartile range of the ensemble spread when comparing the forecasts to observed runoff highlighting the contribution of the simulation error not accounted for in the runoff ensembles. They concluded that the contributions of the precipitation forecasts and the hydrologic simulations to the total error were of similar magnitudes but, for the rising limbs (i.e., the targets of flood forecasting systems), the precipitation forecast errors dominated. A similar analysis of separating the contributions of the precipitation forecast errors and the hydrological simulation errors was performed by *Addor et al.* [2011] who analyzed ensemble runoff forecasts in a 336 km^2 catchment in Switzerland based on precipitation ensemble forecast of a regional climate model (COSMO-LEPS). They found around 14% of the observations were within the interquartile range of the ensemble spread when comparing forecasts to observed runoff, but close to 50% of the forecasts were within the interquartile range of the ensemble spread when comparing the forecasts to simulations. *Zappa et al.* [2011] superposed different sources of uncertainty in a flood forecasting system. They used COSMO-LEPS forecasts and the hydrological model PREVAH for a 186 km^2 catchment in Switzerland. The uncertainty from the meteorological forecasts was represented by the uncertainty of the COSMO-LEPS ensembles propagated through the hydrologic model. The hydrologic model uncertainty was taken into account using a Monte Carlo simulation in which seven parameters of the hydrologic model (relevant for surface runoff generation) were randomly changed. The average runoff ensemble spread for the seven events analyzed was 130 m^3 s^{-1} when PREVAH was coupled with LEPS. When additionally taking the hydrological uncertainty into account the ensemble spread obtained was around 150 m^3 s^{-1}, meaning that the total uncertainty increased about 15%.

[8] The two main objectives of this study are (1) to quantify the contributions of precipitation forecast errors and hydrological simulation errors to the total forecast error, particularly during flood events, and (2) to evaluate the capability of the runoff ensemble forecasts to represent the total runoff forecast uncertainty as a function of lead time. We use a conceptual semidistributed hydrological model [*Blöschl et al.*, 2008] coupled to meteorological inputs (deterministic and ensemble forecasts) based on a combination of ECMWF and ALADIN forecasts [*Haiden et al.*, 2010], which is used operationally in the study area. Around 4 years of forecasts and runoff data at hourly time scale for 43 catchments with areas ranging from 70 to 25,600 km^2 in Austria and Germany are analyzed. Compared to previous studies, this work is based on a more extended amount of data, being the analysis performed at hourly time scale contemporaneously on a considerable number of catchments. Such an extended database allows us to identify scaling properties (with catchment area and lead time) of the total forecast error, of its two components (precipitation forecast and hydrological simulation errors) and of the runoff ensemble forecasts.

2. Study Region, Data, and Meteorological Forecast Inputs

[9] In this study we evaluate the hydrologic forecasts of the flood forecasting system for the Austrian Danube that is currently in operational use and has been developed by the Technical University of Vienna in 2003–2005. The system consists of (1) a meteorological, (2) a hydrologic, and (3) a hydraulic model part. The meteorological forecasts include deterministic and ensemble forecasts of precipitation and deterministic forecasts of air temperature for a lead time of 48 h on an hourly time step; the output of the hydrologic model includes deterministic and ensemble runoff forecasts in the Danube tributaries which are used to run a hydraulic model to estimate runoff and water level for the Danube River.

[10] The region is hydrologically diverse covering large parts of Austria and parts of Bavaria (Figure 1). The west of the region is Alpine with elevations of up to 3800 m asl (above sea level), while the north and east consist of prealpine terrain and lowlands with elevations between 200 and 800 m asl. Mean annual precipitation is between 600 mm yr^{-1} in the east and almost 2000 mm yr^{-1} in the west. The Alpine catchments generally show much higher runoff depths with 1600 mm yr^{-1}, compared to around 100 mm yr^{-1} in the east. Runoff from 43 catchments with sizes ranging from 70 to 25,600 km^2 (median size around 400 km^2) in the study region are used for the evaluation. The small catchments are mostly nested catchments. Land use is mainly agricultural in the lowlands, forested in the medium elevation ranges, and alpine vegetation, rocks, and glaciers in the alpine catchments. For the calibration of the model, meteorological and hydrologic data from 2002 to 2009 were used. For details we refer to *Nester et al.* [2011]. In this paper the analyses of the forecasts are based on a data set consisting of four years of meteorological forecasts (2006–2009). Around 35,000 time steps are analyzed in each catchment. Results are presented for all catchments as well as, in more detail, for four catchments of different size (300 to 1390 km^2) in diverse hydrologic regions (wet and relatively dry). Catchment characteristics and model performance in terms of the Nash-Sutcliffe model efficiency (*nsme*) and the volume error (*VE*), which is used as a measure of bias, of all catchments for the calibration and validation periods (2003–2006 and 2007–2009, respectively) are summarized in Appendix A. Definitions of the performance measures are given in Appendix B.

[11] The meteorological data and forecasts were provided by the Central Institute for Meteorology and Geodynamics (ZAMG) in Vienna using the INCA (integrated nowcasting through comprehensive analysis) system which is discussed in detail by *Haiden et al.* [2011]. The system has been developed for use in mountainous terrain and can be used for the analysis and nowcasting of temperature,

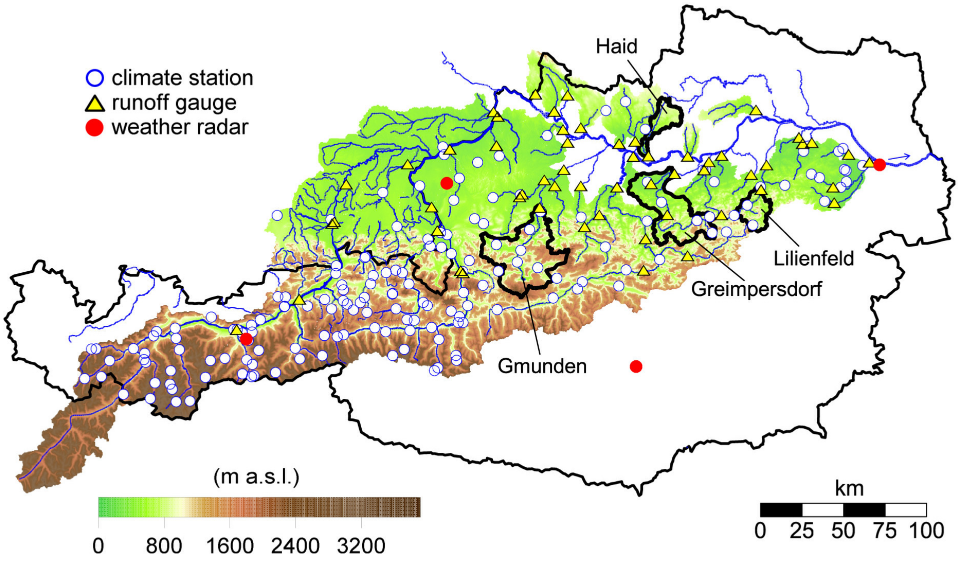

Figure 1. Topography of Austria and parts of southern Germany. The stream gauges used in the study are indicated by triangles, precipitation gauges by white circles, and weather radar stations by red circles. Thin black lines are catchment boundaries, and the thick black lines highlight the catchments Gmunden/Traun, Greimpersdorf/Ybbs, Haid/Naarn, and Lilienfeld/Traisen, used for detailed analyses.

precipitation amount, and wind fields, among others. The analysis part of the system combines surface station data with remote sensing data in a way that the observations at the station locations are reproduced, whereas the remote sensing data provide the spatial structure for the interpolation. The nowcasting part employs classical correlation-based motion vectors derived from previous consecutive analyses. In the case of precipitation the nowcast includes an intensity-dependent elevation effect.

2.1. Observed Precipitation Fields

[12] For each time step, rain gauge data were spatially interpolated on a 1 km grid and combined with radar data as a weighted mean. This approach attempts to combine the quantitative accuracy of rain gauge data with the spatial accuracy of radar data. The weights were derived from a comparison of monthly totals of radar and rain gauge data at the rain gauge locations. In areas where the visibility of the radar is low (e.g., mountainous catchments) the radar weights are small, while in areas with high radar visibility (e.g., lowland catchments) the weights are close to 1 [*Haiden et al.*, 2011]. Currently, 408 online available climate stations are implemented in INCA; 169 of which lie within the model region, which equals to one climate station every 258 km^2. On average, 0.4 stations per 100 km^2 are available in the study region. 70% of the stations are below 1000 m asl, 24% are between 1000 and 2000 m asl and the remaining 6% are above 2000 m asl with the highest station at 3100 m asl.

2.2. Precipitation Forecasts

[13] Deterministic precipitation forecasts are generated over a lead time of 48 h consisting of two components. The first component, termed nowcasts, is obtained by extrapolating the interpolated precipitation field using motion vectors [*Steinheimer and Haiden*, 2007]. The second component consists of the forecasts of the ALADIN and ECMWF numerical weather prediction (NWP) models. The two components are combined as a weighted mean. To allow for a smooth transition between nowcasts and NWP results, the weights are varied as a function of lead time from full weight to the nowcasts during the first 2 h, full weight to the NWP forecasts from 6 h, and a linear transition in between.

2.3. Ensemble Forecasts of Precipitation

[14] The ensemble forecasts of precipitation consist of three components. The first and second components are the nowcasts based on the motion vectors and the deterministic precipitation forecast of the ALADIN model from ZAMG, as for the precipitation forecast. On top of that, to account for small scale spatial uncertainty, the ALADIN forecast are spatially shifted in both the x and y directions to produce 25 pseudo-ensembles. The third component consists of 50 ensemble forecasts from the ECMWF model. The 50 ECMWF ensembles are randomly combined with one of the ALADIN pseudo-ensemble members and with the nowcasts. No uncertainty is assigned to the nowcasts, meaning that up to a lead time of 2 h all ensemble members are identical (zero spread) and the spread increases at longer lead times [*Komma et al.*, 2007]. A verification of INCA precipitation forecasts with a lead time of 12 h showed that in the first 6 h the precipitation amount is underestimated, whereas in the second half of the forecast period the precipitation amount is overestimated which can be attributed to the increasing influence of the NWP models for longer lead times [*Haiden et al.*, 2011].

[15] *Temperature forecasts* are based on a combination of interpolated station data and ALADIN forecasts. No temperature ensembles are generated as their effect on the flood forecasting uncertainty is deemed to be small. Analyses showed that a temperature increase of 1°C increases the forecast uncertainty for a lead time of 48 h on average on

the order of 1% in the months March–May when snow melt is contributing significantly to runoff. We used catchment mean values of precipitation as input into the hydrologic model; for the temperature data elevation was additionally accounted for when averaging over the catchments. All forecast are generated at an hourly time interval.

[16] For the analysis, hourly discharge data from 43 stream gauges were used. The data were checked for errors and in cases where a plausible correction could be made they were corrected. Otherwise they were marked as missing data.

3. Forecast Model Setup and Evaluation Methods

3.1. Forecast Model Setup

[17] The rainfall-runoff model used in this paper is a typical conceptual hydrologic model [*Blöschl et al.*, 2008; *Komma et al.*, 2008]. We used the model in a semidistributed configuration with each catchment divided into elevation zones. The model runs on an hourly time step and includes a snow routine, a soil moisture routine, and a flow routing routine [*Szolgay*, 2004]. Details about calibration and performance of the model are given by *Nester et al.* [2011, 2012]. Two real-time updating procedures are implemented to increase the accuracy of the forecasts. The first procedure is based on ensemble Kalman filtering and is used to assimilate runoff data to update the catchment soil moisture [*Komma et al.*, 2008]. The second procedure is an additive error model that exploits the autocorrelation of the forecast error and involves an exponential decay of the correction [*Komma et al.*, 2007]. The flood forecasting system has been in operational use for the Danube since 2008 and is operated by the state governments of Lower Austria and Upper Austria.

[18] In this study we emulate the real-time mode of flood forecasting in the Austrian Danube tributaries. This means that we use (1) the same updating procedures and (2) the same data sets including the associated uncertainties as the operational system. Errors induced (1) by imperfect water level observations [see, e.g., *Di Baldassarre and Montanari*, 2009], (2) by the estimation of runoff using rating curves, and (3) by data assimilation are neglected. For each forecasting time step t_0 the model is driven by observed precipitation and air temperature. Observed runoff for the same time step is used for the updating. The runoff forecasts are driven by (a) deterministic precipitation forecasts (and air temperature forecasts) and (b) 50 ensemble members of the ensemble precipitation forecasts. The latter give an estimation of the uncertainty distribution of the runoff forecasts over the lead time. In line with other forecasting systems, the uncertainty in the precipitation forecasts was assumed to be the only source of runoff forecast uncertainty, even though uncertainty due to the rainfall-runoff transformation (which we analyze in this study) and the uncertainty conveyed by the data assimilation (e.g., stage measurements and rating-curve errors) may not be negligible, especially during flood events.

[19] Examples of ensemble runoff forecasts for the catchment Greimpersdorf are given in Figure 2 for an event in June 2009. This event was the largest observed event in the study period with a return period of 30–35 years, which was due to local convective storms embedded in a large scale precipitation field. The forecast calculated on 22 June 2009 at 3 a.m. is shown in Figure 2(a). The runoff at the time of the forecast was around 30 m^3 s^{-1}. The cumulative ensemble forecasts of precipitation range between 60 and 100 mm for a lead time of 48 h, while the observed cumulative precipitation was 136 mm. The underestimation is due to heavy convective storms not captured by the precipitation forecasts. This leads to a significant underestimation of

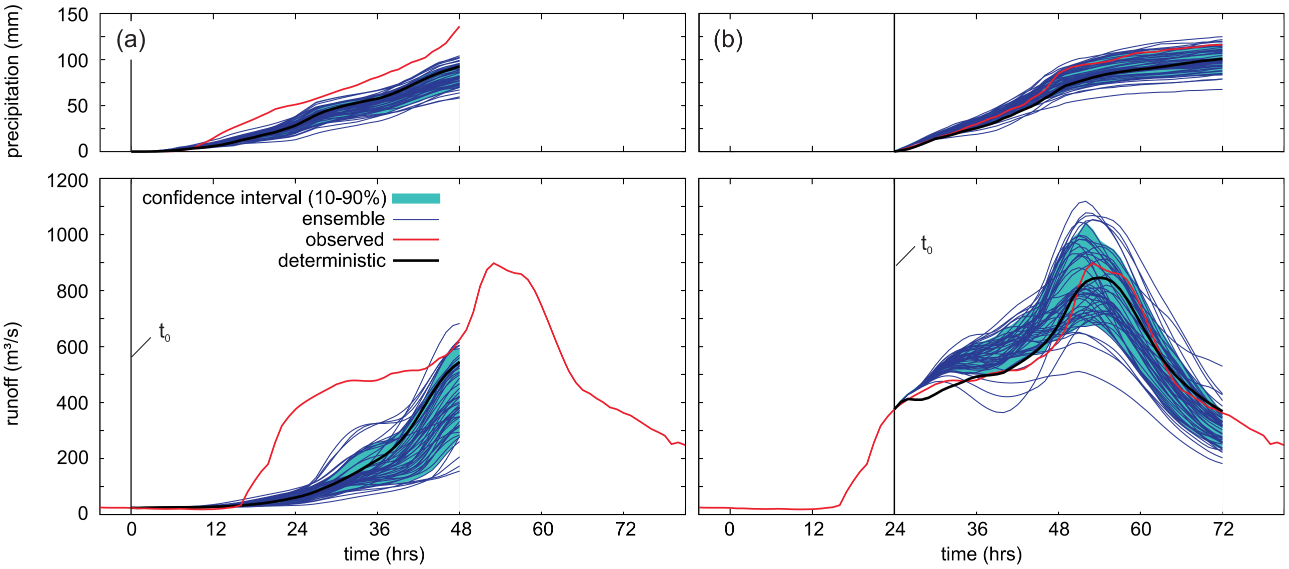

Figure 2. Example for ensemble forecasts at Greimpersdorf/Ybbs (1116 km^2). (a) Forecast time t_0 is 22 June 2009, 3 a.m.; and (b) t_0 is 23 June 2009, 3 a.m. Top panels: cumulative precipitation. Bottom panels: runoff. Red lines are the observations; black lines are the deterministic forecasts; thin blue lines are ensemble forecasts; and 80% confidence intervals are indicated in light blue.

runoff over much of the lead time and to missing the sudden rise in the hydrograph at time 15 h. However, at the end of the 48 h lead time, the runoff of the deterministic forecast run is almost at the level of the observed runoff, as are some ensemble members. Figure 2(b) shows the forecast run 24 h later. The cumulative ensemble forecasts of precipitation range between 84 and 121 mm with an observed precipitation of 116 mm. While the fine scale structure of the event is not fully captured, the overall shape of the hydrograph is captured very well. In this case, the ensemble runs give a very good indication of the forecast errors to be expected. The example illustrates that the performance of the flood forecasts will likely differ between events and can change within a single event. In some cases the ensembles will be representative of the errors but in others they will not.

3.2. Forecast Evaluation Methods

[20] For a first overview of the performance of the ensemble runoff forecasts we used two commonly used analysis methods. The first is the rank histogram or Talagrand diagram [*Talagrand et al.*, 1997, *Hamill*, 2001], where the rank of a verification (e.g., the observed runoff) is tallied relative to the values from an ensemble sorted from lowest to highest for *n* time steps. For an ensemble with 50 members, there are 50 + 1 bins between two ensemble members the observed runoff can fall into. Uniform frequencies for all ranks reflect equiprobability of the observations within the ensemble distribution [*Wilks*, 1995]. The second method is the Brier score (*BS*) [*Brier*, 1950]

$$BS = \frac{1}{N}\sum_{t^*=1}^{N}\left[p(t^*) - o(t^*)\right]^2, \tag{1}$$

where *p* is the forecast probability from the ensemble forecast for time t^* of exceeding a threshold discharge, *o* is a binary value depending on whether the observed discharge at the time step t^* exceeds the threshold discharge ($o = 1$) or does not ($o = 0$), and *N* is the total number of forecasts analyzed. *BS* ranges from 0 to 1. The Brier skill score (*BSS*) measures the improvement of a probabilistic forecast relative to a reference forecast BS_{ref}:

$$BSS = 1 - \frac{BS}{BS_{\text{ref}}}. \tag{2}$$

[21] The range of the *BSS* is $-\infty$ to 1, with the best score equal to 1. Positive scores indicate an improvement over the reference forecast. In order to calculate the *BS* (equation (1)) a threshold discharge needs to be chosen. We focused on high and median flows and selected the 50th and 90th percentile discharges from hourly data consistent with other studies [e.g., *Rousset-Regimbeau et al.*, 2007; *Thirel et al.*, 2008]. In order to calculate the *BSS* (equation (2)) a reference forecast needs to be estimated. We chose the climatological forecast as a reference forecast as proposed by *Stefanova and Krishnamurti* [2002].

[22] To compare our results with other studies, we focused on high and median flows by using the 50th and the 90th percentile derived from hourly observed runoff data from the years 2006–2009. For calculation of the *BSS* [*Stefanova and Krishnamurti*, 2002] the reference forecast BS_{ref} was estimated by using the observed runoff data from 2006 to 2009.

[23] For a more detailed evaluation of the ensemble runoff forecasts, we used a spread-skill analysis [e.g., *Scherrer et al.*, 2004; *Lalaurette et al.*, 2005]. Figure 3 gives the definitions of the terms used. t_0 denotes the time the forecast is made and t^* refers to the time of the predicted runoff. We examined forecasts lead times ($t^* - t_0$) of 1, 3, 6, 12, 24, and 48 h. As a measure of the ensemble spread we used two different measures: (1) the standard deviation $\hat{\sigma}_\varepsilon(t^*)$, and (2) the range between the upper quartile and the lower quartile $\widehat{IQR}_\varepsilon\ (t^*)$ of the runoff of the ensemble members for each point in time t^*. As a measure of skill we used both the standard deviation σ_ε and the IQR_ε of the forecast error $\varepsilon(t^*)$ which is defined as the difference between the observed runoff and the deterministic forecast.

[24] In order to distinguish between different forecast situations, we stratify the analysis of forecast errors in classes

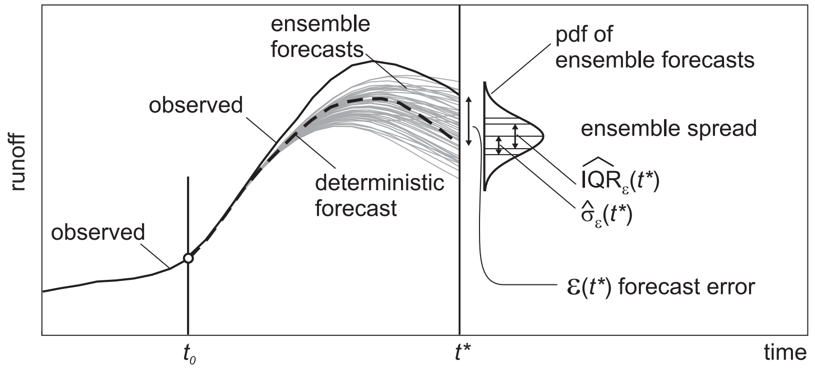
ensemble forecasts
observed
deterministic forecast
observed
runoff
pdf of ensemble forecasts
ensemble spread
$\widehat{IQR}_\varepsilon(t^*)$
$\hat{\sigma}_\varepsilon(t^*)$
$\varepsilon(t^*)$ forecast error
t_0
t^*
time

Figure 3. Definitions of the time t_0 the forecast is made and the time t^* for which the forecast is made. Two measures are used for the ensemble spread at time step t^*: (1) $\hat{\sigma}_\varepsilon(t^*)$, and (2) $\widehat{IQR}_\varepsilon\ (t^*)$. The forecast error is denoted as $\varepsilon(t^*)$.

of runoff ensemble spreads. Even though classes of precipitation ensemble spreads could have been used instead, the runoff ensemble spreads allow for taking into account not only precipitation induced but also snow melt induced flood events. Small ensemble spreads are more likely to occur when the runoff is constant or receding, while large ensemble spreads are more likely to occur when runoff is rising. Since floods are of interest here, the analysis of forecast errors associated to large ensemble spreads will be emphasized in the following. We grouped the forecast time steps t^* (for each lead time separately) into 10 classes according to the ensemble spread of that time step. For example, class 1 represents 10% of the time steps with the smallest ensemble spread and class 10 represents 10% of the time steps with the largest ensemble spread. Each class had the same number of time steps of $n = 3500$. For each class j we estimated the mean ensemble spread of the class j as

$$\bar{\hat{\sigma}}_{\varepsilon,j} = \frac{1}{2}(\max \hat{\sigma}_{\varepsilon,j} + \min \hat{\sigma}_{\varepsilon,j}) \tag{3}$$

and

$$\overline{\widehat{IQR}}_{\varepsilon,j} = \frac{1}{2}(\max \widehat{IQR}_{\varepsilon,j} + \min \widehat{IQR}_{\varepsilon,j}). \tag{4}$$

[25] Similarly, we calculated the standard deviation σ_ε of the forecast error $\varepsilon(t^*)$ over the same time steps:

$$\sigma_{\varepsilon,j} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}[\varepsilon(t^*) - \bar{\varepsilon}]^2}, \tag{5}$$

with $\bar{\varepsilon}$ as the mean forecast error over the n time steps. We also calculated the interquartile range $IQR_{\varepsilon,j}$ of the forecast error $\varepsilon(t^*)$ for each class j. If the ensemble forecasts fully portray the forecast errors, i.e., the ensembles are perfect, the ensemble spread and the forecast errors are equal for all classes j. It is worth noting that, differently from the *BSS* measure, the forecast errors and the ensemble spread do not account for biases of the forecast estimation. By comparing forecast errors and ensemble spread we aim to assess if the runoff ensemble forecasts and the total runoff forecast errors have the same spread and therefore the first can be deemed representative of the second.

[26] The other main objective of the paper is to quantify the contributions of precipitation forecast errors and hydrological simulation errors to the total forecast error. We analyzed the errors for two cases of runoff forecasts.

[27] 1. In the first case we used forecasts of deterministic precipitation as an input to the runoff model.

[28] 2. In the second case we used observed (and interpolated) precipitation data.

[29] These two cases allow us to examine the precipitation forecast errors separately from the hydrological simulation errors. Because of the nonlinearity of the runoff processes we consider it more appropriate to test the precipitation forecast errors via their effect on runoff rather than directly by comparing them against rain gauge data. This also allows us to compare the precipitation forecast errors directly with the hydrological simulation errors. This has been done looking in particular to the largest of ensemble spread classes since floods are of interest here.

4. Results

4.1. Forecast Performance Evaluated by Means of the Brier Skill Score

[30] An overall view of the performance of the ensemble runoff forecasts, irrespective of discharge and ensemble spread, is given in this section. Figure 4 shows a Talagrand diagram for a forecast lead time of 48 h. Q5, Q25, Q50, Q75, and Q95 are given for the 43 catchments analyzed in this study. If the equiprobability of the observations within the ensemble distribution is given, the ratio of frequency to expected frequency of the observed runoff falling in one of the ranks of the ensemble members should be unity. The diagram, however, is U-shaped for all lead times (lead times

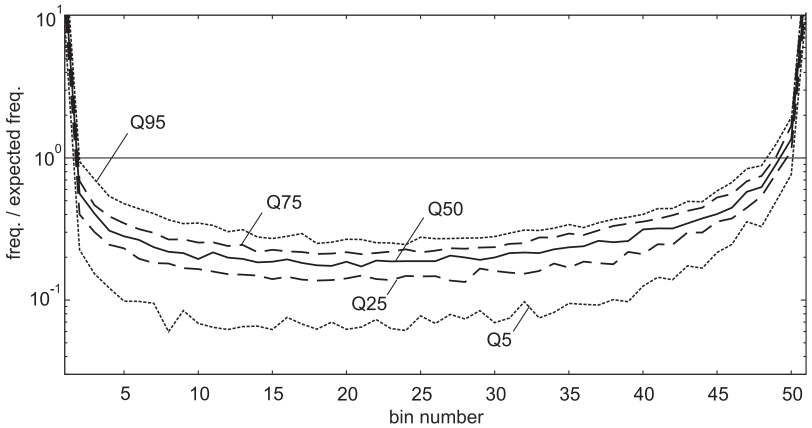

Figure 4. Talagrand diagram for runoff ensemble forecasts for a lead time of 48 h for all 43 catchments. Median values are shown as continuous line, 25% and 75% quantiles as dashed lines, and 5% and 95% quantiles as dotted lines. The number of the bin which is between two ensemble members is plotted against the ratio of frequency to expected frequency of the observations falling into each bin for n time steps.

from 1 to 24 h are not shown in Figure 4). This means that the observed runoff is too often lower or higher than the lowest or highest ensemble member; the ensemble spread is too narrow. As the lead time increases the U-shape becomes less distinct which means that the ensemble spread better captures the uncertainty in the forecasts. However, there are differences between the catchments which is shown in Figure 4 with the different quartiles. For a lead time of 48 h, the ratio of frequency to expected frequency in the middle bins of ranked ensembles members is between 0.15 and 0.22 for 50% of the catchments (Q75–Q25). In some catchments the ratio of frequency to expected frequency is below 0.1 for the middle bins. The mean value of observed runoff outside the entire range covered by the 50 ensemble members is 70%, while the figure for a perfect ensemble is 2%. It is clear that on average over all time steps the ensemble forecasts do not capture the entire range of forecast errors.

[31] Figure 5 shows the Brier skill score *BSS* for all catchments analyzed as a function of lead time. Figure 5(a) gives the *BSS* for a 50% percentile of runoff as reference forecast, i.e., it relates to medium and low runoff. Figure 5(b) gives the *BSS* for a 90% percentile, i.e., it relates to high flows. For both percentiles the *BSS* decreases with increasing lead time. Clearly as the lead time increases, the skill of the ensemble forecasts to match the forecast errors decreases. Overall, the skill for medium runoff (50% percentile) is higher than for high runoff (90% percentile). For most of the catchments the 50% *BSS* ranges between 0.8 and 0.1 at a lead time of 48 h, while the 90% *BSS* ranges between 0.6 and 0.0. For high flows it is more difficult for the ensemble forecasts to match the forecast errors than for medium flows.

[32] There are a small number of catchments (e.g., Greimpersdorf indicated by the black dashed line) where the *BSS* rapidly decreases after the first hour and increases after 12 h (Figure 5(a)). This is because low and medium flows in these catchments are influenced by regulations of hydropower plants which have a half-daily cycle not represented in the hydrologic model. For high flows the influence of the regulations is less apparent (Figure 5(b)). While it would be easily possible to model these cycles provided the information is available from the hydropower operators it is not relevant for the flood forecast which are the purpose of this study.

4.2. Ensemble Spread Versus Total Forecast Error

[33] As discussed in section 3.2, the spread-skill analysis provides more detailed insight into the performance of the ensemble forecasts and the contributions of the various error sources. In Figure 6 the midpoint of each class of ensemble spread $\overline{\hat{\sigma}}_\varepsilon$ and the corresponding spread of the forecast error σ_ε are shown as a function of lead time for the four example catchments highlighted in Figure 1. For ease of comparison, both the ensemble spread and forecast errors were scaled by the mean annual runoff of each catchment. All time steps for every lead time were assigned into one of 10 classes of equal size according to the ensemble spread $\overline{\hat{\sigma}}_\varepsilon$. Class 1 represents 10% of the time steps with the smallest ensemble spread and class 10 represents 10% of the time steps with the largest ensemble spread. Then we calculated the midpoint of the ensemble spreads for each class and the standard deviation of the forecast errors for each class according to equations (3) and (5). Figure 6 shows the results for the lead times 12 h (light gray squares), 24 h (dark gray triangles), and 48 h (black circles) hours. If the ensemble standard deviation and the forecast error standard deviation match, the points are close to the 1 : 1 line. On average, the standard deviation of the forecast error is 2–3 times bigger than the ensemble standard deviation. For short lead times, in particular the ensemble spread is small. This is because the ensemble spread is related to the precipitation forecasts, while for lead times shorter than the catchment response time the runoff mainly depends on observed precipitation [*Komma et al.*, 2007]. For lead times of 48 h the errors are twice and up to five times larger than the ensembles. Several reasons contribute to this. First, parts of the errors can be explained by the fact that we used real-time data and we only corrected obvious errors, whereas runoff

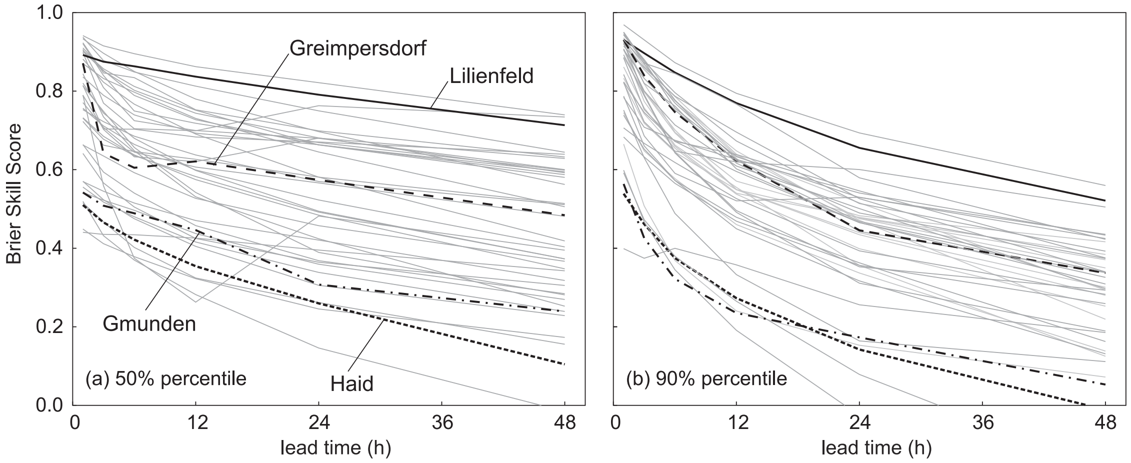

Figure 5. Brier skill score *BSS* as a function of lead time, computed using observed river flow as reference for a runoff exceeding the (a) 50% percentile and (b) 90% percentile. The focus in the left figure is on predicting low and medium runoff, the focus in the right is on predicting high runoff. Thick lines refer to the catchments used for detailed analyses.

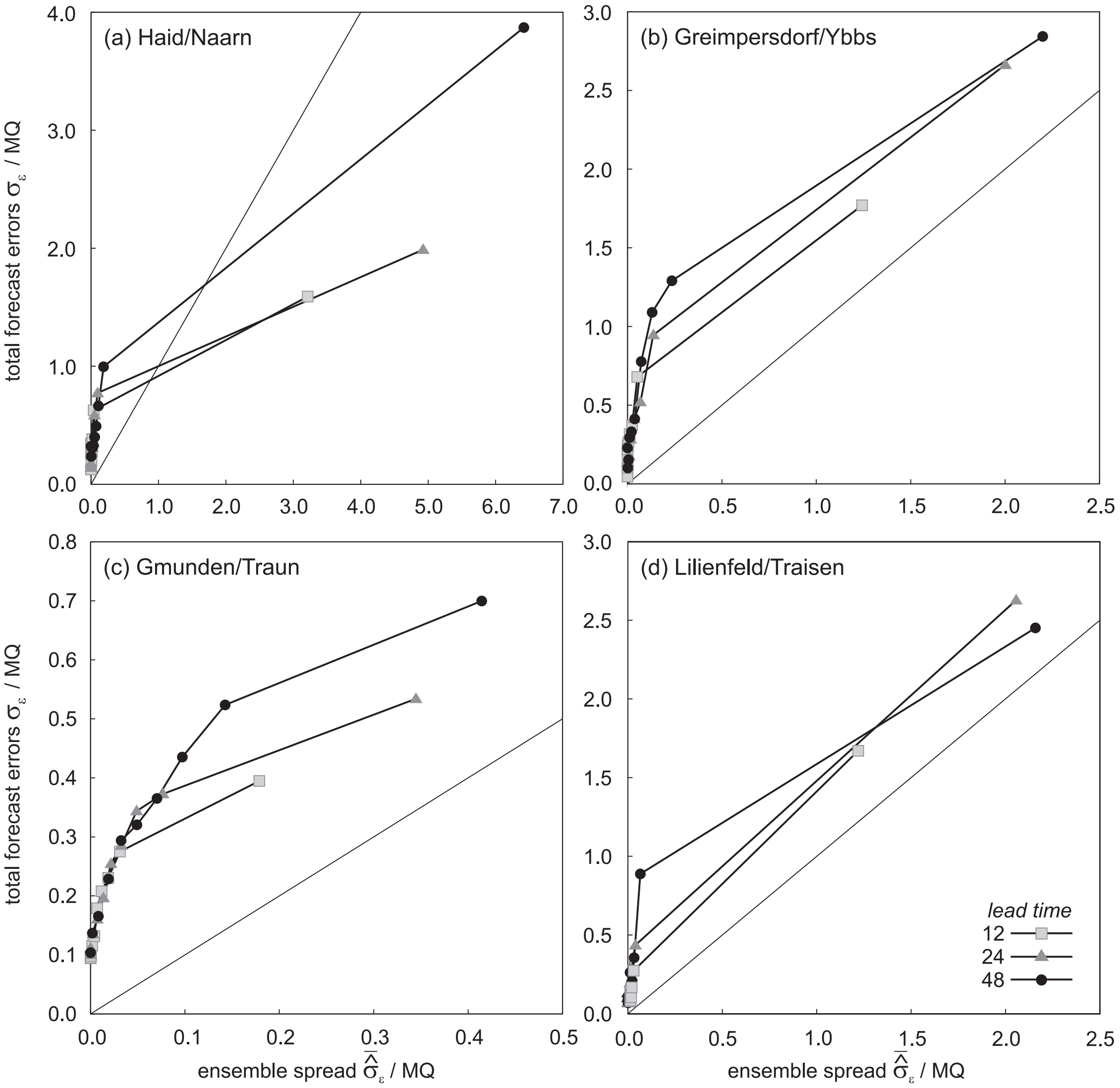

Figure 6. Ensemble spread $\hat{\overline{\sigma}}_\varepsilon$ versus standard deviation of total forecast errors σ_ε, both scaled by mean runoff (*MQ*). Ensemble spread is plotted at the midpoint of the classes. Light gray squares indicate a lead time of 12 h, dark gray triangles 24 h, and black circles 48 h. The thin line is the 1 : 1 line.

variations on the order of a few m^3 s^{-1} were not corrected. Second, short events with fast discharge increase in the catchments Haid (Figure 6(a)) and Lilienfeld (Figure 6(d)) are underestimated by the forecasts. Third, the runoff at the gauge Gmunden (Figure 6(c)) is influenced by the retention effects of a lake in the catchments. Interestingly the shapes of the forecast errors and the ensembles are remarkably similar. In Haid both the forecast errors and the ensemble spread increases significantly for lead times beyond 24 h. In Lilienfeld both remain constant. The similarity of the shapes suggests that the ensembles do capture the important characteristics of the errors.

[34] In the catchment Haid (Figure 6(a)), the ensemble spread does not capture the forecast errors in the first nine classes. The ensemble spread is around 10 times smaller than the forecast errors, whereas in the class with the largest ensemble forecasts the spread is larger than the forecast error by a factor of 2.0–2.5. The forecast errors almost double in this class when increasing the lead time from 24 to 48 h. The catchments Greimpersdorf (Figure 6(b)) and Gmunden show a similar behavior. The ensemble spread in the first 9 classes is up to 5 times smaller than the forecast errors. With increasing ensemble spread this factor decreases and the values are much closer to the 1 : l-line, indicating that the ensemble spread is almost able to capture the forecast errors when the spread is large.

[35] Figure 7 shows the spread-skill analysis using the IQR as a measure for the ensemble spread and the forecast errors. Again, if the ensemble spread and the forecast errors match, the points are on the 1 : 1 line. The ensemble spread using the IQR as measure is about 2 times larger than the ensemble spread using the standard deviation in Figure 6. The forecast errors using the IQR are about half of the forecast errors using the standard deviation as measure. However, the factors between ensemble spread and forecast error are of similar magnitude as in Figure 6. It is important to note that the classes of ensemble spread used in Figure 7 do not necessarily include the same time steps as in Figure 6.

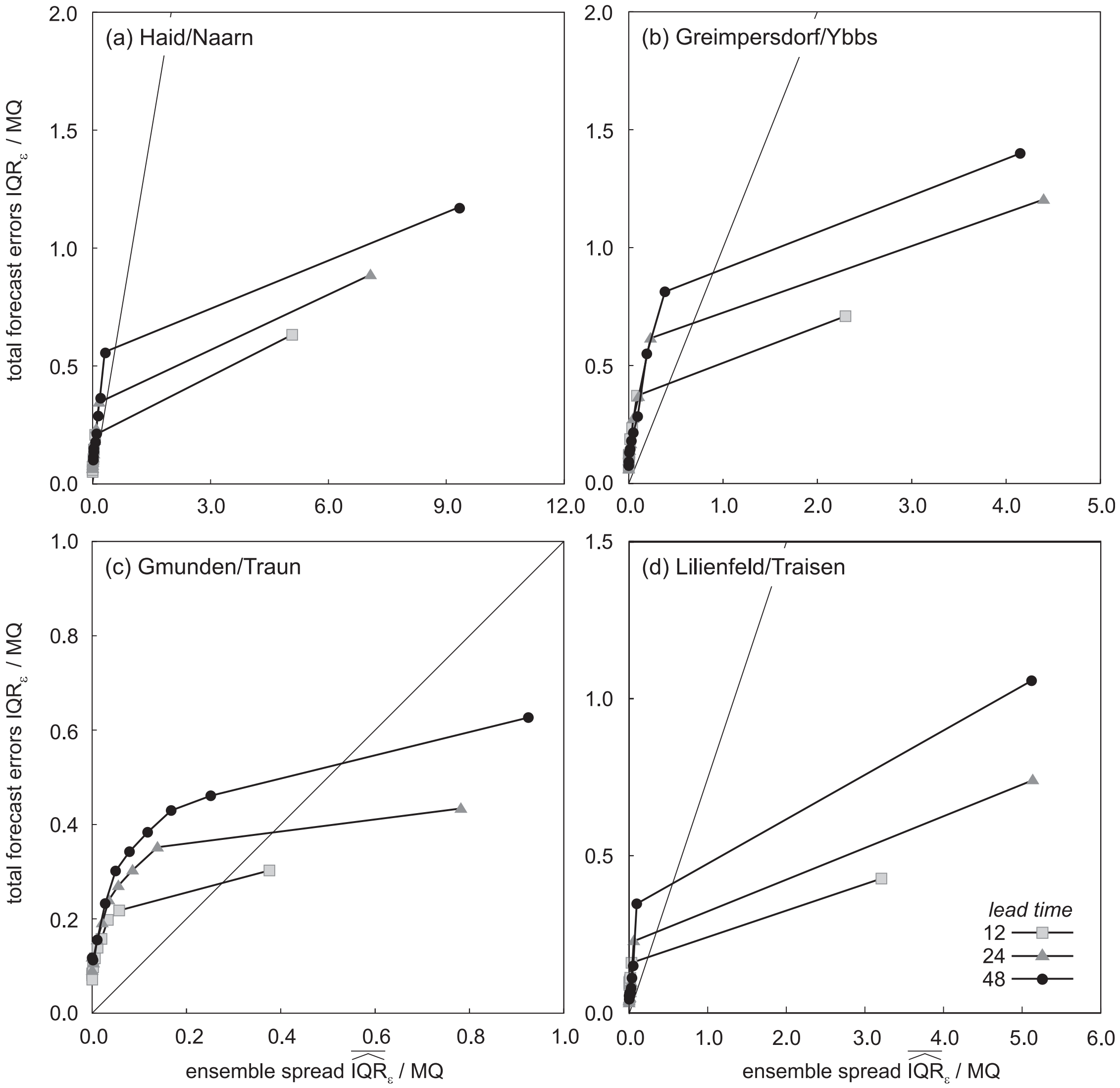

Figure 7. Same as Figure 6, but with the interquartile range as a measure for the ensemble spread $\widehat{IQR}_\varepsilon$ and the total forecast errors IQR_ε.

Despite the different values of ensemble spread and forecast errors the figures look similar, but there are some small differences between the shapes in Figures 6 and 7. For example, for the class of the largest 10% of ensembles at the gauge Greimpersdorf (Figure 7(b)) the ensemble spread for a lead time of 24 h is larger than then ensemble spread for 48 h using the IQR which can be attributed to the fact that the classes do not include the same time steps.

[36] Figure 8(a) shows the CDFs of the Spearman's rank correlation coefficient r_s between the ensemble standard deviation $\overline{\hat{\sigma}}_\varepsilon$ and the standard deviation of the total forecast error σ_ε for all 43 catchments analyzed for different lead times. It shows that for a lead time of 12 h 75% of the Spearman's rank correlation coefficients r_s are larger than 0.88, for a lead time of 24 h 75% of the values are larger than 0.95, and for a lead time 48 h 75% of the values are larger than 0.92. Similar results are obtained for the Spearman's rank correlation coefficient r_s between the ensemble interquartile range $\overline{\widehat{IQR}}_\varepsilon$ and the interquartile range of the total forecast error IQR_ε (Figure 8(b)) For a lead time of 12 h 75% of the Spearman's rank correlation coefficients r_s are larger than 0.95, for a lead time of 24 h 75% of the values are larger than 0.97, and for a lead time 48 h 75% of the values are larger than 0.98. This indicates that although the ensemble spreads are too narrow to capture the total forecast errors they still are a good indicator of the forecast error.

[37] The results of the spread-skill analyses using the standard deviation and the IQR as measure for the ensemble spread are consistent in terms of the shape of the figures. For the first nine classes, the forecast errors are up to 10 times larger than the ensemble spread. For the 10% of the largest ensemble spreads, this factor is lower. Analyses showed that $\overline{\widehat{IQR}}_\varepsilon$ is about 2 times larger than $\overline{\hat{\sigma}}_\varepsilon$, whereas IQR_ε is about 2 times smaller than σ_ε. As $\overline{\widehat{IQR}}_\varepsilon$ and $\overline{\hat{\sigma}}_\varepsilon$ are correlated, we will focus on $\overline{\hat{\sigma}}_\varepsilon$ as a measure of the ensemble spread for the remainder of this work.

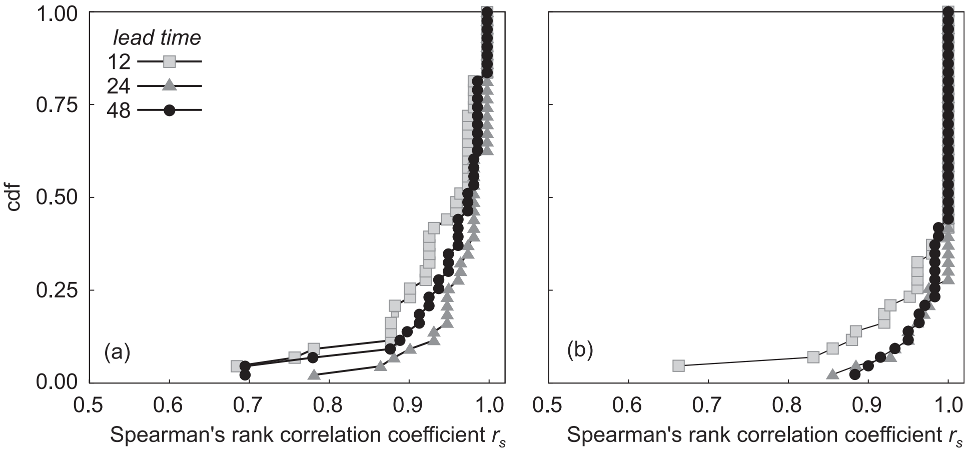

Figure 8. CDFs of the Spearman rank correlation coefficient r_s between ensemble spread and total forecast error for all 43 catchments analyzed. (a) Ensemble spread $\overline{\hat{\sigma}}_\varepsilon$ and total forecast error σ_ε, and (b) ensemble spread $\overline{\widehat{IQR}}_\varepsilon$ and total forecast error IQR_ε. Light gray squares indicate a lead time of 12 h, dark gray triangles 24 h, and black circles 48 h. Each point indicates the Spearman rank correlation coefficient r_s of a single catchment.

4.3. Contributions to the Forecast Error

[38] Figure 6 shows the total forecast errors where no distinction between the individual error sources is made. It is now of interest to examine the contributions of the precipitation forecasts and the hydrologic simulations to the total forecast errors. We analyzed the errors for two cases of runoff forecasts: (a) In the first case we used forecasts of deterministic precipitation as input in the runoff model as in Figures 6 to 8. (b) In the second case we used observed (and interpolated) precipitation data. In the second case, the precipitation forecast error is absent and the entire error is what we term hydrological simulation error. This error is due to precipitation measurement and interpolation, water level measurement and estimation of discharge and the structure and the parameters of the runoff model (Table 1). The first case also includes error components from the parameters and the structure of the atmospheric model and the initial conditions.

[39] For each time step we calculated the differences between observed runoff and the two cases of runoff forecasts. Using the same 10 classes of ensemble spread, we calculated the standard deviation of the hydrological simulation error σ_{hysim}. Assuming that the precipitation forecast errors and the hydrological simulation errors are independent, the variances are additive and the precipitation forecast error standard deviation σ_{pfor} can be calculated as

$$\sigma_{\text{pfor}} = \sqrt{{\sigma_\varepsilon}^2 - {\sigma_{\text{hysim}}}^2}. \tag{6}$$

[40] Note that equation (6) can be easily generalized to incorporate dependence between the errors. For instance, *Di Baldassarre and Montanari* [2009] assumed perfect positive correlation between the errors when dealing with rating curves uncertainty in river flow data. The individual process contributions to the error (Table 1) do not suggest any direct dependencies, as the hydrological simulation errors are mainly related to the hydrological model while the precipitation forecast errors are mainly related to the atmospheric model which is parameterized independently. A detailed error analysis was, however, considered beyond of the scope of this paper.

[41] We do not directly compare precipitation forecasts with precipitation measurements. The main reason is that the time scales relevant for the comparison depend on the catchment response. For example, in a catchment with a fast response, one would have to compare, say, precipitation forecast with a 6 h aggregation level, while for slowly responding catchments the aggregation level would have to be 24 h or more. Also, the nonlinearity of the runoff processes does not allow a direct comparison of rainfall errors (unit mm) with runoff errors (unit m^3 s^{-1}). We therefore consider it more appropriate to back-calculate the contribution of the precipitation forecasts from equation (6).

[42] Figure 9 shows the contributions to the total forecast errors for the four catchments. For clarity, only the results for the lead time of 48 h are shown. For small ensemble spreads the entire error is made up of hydrological simulation error. For larger ensemble spreads, the contribution of the precipitation forecast error increases and for the largest ensemble class, it is larger than the hydrological simulation error. This is the class of major interest because the large ensemble spreads are assumed to typically occur during the rising limbs of events, which could reveal to be the flood events that we want to forecast, while the small ensemble spreads typically occur during recessions or constant runoff periods. There are some apparent differences between the catchments. In the smaller catchments (Haid and Lilienfeld) the hydrological simulation error is larger than the precipitation forecast errors for all but the largest ensemble class. For Haid the precipitation forecast errors for of the largest ensemble class is about twice the hydrological simulation error, while for Lilienfeld they are similar. This is because Haid is drier than Lilienfeld (380 mm mean annual

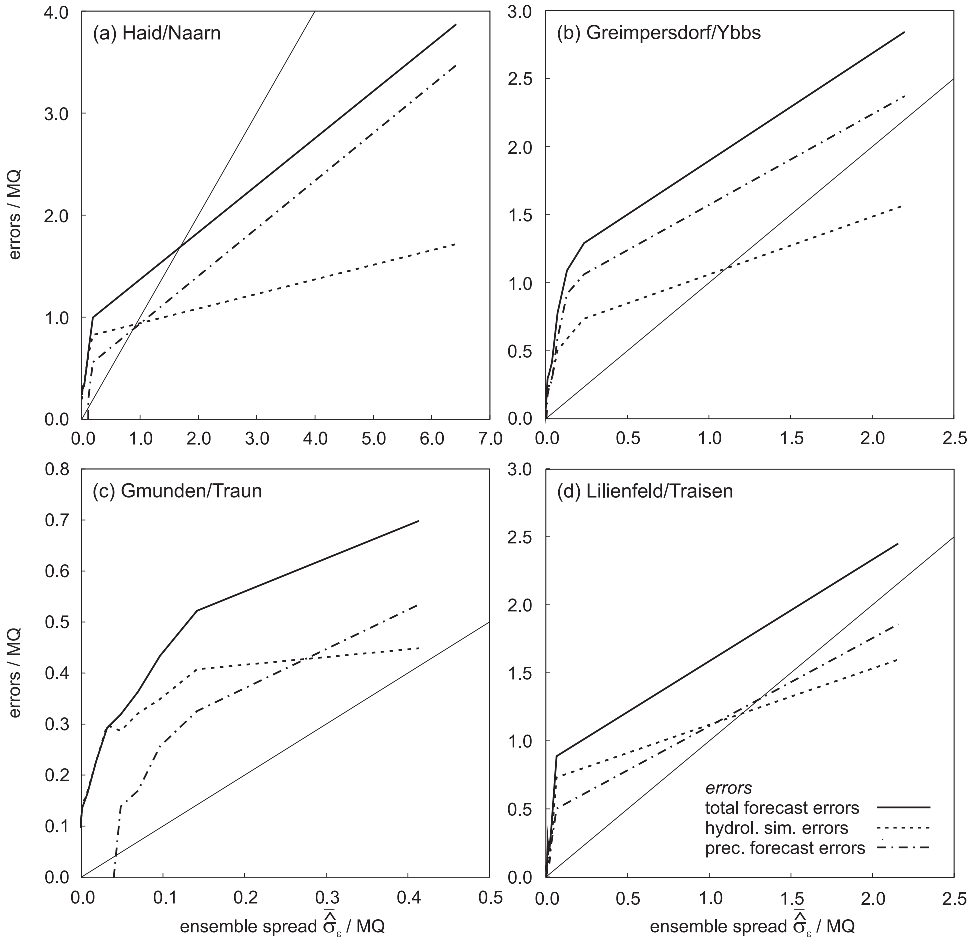

Figure 9. Components of the forecast errors for a lead time of 48 h for four catchments. Total forecast errors σ_ε are indicated as solid lines (and are identical with black circles in Figure 5), hydrological simulation errors σ_{hysim} are indicated as dashed lines, and precipitation forecast errors σ_{pfor} as dashed-dotted lines. Ensemble spread is plotted at the midpoint of the classes. The thin line is the 1:1 line.

runoff as opposed to 860 mm in Lilienfeld), so one would expect larger hydrological simulation errors [*Nester et al.*, 2011]. The errors in Greimpersdorf are similar to the errors in Lilienfeld. Both catchments have similar mean annual runoff depths and mean annual precipitation, however, Greimpersdorf is 3.5 times the size of Lilienfeld. The precipitation forecast errors are somewhat larger in Greimpersdorf, which can be attributed to the smaller number of precipitation stations per 100 km^2 in the catchments (1.2 in Lilienfeld and 0.4 in Greimpersdorf). This means that the spatial pattern of, e.g., small scale summer storms can be captured better in the catchment with higher precipitation density. The runoff in Gmunden is influenced by the retention effects of a lake, which is not explicitly modeled. However, the lake retention affects mainly low runoff situations, and flood peaks are not affected. This explains that for the small classes of ensemble spread the precipitation forecast error is zero, and the hydrological simulation error makes up 100% of the total forecast error.

[43] To assess the error contributions for all catchments, the forecast errors scaled with mean catchment runoff have been plotted against catchment size in Figure 10 for different lead times. As the main interest in this study is on the forecasting of floods the values of the top ensemble spread (largest 10%) are shown. Light gray squares indicate a lead time of 12 h, dark gray triangles represent a lead time of 24 h, and black circles stand for a lead time of 48 h. A linear regression was fitted to the errors of the individual catchments in the logarithmic domain:

$$\frac{\sigma_\varepsilon}{MQ} = \alpha \cdot A^\beta + \varsigma, \tag{7}$$

where α (dimension km$^{-2\beta}$) and β (dimensionless) are coefficients, A is the catchment area (km^2), MQ is the mean catchment runoff (m^3 s^{-1}), and ς is the dimensionless error of the regression. The gray shades of the regression lines in Figure 10 match those of the symbols for different lead times. All errors decrease very clearly with catchment area, although the rate of decrease differs with the error component and the lead time (Tables 2 and 3). The precipitation forecast errors (Figure 10(c)) decrease with catchment area. The precipitation forecast errors also decrease with decreasing lead time (from 48 to 12 h). As the lead times get close to the catchment response time, any errors of forecasted precipitation will no longer affect the runoff

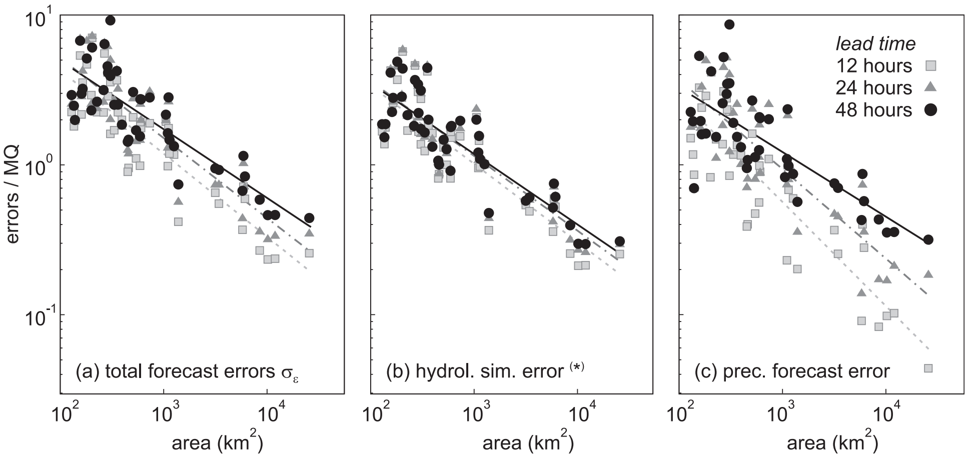

Figure 10. Errors scaled by mean catchment runoff versus catchment area for the top class (largest 10%) of ensemble spreads for 43 catchments. (a) Total forecast errors σ_ε, (b) hydrological simulation errors σ_{hysim}, and (c) precipitation forecast errors σ_{pfor}. The regression lines relate to different forecast lead times according to the gray scale. (*) The hydrological simulation error includes precipitation measurement and interpolation errors.

forecasts. This is particularly the case for the large catchments where the response times are longer than in the small catchments. Because of this the 12 h precipitation forecast errors in the large catchments are very small and therefore the dependence on area is stronger ($\beta = -0.695$) than for 48 h ($\beta = -0.433$). Table 2 gives the values of the slopes and the 90% confidence interval for the slope parameter of the regression [*Kottegoda and Rosso*, 1997, p. 352]. For the precipitation forecast error, the slopes of the regressions are significantly different for the lead times analyzed. The hydrological simulation errors (Figure 10(b)) decrease with catchment area due to the aggregation effects, while there is much less dependence on the lead time. This can be attributed to the updating procedures implemented in the model. *Komma et al.* [2008] showed that these updating procedures can reduce the error in particular for short lead times. For short lead times (up to 8–10 h) the additive error model reduces the errors significantly, whereas the Ensemble Kalman filtering reduces errors for the entire lead time. Without updating the model states one would not expect any dependence as these are strictly simulations. The decrease in the errors from 48 to 12 h (e.g., 1.18 to 1.02 for catchment areas of 1000 km^2) points to the value of the updating procedure for cases when the ensemble spread is large (top 10% of ensemble spreads). For the catchment areas of 10,000 km^2 the relative effect of the updating is about twice as big (0.40 to 0.31 for catchment areas of 1000 km^2) which is related to the longer response times and therefore longer autocorrelation in the hydrographs of the large catchments. For the hydrological simulation errors, the slopes of the regressions for the different lead times are not significantly different (i.e., the 90% confidence intervals for each lead time contain the estimated slopes for the other lead times). The total forecast errors (Figure 10(a)) are the combined result of the two error components. There is again a strong dependence on catchment area and a moderate dependence on the forecast lead time. The slopes of the regressions show a similar behavior as the slopes of the precipitation forecast error.

[44] It is now of interest to compare the error components as a function of catchment scale and lead time. For the 48 h lead time the precipitation forecast errors and hydrological simulation errors are of similar magnitudes. As the lead time decreases, the hydrological simulation errors change little while the precipitation forecast errors do, in particular in the large catchments. Obviously, for very short lead times the precipitation forecast errors would be zero. It is important to note, however, that this analysis is for those 10% of the time steps with the largest ensemble spreads, i.e., for a total of 36 days per year which not only includes floods. If individual large events were examined,

Table 2. Slopes β (–) of the Regression (Equation (7)) of Ensemble Spread $\overline{\sigma}_\varepsilon$, Total Forecast Errors σ_ε, Hydrological Simulation Errors σ_{hysim}, and Precipitation Forecast Errors σ_{pfor} (All Scaled by Mean Catchment Runoff) Against Catchment Area As Shown in Figures 10 and 11 for the Top Class (Largest 10%) of Ensemble Spreads[a]

Lead Time (h)	Slope β (–) of $\overline{\sigma}_\varepsilon$ (90% Confidence Interval)	Slope β (–) of σ_ε (90% Confidence Interval)	Slope β (–) of σ_{hysim} (90% Confidence Interval)	Slope β (–) of σ_{pfor} (90% Confidence Interval)
12	−0.754 (−0.603, −0.904)	−0.555 (−0.475, −0.635)	−0.522 (−0.441, −0.603)	−0.695 (−0.590, −0.800)
24	−0.626 (−0.497, −0.754)	−0.534 (−0.452, −0.617)	−0.499 (−0.422, −0.576)	−0.593 (−0.481, −0.705)
48	−0.557 (−0.426, −0.687)	−0.458 (−0.383, −0.534)	−0.471 (−0.401, −0.541)	−0.433 (−0.337, −0.529)

[a] Values in parentheses are the 90% confidence intervals for the mean of the regressions.

Table 3. Mean Ensemble Spread $\overline{\overline{\sigma}}_{\varepsilon}$, Total Forecast Errors σ_{ε}, Hydrological Simulation Errors σ_{hysim}, and Precipitation Forecast Errors σ_{pfor} (All Scaled by Mean Catchment Runoff) From the Regressions Against Catchment Area As in Figures 10 and 11 for the Top Class (Largest 10%) of Ensemble Spreads

Lead Time (h)	10^2 km^2 $\overline{\overline{\sigma}}_{\varepsilon}/\sigma_{\varepsilon}/\sigma_{\mathrm{hysim}}/\sigma_{\mathrm{pfor}}$	10^3 km^2 $\overline{\overline{\sigma}}_{\varepsilon}/\sigma_{\varepsilon}/\sigma_{\mathrm{hysim}}/\sigma_{\mathrm{pfor}}$	10^4 km^2 $\overline{\overline{\sigma}}_{\varepsilon}/\sigma_{\varepsilon}/\sigma_{\mathrm{hysim}}/\sigma_{\mathrm{pfor}}$
12	4.97/4.27/3.38/2.80	0.88/1.19/1.02/0.56	0.15/0.33/0.31/0.11
24	6.71/5.17/3.61/3.66	1.59/1.51/1.15/0.93	0.38/0.44/0.36/0.24
48	6.32/4.97/3.49/3.27	1.75/1.73/1.18/1.21	0.49/0.60/0.40/0.45

and in particular the rising limbs, the relative magnitudes of the two error sources may change with the precipitation forecast errors becoming much more important than the hydrological simulation errors [see e.g., *Blöschl et al.*, 2008].

[45] Figure 11 shows a similar analysis as Figure 10, however with the ensemble spread plotted against the catchment area. Tables 2 and 3 give the associated slopes and magnitudes of the ensemble spread along with those of the total forecast, the hydrologic simulation, and the precipitation forecast errors. Overall, the scaling characteristics of the ensemble spreads are very similar as those of the forecast errors. The decrease with catchment area is very similar. The slope of the dependency between ensemble spread and area also increases with decreasing lead time, similar to that of the forecast error, although it is somewhat steeper for the shortest lead time. Similarly the magnitudes of the ensemble spread and the total forecast errors compare well for all lead times and catchment areas with a tendency of underestimates for shorter lead times an larger catchment areas.

[46] Figure 12 summarizes the spread-skill relationship for a lead time of 48 h for all analyzed catchments. Only the time steps corresponding to the top 10% of the ensemble spreads are represented. Figure 12(a) shows the relation between ensemble spread and total forecast error, Figure 12(b) shows the relation between ensemble spread and precipitation forecast error, estimated according to equation (6). The size of the circles indicates the catchment area. The thin line indicates the 1:1 line.

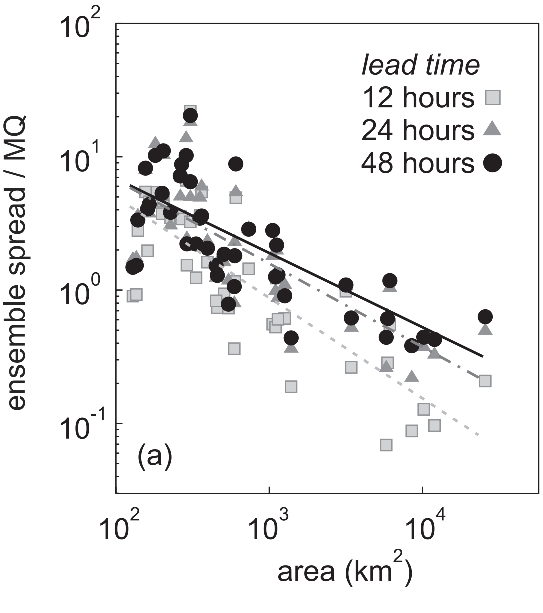

Figure 11. Ensemble spread scaled by mean catchment runoff versus catchment area for the top class (largest 10%) of ensemble spreads for 43 catchments.

[47] From Figure 12(a) it becomes apparent that on average the total forecast errors and the ensemble spreads are of similar magnitude for all catchments as the points are close to the 1:1 line. For small catchments the ensemble spread is on average larger than the total forecast error by a factor of 1.4, and for large catchments the factor is 0.9. In Figure 12(b) the points are somewhat farther from the 1:1 line, especially for the small catchments where the ensemble spread is bigger than the precipitation forecast error by an average factor of 2.3. For the large catchments, the factor is 1.3. This indicates that on average, the ensemble spread is representing the precipitation forecast errors and the total forecast errors for the top 10% of the ensemble spreads, meaning that the model uncertainty has not to be considered for the case of large ensemble spreads.

5. Discussion

[48] Even though not a main objective of this study, we discuss hereafter the overall performance of the flood forecasting system since it is operational. For a lead time of 24 h we obtain mean Brier Skill Scores *BSS* of 0.40 and, for a lead time of 48 h, the mean value was 0.25 when using the 90% percentile of runoff as reference forecast. These values are consistent with performances of other forecasting systems reported in literature. *Addor et al.* [2011] showed similar *BSS* values in the range of 0.30–0.50 for a lead time of one day and 0.10–0.30 for a lead time of 2 days, when evaluating the PREVAH model with COSMO-LEPS ensembles run on an hourly time step in a 336 km^2 catchment. *Rousset-Regimbeau et al.* [2007] and *Thirel et al.* [2008] evaluated ensemble runoff forecasts for 900 French catchments with areas ranging from 240 to 112,000 km^2. They used ECMWF forecasts as input into a coupled land surface and hydrogeological model run on a 3 h time step. Slightly larger BSS values can be expected due to averaging effects [e.g., *Skøien et al.*, 2003] in both time and scale. Indeed, *Rousset-Regimbeau et al.* [2007] who analyzed ensemble forecasts with a lead time up to 10 days found *BSS* values in the range of 0.4 to 1.0 for one-day runoff forecasts and in the range of 0.3 to 1 for five-day forecasts using the 90% percentile of runoff as reference forecast. Larger *BSS* values were obtained in large catchments, while lower *BSS* were generally found in smaller basins. Also *Thirel et al.* [2008] who focused on short range forecasts up to 2 days reported mean *BSS* of 0.90 for the first day of the forecast and 0.85 for the second day of the forecast.

[49] The first main objective of this study is to *quantify the contributions of precipitation forecast errors and hydrological simulation errors to the total forecast error*, particularly during flood events. We compare estimated runoff obtained by using forecasted precipitation and observed precipitation as input into the hydrologic model. To distinguish between different forecast situations, we stratify the analysis of forecast errors in classes of

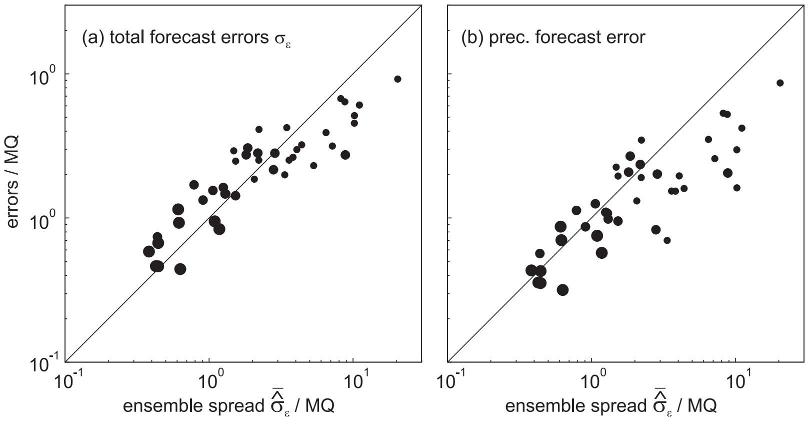

Figure 12. (a) Ensemble spread, scaled by mean runoff versus total forecast errors, scaled by mean runoff for the top 10% of the ensemble. (b) Ensemble spread, scaled by mean runoff versus precipitation forecast error, scaled by mean runoff for the top 10% of the ensemble spreads. Only the lead time of 48 h is represented. The size of the circles indicates the size of the catchments.

ensemble spreads. As expected, the analyses revealed that for small ensemble spreads, which indicate a small meteorological uncertainty and are more likely to occur when the runoff is constant or falling, the hydrological simulation error accounts for almost 100% of the total errors. With increasing ensemble spread the uncertainty from meteorology increases and is the main source of uncertainty for large ensemble spreads, which are more likely to occur when runoff is growing and therefore in situations of flood prognosis. For the largest 10% of the ensemble spreads in the four focus catchments, the contributions of the precipitation forecast errors account for 60%–85% of the error variance, whereas the hydrological simulation errors account for 15%–40% of the error variance. *Olsson and Lindström* [2008] found that for all phases of runoff the contributions of the meteorological and hydrological simulation errors are similar, and only for the rising limb the uncertainty in the meteorological forecasts dominated. However, *Olsson and Lindström* [2008] used a daily time step which reduces the anthropogenic variability introduced by the operation of reservoirs and lakes which are not taken into account in this study.

[50] For short lead times, the hydrological simulation error is the main source of uncertainty, as would be expected. For a lead time of 12 h, the ratio of the hydrological simulation error and the precipitation forecast error increases from 1.2 to 2.7 with the catchment size increasing from 100 to 10,000 km^2. For long lead times, the precipitation forecast error is dominant. For lead times of 48 h, the ratio of hydrological simulation error to precipitation forecast error decreases from 1.1 to 0.9 with increasing catchment size. This is due to two reasons: (1) for short lead times the uncertainty in the precipitation forecasts is small as no uncertainty is attributed to the meteorological ensemble in the first two hours [*Komma et al.*, 2007] and (2) the response time of the catchments is longer than the lead time of the forecasts, meaning that it takes the input variability from the meteorological ensemble forecasts longer to reach the basin outlet than the lead time of the forecast [*Renner et al.*, 2009].

[51] All errors decrease clearly with increasing catchment area and decreasing lead time. The decrease of precipitation forecast errors with catchment area can be attributed to averaging effects as discussed by *Sivapalan* [2003] and *Skøien and Blöschl* [2006] (i.e., the catchment acts as a space-time filter on precipitation, especially for increasing catchment areas, meaning that for bigger catchments the accuracy required in predicted spatial location and temporal dynamics of precipitation is lower). The precipitation forecast error scales with $\beta = -0.695$ for a lead time of 12 h, and $\beta = -0.433$ for a lead time of 48 h. The smaller errors for short lead times and large catchments can be attributed to the fact that runoff in large catchments does not depend much on the future precipitation at short lead times, but on the observed precipitation. The runoff in small catchments, which have shorter response times, are more dependent on the future precipitation, even at short lead times. In fact, a lead time of 12 h can be deemed large in a small catchment, characterized by a small response time. For the 48 h lead time, the runoff in large catchments is affected more by the future precipitation, which is reflected in the smaller scaling factor, because 48 h can be deemed as short in comparison to large response times. The hydrological simulation error (Figure 10(b)) also decreases with catchment area, but the dependence on the lead time is smaller ($\beta = -0.522$ for 12 h lead time and $\beta = -0.471$ for 48 h lead time). Two reasons contribute to this: (1) the more linear (smoothed) response in large catchments is easier to model than the larger nonlinearities in small catchments [e.g., *Rogger et al.*, 2012] and (2), as showed by *Komma et al.* [2008], updating procedures can reduce the error of forecasts in particular for short lead times and large catchments. *Komma et al.* [2008] analyzed rising limbs and showed errors 12% smaller compared to forecasts without updating for a lead time of 12 h. The total forecast errors as a combination of the two components show again a strong dependence on catchment area and a moderate dependence on the forecast lead time ($\beta = -0.555$ and -0.458 for lead times of 12 and 48 h, respectively). Having the variability of forecast errors decreasing with catchment size is partly due to the fact that the variability of streamflow is lower in large

catchments. In fact the coefficient of variation of the entire runoff hydrographs scales with −0.254 with catchment area for the 43 catchments in this study. This slope is much lower than the one of a random field (scaling factor −0.5) because of the spatial and temporal correlation of rainfall and runoff production over the catchments [*Viglione et al.*, 2010a, 2010b]. Similar values are found by *Merz and Blöschl* [2003] who focused on the evaluation of mean annual flood and further distinguished different runoff situations. They found coefficients of variation in the range of −0.205 for snowmelt induced floods and −0.413 for flash floods, which are less spatially and temporally organized [*Viglione et al.*, 2010b]. Since the scaling factors for the forecast errors are higher (in absolute value), we can conclude that the performance of the forecasting system increases with catchment size.

[52] The second objective of this study is to *evaluate the capability of the runoff ensemble forecasts to represent the total forecast error* as a function of lead time. Ensemble forecasts have been considered a suitable tool for quantifying and communicating the uncertainties of forecasts [see e.g., *Hlavcova et al.*, 2006; *Demeritt et al.*, 2007], as the spread of the ensemble members can be used as a measure of forecast uncertainty [*Buizza*, 2003]. In the studies of *Johnell et al.* [2007], which is based on ECMWF ensemble forecasts and *Jaun and Ahrens* [2009], which is based on downscaled ECMWF ensemble forecasts, forecast errors increased with increasing ensemble spread and with increasing runoff for all catchments. *Johnell et al.* [2007] showed an increase of the mean absolute error of the ensemble median with increasing ensemble spread class from 5% to 30%, averaged over all catchments which is smaller than the mean absolute error of the deterministic forecasts in this study. For a lead time of 48 h, we have estimated values in the order of 8% to 40% for the large catchments, and values in the order of 15% to 140% for small catchments. The larger errors can be attributed to the facts that we use an hourly time step and *Johnell et al.* [2007] used a daily time step for estimating runoff. *Komma et al.* [2007] analyzed the forecasts of five flood events in a 600 km^2 catchment in Austria and found mean normalized absolute errors of about 40% when evaluating the entire events at a lead time of 48 h, which is somewhat lower than the values in this study for medium catchments. In the study of *Komma et al.* [2007] the number of precipitation stations per 100 km^2 is 1.3, while in this study on average 0.4 stations per 100 km^2 are available. An increasing number of precipitation stations per catchment as discussed by *Merz et al.* [2009] allows better estimates of catchment precipitation, which further reduces the forecast errors in larger catchments. *Jaun and Ahrens* [2009] who evaluated daily forecasts for 23 catchments in Switzerland show ensemble spreads and forecast errors of similar magnitude for large ensemble spreads. For small ensemble spreads positive forecast errors were 1.5 to 5 times larger than the (small) ensemble spread, and with increasing ensemble spread the factor between error and spread decreased to a value of 0.9 for large ensemble spreads. For negative forecast errors, the factor was larger in the range of 1.5 to 100 for small ensemble spreads and decreased to a factor of 1.3 for large ensemble spreads. *Jaun and Ahrens* [2009] concluded that the uncertainty is covered by the ensemble with appropriate spread. For small ensemble spreads we observe total forecast errors which are larger than the ensemble spread by a factor in the range of 10 to 100 for all lead times. For large ensemble spreads the total forecast errors and the ensemble spreads are much closer to the 1:1 line. On average, for a lead time of 48 h the ensemble spread is larger than the total forecast error by a factor of 1.4 for small catchments, and for large catchments the factor is 0.9. This indicates that the ensemble spread is representative of the total forecast errors in situations with large ensemble spreads, which are of particular interest for flood forecasting, but there is still potential to improve the spread-skill relationship also for small ensemble spreads [see e.g., *Schaake et al.*, 2004; *Olsson and Lindström*, 2008]. Even if the ensemble spread does not always capture the magnitude of the forecast error, there is a clear correlation between the two, as shown by calculating the Spearman's rank correlation, meaning that the ensemble spread can always be used as an index of forecast errors. For the ensemble spread the scaling factor is large in absolute values for short lead times (−0.754 for 12 h), for the same reason, i.e., the runoff in large catchments does not depend much on the future precipitation at short lead times, but on the observed precipitation.

6. Conclusions

[53] In this study we perform an error analysis on the forecasts of an operational flood forecasting system for the Danube tributaries in Austria and Germany. We carried out a spread-skill analysis with two different measures for the spread, the standard deviation and the interquartile range. Both analyses are consistent in terms of the results: For the 90% of the time steps with small ensemble spreads the forecast errors are larger than the ensemble spread and for the 10% of the time steps with the largest ensemble spreads the forecast errors and the ensemble spread are of similar magnitude. For flood forecasting we are mainly interested in the 10% of the time steps with the largest ensemble spreads, which are typically occurring during the rising limbs of flood events. For these time steps, clear scaling relationships of the forecast error components with catchment area have been found and discussed. The results indicate that the forecast error components, the hydrological simulation error, and the precipitation forecast error, decrease with increasing catchment size. As one would expect, the error component of the hydrological simulation does not differ significantly with increasing lead time, whereas the error component of the precipitation forecast differs significantly with changing lead time. For short lead times, the ratio of the hydrological simulation error to precipitation forecast error is 1.2 to 2.7 with increasing catchment size from 100 to 10,000 km^2. For long lead times the ratio of hydrological simulation error to precipitation forecast error decreases from 1.1 to 0.9 with increasing catchment size. A similar scaling is also found for ensemble spreads, which are shown to represent quantitatively the total forecast error when forecasting floods.

[54] We believe that this kind of scaling analysis of the forecast errors should be performed in other case studies as well, e.g., in other climates and using different models. Comparing different studies is needed for understanding, which is the idea underlying the so called comparative hydrology, in which simple indices are used for quantifying similarities of processes and models across scales [*McDonnell and Woods*, 2004; *Blöschl*, 2006].

Appendix A

[55] A summary of catchment characteristics (area, mean annual precipitation, runoff and temperature, and difference in elevation) as well as model performance measures (*Nester et al.* [2011]) is given in Table A1.

Appendix B

[56] Statistical measures used to evaluate the model performance include the *Nash and Sutcliffe* [1970] coefficient of efficiency (*nsme*):

$$nsme = 1 - \frac{\sum_{i=1}^{n} (Q_{\mathrm{sim},i} - Q_{\mathrm{obs},i})^2}{\sum_{i=1}^{n} (Q_{\mathrm{obs},i} - \overline{Q_{\mathrm{obs}}})^2}, \quad \text{(B1)}$$

where $Q_{\mathrm{obs},i}$ and $Q_{\mathrm{sim},i}$ are observed and simulated runoff at hour i, respectively, and $\overline{Q_{\mathrm{obs}}}$ is the mean observed runoff over the calibration or validation period of n hours. *nsme* values can range from $-\infty$ to 1. A perfect match between simulation and observation implies $nsme = 1$; $nsme = 0$ indicates that the model predictions are as accurate as the mean of the observed data, and $nsme < 0$ occurs when the observed mean is a better predictor than the model.

[57] As a measure of bias the volume error *VE* was used:

$$VE = \frac{\sum_{i=1}^{n} Q_{\mathrm{sim},i} - \sum_{i=1}^{n} Q_{\mathrm{obs},i}}{\sum_{i=1}^{n} Q_{\mathrm{obs},i}}. \quad \text{(B2)}$$

[58] The value can be positive or negative, with a *VE* of an unbiased model being 0. Values larger and smaller than 0 imply over- and underestimation, respectively.

Table A1. Stream Gauges Used for Detailed Analyzes[a]

Gauge/Catchment	Area (km^2)	MAP (mm yr^{-1})	MAR (mm yr^{-1})	MAT (°C)	Δh (m)	*nsme* (calib/valid)	*VE* (calib/valid)
Molln/Steyrling	129	1705	865	7.3	1349	0.58/0.67	0.30/0.14
Erlaufboden/Erlauf	136	1590	1190	6.1	1290	0.73/0.66	0.01/−0.05
Oberkappel/Ranna	139	1055	620	8.2	464	0.69/0.57	−0.02/−0.20
Krenstetten/Urlbach	156	1050	430	8.6	475	0.61/0.72	0.16/0.08
Kremsmünster/Krems	161	1260	540	8.7	931	0.66/0.72	0.06/0.05
Haging/Antiesen	165	1030	440	8.9	430	0.70/0.70	0.03/0.19
Cholerakapelle/Schwechat	181	890	260	8.9	609	0.79/0.50	0.14/−0.01
Obermühl/Kl. Mühl	200	950	480	8.6	626	0.64/0.52	−0.01/−0.10
Siegersdorf/Gr. Tulln	204	815	200	9.2	720	0.46/0.64	0.60/0.35
Rottenegg/Rodl	227	960	385	7.7	765	0.62/0.36	0.05/0.12
St. Georgen/Gusen	263	860	250	8.2	655	0.70/0.68	0.11/−0.06
Atzenbrugg/Perschling	268	820	200	9.1	593	0.62/0.72	0.25/0.22
Hirtenberg/Triesting	287	960	230	8.7	758	0.58/0.63	−0.08/−0.20
Hofstetten/Pielach	290	1425	690	8.2	1002	0.81/0.62	−0.05/0.00
Imbach/Krems	306	720	210	7.7	704	0.33/0.42	-0.22/0.00
Haid/Naarn*	306	915	380	7.3	810	0.75/0.65	0.19/0.19
Lilienfeld/Traisen*	333	1440	860	7.3	1275	0.85/0.69	−0.12/0.00
Pfaffing/Aschach	353	975	330	8.5	419	0.55/0.49	0.30/0.40
Fraham/Innbach	362	940	340	8.5	509	0.62/0.60	−0.06/0.12
Obergäu/Lammer	395	1860	1215	5.6	1910	0.59/0.69	0.10/0.01
Teufelmühle/Gr. Mühl	450	1055	580	7.6	896	0.72/0.49	0.13/−0.02
Penningersteg/Alm	459	1600	960	7.7	2038	0.46/0.62	0.17/0.19
Opponitz/Ybbs	507	1800	1140	6.6	1471	0.84/0.79	0.17/0.07
Klaus/Steyr	542	1750	1530	6.6	1946	0.59/0.58	0.02/−0.21
Wildalpen/Salza	592	1630	1150	5.9	1942	0.74/0.73	0.06/0.00
Niederndorf/Erlauf	595	1460	750	7.6	1551	0.79/0.73	0.09/−0.03
Schwertberg/Aist	605	885	305	7.6	788	0.69/0.65	0.06/−0.02
Windpassing/Traisen	735	1270	640	7.9	1354	0.80/0.75	−0.09/−0.10
Ruhstorf/Rott	1,052	840	220	8.4	223	0.58/0.07	0.03/0.13
Rosenheim/Mangfall	1099	1520	430	7.9	1386	0.71/0.36	−0.13/0.03
Greimpersdorf/Ybbs*	1,116	1480	840	7.5	1607	0.86/0.80	−0.02/0.02
Siezenheim/Saalach	1,139	1730	910	5.9	2370	0.71/0.74	0.08/0.05
Fischerau/Ager	1,260	1360	720	8.2	1403	0.77/0.43	0.04/0.03
Gmunden/Traun*	1,390	1810	1425	6.3	2499	0.76/0.78	−0.01/−0.10
Golling/Salzach	3,161	1630	1100	3.9	3059	0.63/0.73	0.03/0.02
Wels/Traun	3,426	1600	855	7.4	2603	0.82/0.85	0.06/0.02
Innsbruck/Inn	5,792	1110	825	1.5	3357	0.71/0.74	−0.04/0.03
Steyr/Enns	5,915	1510	1060	6.1	2556	0.80/0.71	0.09/−0.02
Oberndorf/Salzach	6,120	1650	1020	5.1	3156	0.65/0.74	0.07/−0.02
Brixlegg/Inn	8,503	1190	900	2.0	3420	0.73/0.76	−0.04/0.00
Rosenheim/Inn	10,186	1285	820	3.4	3473	0.68/0.76	0.08/0.03
Wasserburg/Inn	11,977	1270	800	3.9	3495	0.65/0.73	−0.06/0.03
Schärding/Inn	25,664	1340	830	5.1	3632	0.77/0.80	0.00/0.07

[a] MAP, MAR, and MAT (2002–2009) is mean annual precipitation, runoff, and temperature, respectively. Δh is the difference in elevation within the catchment. *nsme* stands for Nash-Sutcliffe model efficiency, *VE* stands for volume error (figures are for the calibration period/validation period) (from *Nester et al.* [2011a]). *denote catchments used for detailed analyzes.

[59] **Acknowledgments.** We would like to thank the Hydrographic Services of Lower Austria and Upper Austria for providing the discharge data and the Central Institute for Meteorology and Geodynamics (ZAMG) in Vienna for providing the meteorological data. The study was performed as part of developing an operational flood foresting system for the Austrian Danube and tributaries.

References

Addor, N., S. Jaun, F. Fundel, and M. Zappa (2011), An operational hydrological ensemble prediction system for the city of Zurich (Switzerland): Skill, case studies and scenarios, *Hydrol. Earth Syst. Sci.*, *15*, 2327–2347, doi:10.5194/hess-15-2327-2011.

Blöschl, G. (2006), Hydrologic synthesis: Across processes, places, and scales, *Water Resour. Res. 42*(3), W03S02, doi:10.1029/2005WR004319.

Blöschl., G. (2008), Flood warning—On the value of local information, *Int. J. River Basin Manage.*, *6*(1) 41–50, doi:10.1080/15715124.2008.9635336.

Blöschl, G., C. Reszler, and J. Komma (2008), A spatially distributed flash flood forecasting model, *Environ. Model. Software*, *23*(4), 464–478.

Brier, G. W. (1950), Verification of forecasts expressed in terms of probability, *Mon. Weather Rev.*, *78*, 1–3.

Buizza, R., (2003), *Encyclopaedia of Atmospheric Sciences*, Chap. Weather Prediction: Ensemble Prediction, pp. 2546–2557, Academic, London.

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

Cloke, H. L., and F. Pappenberger (2009), Ensemble flood forecasting: A review, *J. Hydrol.*, *375*, 613–626, doi:10.1016/j.jhydrol.2009.06.005.

Coccia, G., and E. Todini (2011), Recent developments in predictive uncertainty assessment based on the model conditional processor approach, *Hydrol. Earth Syst. Sci.*, *15*, 3253–3274, doi:10.5194/hess-15-3253-2011.

Demeritt, D., H. Cloke, F. Pappenberger, J. Thielen, J. Bartholmes, and M. H. Ramos (2007), Ensemble predictions and perceptions of risk, uncertainty, and error in flood forecasting, *Environ. Haz.*, *7*(2), 115–127, doi:10.1016/j.envhaz.2007.05.001.

Di Baldassarre, G., and A. Montanari (2009), Uncertainty in river discharge observations: A quantitative analysis, *Hydrol. Earth Syst. Sci.*, *13*, 913–921, doi:10.5194/hess-13-913-2009.

Gouweleeuw, B. T., J. Thielen, G. Franchello, A. P. J. De Roo, and R. Buizza (2005), Flood forecasting using medium-range probabilistic weather prediction, *Hydrol. Earth Syst. Sci.*, *9*(4), 365–380.

Grimit, E. P., and C. F. Mass (2007), Measuring the ensemble spread–error relationship with a probabilistic approach: Stochastic ensemble results, *Mon. Weather Rev.*, *135*, 203–221, doi:10.1175/MWR3262.1.

Haiden, T., A. Kann, K. Stadlbacher, M. Steinheimer, and C. Wittmann (2010), Integrated nowcasting through comprehensive analysis (INCA)—System overview, ZAMG Report, 60 pp., [available at http://www.zamg.ac.at/fix/INCA_system.doc (last accessed 14 November 2011)], Central Institute for Meteorology and Geodynamics, Vienna, Austria.

Haiden, T., A. Kann, C. Wittmann, G. Pistotnik, B. Bica, and C. Gruber (2011), The integrated nowcasting through comprehensive analysis (INCA) system and its validation over the eastern Alpine region, *Weather Forecasting*, *26*, 166–183, doi:10.1175/2010WAF2222451.1.

Hamill, T. (2001), Interpretation of rank histograms for verifying ensemble forecasts, *Mon. Weather Rev.*, *129*, 550–560, doi:10.1175/1520-0493(2001)129<0550:IORHFV>2.0.CO;2.

Hamill T. M., R. Hagedorn, and J. S. Whitaker (2008), Probabilistic forecast calibration using ECMWF and GFS ensemble reforecasts. Part II: Precipitation, *Mon. Weather Rev.*, *136*, 2620–2632, doi:10.1175/2007MWR2411.1.

Hlavcova, K., J. Szolgay, R. Kubes, S. Kohnova, and M. Zvolensky (2006), Routing of numerical weather predictions through a rainfall-runoff model, in *Transboundary Floods: Reducing Risks through Flood Management*, edited by J. Marsalek, G. Stancalie, and G. Balint, pp. 79–90, NATO Science Series: IV: Earth and Environmental Sciences, Springer, Dordrecht, The Netherlands, doi:10.1007/1-4020-4902-1_8.

Hopson, T. M., and P. J. Webster (2010), A 1–10-day ensemble forecasting scheme for the major river basins of Bangladesh: Forecasting severe floods of 2003–07, *J. Hydrometeorol.*, *11*(3), pp. 618–641.

Jaun, S., and B. Ahrens (2009), Evaluation of a probabilistic hydrometeorological forecast system, *Hydrol. Earth Syst. Sci.*, *13*(7), pp. 1031–1043.

Johnell, A., G. Lindström, and J. Olsson (2007), Deterministic evaluation of ensemble runoff predictions in Sweden, *Nord. Hydrol.*, *38*(4–5), 441–450, doi:10.2166/nh.2007.022.

Komma, J., C. Reszler, G. Blöschl, and T. Haiden (2007), Ensemble prediction of floods—Catchment non-linearity and forecast probabilities, *Nat. Haz. Earth Sys. Sci.*, *7*, 431–444, doi:10.5194/nhess-7-431-2007.

Komma, J., G. Blöschl, and C. Reszler (2008), Soil moisture updating by ensemble Kalman filtering in real-time flood forecasting, *J. Hydrol.*, *357*, 228–242.

Kottegoda, N. T., and R. Rosso (1997), *Statistics, Probability and Reliability for Civil and Environmental Engineers*, McGraw-Hill, New York.

Krzysztofowicz, R. (2001), Integrator of uncertainties for probabilistic river stage forecasting: Precipitation-dependent model, *J. Hydrol.*, *249*, 69–85, doi:10.1016/S0022-1694(01)00413-9.

Krzysztofowicz, R., and K. S. Kelly (2000), Hydrologic uncertainty processor for probabilistic river stage forecasting, *Water Resour. Res.*, *36*(11), 3265–3277, doi:10.1029/2000WR900108.

Lalaurette, F., J. Bidlot, L. Ferranti, A. Ghelli, F. Grazzini, M. Leutbecher, J.-E. Paulsen, and P. Viterbo (2005), Verification statistics and evaluations of ECMWF forecasts in 2003–2004, *Tech. Rep. 463*, Eur. Cent. for Medium-Range Weather Forecasts, Reading, UK. [available at http://www.ecmwf.int/publications/library/ecpublications/pdf/tm/401-500/tm463.pdf.]

McDonnell, J. J., and R. Woods (2004), On the need for catchment classification, *J. Hydrol.*, *299*(1–2), 2–3, doi:10.1016/j.jhydrol.2004.09.003.

Merz, R., and G. Blöschl (2003), A process typology of regional floods, *Water Resour. Res.*, *39*(12), 1340, doi:10.1029/2002WR001952.

Merz, R., J. Parajka, and G. Blöschl (2009), Scale effects in conceptual hydrologic modeling, *Water Resour. Res. 45*, W09405, doi:10.1029/2009WR007872.

Montanari, A. (2007), What do we mean by 'uncertainty'? The need for a consistent wording about uncertainty assessment in hydrology, *Hydrol. Processes*, *21*, 841–845, doi:10.1002/hyp.6623.

Montanari, A., and A. Brath (2004), A stochastic approach for assessing the uncertainty of rainfall-runoff simulations, *Water Resour. Res.*, *40*, W01106, doi:10.1029/2003WR002540.

Montanari, A., and G. Grossi (2008), Estimating the uncertainty of hydrological forecasts: A statistical approach, *Water Resour. Res.*, *44*, W00B08, doi:10.1029/2008WR006897.

Montanari, A., C. A. Shoemaker, and N. van de Giesen (2009), Introduction to special section on uncertainty assessment in surface and subsurface hydrology: An overview of issues and challenges, *Water Resour. Res.*, *45*, W00B00, doi:10.1029/2009WR008471.

Nester, T., R. Kirnbauer, D. Gutknecht, and G. Blöschl (2011), Climate and catchment controls on the performance of regional flood simulations, *J. Hydrol. 402*, 340–356, doi:10.1016/j.jhydrol.2011.03.028.

Nester, T., R. Kirnbauer, J. Parajka, and G. Blöschl (2012), Evaluating the snow component of a flood forecasting model, *Hydrol. Res.*, doi:10.2166/nh.2012.041, in press.

Olsson, J., and G. Lindström (2008), Evaluation and calibration of operational hydrologic ensemble forecasts in Sweden, *J. Hydrol. 350*, 14–24, doi:10.1016/j.jhydrol.2007.11.010.

Pappenberger, F., K. J. Beven, N. M. Hunter, P. D. Bates, B. T. Gouweleeuw, J. Thielen, and A. P. J. de Roo (2005), Cascading model uncertainty from medium range weather forecasts (10 days) through a rainfall-runoff model to flood inundation predictions within the European Flood Forecasting System (EFFS), *Hydrol. Earth Syst. Sci.*, *9*(4), 381–393.

Renner, M., M. G. F. Werner, S. Rademacher, and E. Sprokkereef (2009), Verification of ensemble flow forecasts for the River Rhine, *J. Hydrol.*, *376*, 463–475, doi:10.1016/j.jhydrol.2009.07.059.

Rogger, M., H. Pirkl, A. Viglione, J. Komma, B. Kohl, R. Kirnbauer, R. Merz, and G. Blöschl (2012), Step changes in the flood frequency curve: Process controls, *Water Resour. Res.*, *48*, W05544, doi:10.1029/2011WR011187.

Roulin, E. (2007), Skill and relative economic value of medium-range hydrologic ensemble predictions, *Hydrol. Earth Syst. Sci.*, *11*, 725–737, doi:10.5194/hess-11-725-2007.

Roulin, E., and S. Vannitsem (2005), Skill of medium-range hydrologic ensemble predictions, *J. Hydrometeorol.*, *6*(5), 729–744.

Rousset-Regimbeau, F., F. Habets, E. Martin, and J. Noilhan (2007), Ensemble runoff forecasts over France, ECMWF Newsletter No. 111, [available at http://www.ecmwf.int/publications/newsletters/pdf/111.pdf (last accessed 29 September 2011)], Eur. Cent. for Medium-Range Weather Forecasts, Reading, UK.

Schaake, J., S. Perica, M. Mullusky, J. Demargne, E. Welles, and L. Wu (2004), Pre-processing of atmospheric forcing for ensemble runoff prediction, Proceedings of the 84th AMS Annual Meeting held in Seattle, WA, January 2004, 5 pp., [available at http://ams.confex.com/ams/pdfpapers/72172.pdf (last accessed 29 September 2011)], Am. Meteorol. Soc., Boston, Mass.

Scherrer, S. C., C. Appenzeller, P. Eckert, and D. Cattani (2004), Analysis of the spread–skill relations using the ECMWF ensemble prediction system over Europe, *Weather Forecast.*, *19*, 552–565.

Schmeits M. J., and K. J. Kok (2010), A comparison between raw ensemble output, (modified) Bayesian model averaging, and extended logistic regression using ECMWF ensemble precipitation reforecasts, *Mon. Weather Rev.*, *138*, 4199–4211, doi:10.1175/2010MWR3285.1.

Sivapalan, M. (2003), Process complexity at hillslope scale, process simplicity at the watershed scale: is there a connection?, *Hydrol. Processes*, *17*, 1037–1041.

Skøien, J. O., and G. Blöschl (2006), Catchments as space-time filters—A joint spatio-temporal geostatistical analysis of runoff and precipitation, *Hydrol. Earth Syst. Sci.*, *10*, 645–662.

Skøien, J. O., G. Blöschl, and A. W. Western (2003), Characteristic space scales and timescales in hydrology, *Water Resour. Res. 39*(10), 1304, doi:10.1029/2002WR001736.

Stefanova, L., and T. N. Krishnamurti (2002), Interpretation of seasonal climate forecast using Brier skill score, *J. Climate*, *15*, 537–544.

Steinheimer, M., and T. Haiden (2007), Improved nowcasting of precipitation based on convective analysis fields, *Adv. Geosci.*, *10*, 125–131.

Szolgay, J. (2004), Multilinear flood routing using variable travel-time discharge relationships on the Hron river, *J. Hydro. Hydromech.*, *52*, 4.

Talagrand, O., R. Vautard, and B. Strauss (1997), Evaluation of probabilistic prediction systems, in *Proceedings, ECMWF Workshop on Predictability*, pp. 1–25, Eur. Cent. for Medium-Range Weather Forecasts, Reading, UK.

Thielen, J., J. Bartholmes, M.-H. Ramos, and A. de Roo (2009), The European Flood Alert System—Part 1: Concept and development, *Hydrol. Earth Syst. Sci.*, *13*, 125–140, www.hydrol-earth-syst-sci.net/13/125/2009.

Thirel, G., F. Rousset-Regimbeau, E. Martin, and F. Habets (2008), On the impact of short-range meteorological forecasts for ensemble runoff predictions, *J. Hydrometeorol.*, *9*, 1301–1317.

Verbunt, M., A. Walser, J. Gurtz, A. Montani, and C. Schär (2007), Probabilistic flood forecasting with a limited-area ensemble prediction system: Selected case studies, *J. Hydrometeorol.*, *8*, 897–909, doi:10.1175/JHM594.1.

Viglione, A., G. B. Chirico, R. Woods, and G. Blöschl (2010a), Generalised synthesis of space–time variability in flood response: An analytical framework, *J. Hydrol. 394*(1–2), 198–212. doi:10.1016/j.jhydrol.2010.05.047.

Viglione, A., G. B. Chirico, J. Komma, R. Woods, M. Borga, and G. Blöschl (2010b), Quantifying space-time dynamics of flood event types, *J. Hydrol.*, *394*(1-2), 213–229, doi:10.1016/j.jhydrol.2010.05.041.

Weerts, A. H., H. C. Winsemius, and J. S. Verkade (2011), Estimation of predictive hydrological uncertainty using quantile regression: Examples from the National Flood Forecasting System (England and Wales), *Hydrol. Earth Syst. Sci.*, *15*, 255–265, doi:10.5194/hess-15-255-2011.

Wilks, D. S. (1995), *Statistical Methods in the Atmospheric Sciences: An Introduction*, 467 pp., Academic, San Diego, Calif.

Zappa, M., S. Jaun, U. Germann, A. Walser, and F. Fundel (2011), Superposition of three sources of uncertainties in operational flood forecasting chains, *Atmos. Res. 100*, 246–262, doi:10.1016/j.atmosres.2010.12.005.