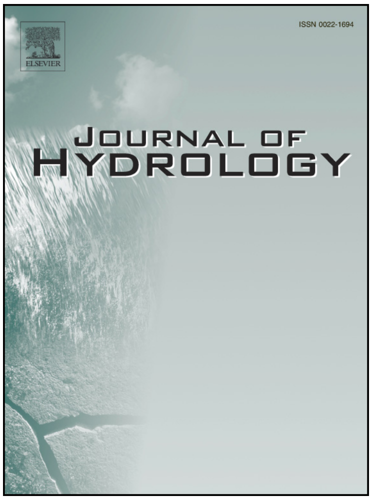


# Evaluation of ensemble streamflow predictions in Europe

Lorenzo Alfieri, Florian Pappenberger, Fredrik Wetterhall, Thomas Haiden, David Richardson, Peter Salamon

PII: S0022-1694(14)00495-8
DOI: http://dx.doi.org/10.1016/j.jhydrol.2014.06.035
Reference: HYDROL 19709

To appear in: *Journal of Hydrology*

Received Date: 2 December 2013
Revised Date: 7 May 2014
Accepted Date: 21 June 2014

Please cite this article as: Alfieri, L., Pappenberger, F., Wetterhall, F., Haiden, T., Richardson, D., Salamon, P., Evaluation of ensemble streamflow predictions in Europe, *Journal of Hydrology* (2014), doi: http://dx.doi.org/10.1016/j.jhydrol.2014.06.035

## Evaluation of ensemble streamflow predictions in Europe

Lorenzo Alfieri[1,2], Florian Pappenberger[1], Fredrik Wetterhall[1], Thomas Haiden[1], David Richardson[1], and Peter Salamon[2].

[1] European Centre for Medium-Range Weather Forecasts, Reading, UK

[2] European Commission - Joint Research Centre, Ispra, Italy

Correspondence to: Lorenzo Alfieri
European Commission - Joint Research Centre
TP 122, Via E. Fermi 2749
21027 Ispra (VA)
Italy
Tel: +39 0332 78 3835
Email: lorenzo.alfieri@jrc.ec.europa.eu

## Abstract

In operational hydrological forecasting systems, improvements are directly related to the continuous monitoring of the forecast performance. An efficient evaluation framework must be able to spot issues and limitations and provide feedback to the system developers. In regional systems, the expertise of analysts on duty is a major component of the daily evaluation. On the other hand, large scale systems need to be complemented with semi-automated tools to evaluate the quality of forecasts equitably in every part of their domain.

This article presents the current status of the monitoring and evaluation framework of the European Flood Awareness System (EFAS). For each grid point of the European river network, 10-day ensemble streamflow predictions are evaluated against a reference simulation which uses observed meteorological fields as input to a calibrated hydrological model. Performance scores are displayed over different regions, forecast lead times, basin sizes, as well as in time, considering average scores for moving 12-month windows of forecasts. Skilful predictions are found in medium to large rivers over the whole 10-day range. On average, performance drops significantly in river basins with upstream area smaller than 300 km$^2$, partly due to underestimation of the runoff in mountain areas. Model limitations and recommendations to improve the evaluation framework are discussed in the final section.

Keywords: flood early warning; ensemble streamflow predictions; CRPS; skill scores; distributed hydrological modelling.

## 1 Introduction

Operational hydrological forecasting systems play a key role in the water resources management and in the preparedness against extreme events. Assessing their performance is crucial for the error diagnostic and in the planning of development work to improve the system accuracy and extend the forecast lead time. A vast number of regional and national hydro-meteorological centres have flood forecasting and early warning systems in place based on weather predictions (see Alfieri et al., 2012 for a recent review of European systems). At the same time, the number of ensemble-based systems is increasing (Cloke and Pappenberger, 2009; Wetterhall et al., 2013), with the aim of describing part of the uncertainty embedded in the forecasts. The evaluation of the forecast accuracy is regularly performed in many operational systems, where verification scores need to be complemented by the local knowledge and experience of analysts on duty. Further, skill scores are rarely displayed publicly, to prevent misinterpretation of results and avoid the need for simplifying their information content for a wider recipient of users. Yet, reporting on past performance by means of verification scores is listed as one of the main priorities of users, to increase the trust in forecasting systems (Wetterhall et al., 2013).

Assessing the forecast performance over large domains raises the challenge of comparing river points with different upstream area and hydrological regimes. In these cases, a widespread approach to tackle the forecast verification is to compute scores based on the probability of thresholds exceedance (e.g., warning levels), that can be defined in a consistent way for every point. While this is a standard practice for early warning systems (e.g., Bartholmes et al., 2009; Gourley et al., 2012), it is also applied to the verification of categorical events for any set of thresholds (Thirel et al., 2008). If quantitative values are considered, the choice of performance scores becomes wider (Legates and McCabe, 1999;

Wilks, 2006), though only a relatively small subset is specifically dedicated to evaluate the quality of ensemble forecasts (Brown et al., 2010). The comparison of forecast skill in several river sections is often performed through benchmarking against simplified simulations (Pappenberger et al., submitted), previous model versions (Arheimer et al., 2011), different input data (e.g., Renner et al., 2009), or climatological values (Demargne et al., 2010; Verkade et al., 2013; Wood et al., 2005). An alternative method consists in normalizing forecasts and reference values before the evaluation (Pappenberger et al., 2010). Trinh et al. (2013) used a similar concept to propose a modified Continuous Ranked Probability Score (CRPS) which is suitable to compare forecast performance at different river sections. In operational systems, the forecast performance must be monitored and updated continuously in time. Hence, a skill assessment based on different scores and benchmarks (e.g., Alfieri et al., 2013a; Randrianasolo et al., 2010) is often preferred in order to analyze different aspects of the forecast performance at several locations and quickly detect trends over time or weaknesses.

In 2012, after the transfer of the EFAS operational suite to the European Centre for Medium-Range Weather Forecasts (ECMWF), a commitment was made to set up an evaluation framework of the hydrological forecasts, in order to monitor their performance over time and after major system updates. The idea was to implement an automated procedure to regularly produce and update summary skill scores for the whole computation domain, able to spot a variety of possible problems and address subsequent in-depth analysis. Among the main challenges to face was the choice of appropriate skill scores, the handling of large data sets, and the visualization of results through concise and intuitive graphs.

This article presents the current status of implementation of such an evaluation framework, after one year of operational runs at ECMWF. Streamflow forecasts at every grid point of the river network are verified against a reference simulation which uses observed meteorological fields as input to a calibrated hydrological model.

## 2 Data and Methods

### 2.1 Model framework

The main components of the EFAS hydro-meteorological forecasting chain are: a) a hydrological model, b) weather forecasts, and c) meteorological observations, to update the initial model states and for verification purpose (see Figure 1). Each of these three components has inherent uncertainty, which can be described in the modelling framework and propagated to the output discharge. The current EFAS system is a multi-model ensemble approach, in that it accounts for the uncertainty of input weather forecasts using model runs from different meteorological centres in Europe. These include two deterministic forecasts, from the ECMWF (ECMWF-HiRes, Miller et al., 2010) and from the German Weather Service (DWD, see Majewski et al., 2002; Steppeler et al., 2003), and two ensemble forecasts, from the COSMO Consortium (COSMO-LEPS, Marsigli et al., 2005) and from ECMWF (ECMWF-ENS, Miller et al., 2010). The version of the evaluation framework presented here is based on the performance of the ECMWF-ENS forecasts only, though it is foreseen to extend it to include the other model simulations. The system setup and additional details on how weather forecasts are handled in EFAS are documented in the published literature (Bartholmes et al., 2009; Pappenberger et al., 2010; Thielen et al., 2009), therefore we refer the reader to these articles for additional information not included in the present work, and focus on the analysis of the evaluation framework.

*Figure 1: Schematic view of the EFAS hydro-meteorological forecasting system.*

## 2.2 Meteorological data

ECMWF-ENS is a 51-member ensemble forecast run twice per day, at 00 UTC and 12 UTC as part of the operational production suite of ECMWF Integrated Forecast System (IFS, see Bechtold et al., 2014; Miller et al., 2010). ENS forecasts are run globally at T639 spectral resolution, corresponding to about 32 km horizontal resolution, with forecast lead time (LT) up to 10 days. After day 10, the model run is extended up to day 15 (day 32 twice per week) at a coarser horizontal resolution of about 65 km. Currently, EFAS uses only the first 10 days of forecast as input to the hydrological model. For this work, ENS forecasts from January 2009 to the present were extracted and used in the hydrological simulations, considering those available at the time of the forecasts (i.e., no reforecast with more recent IFS versions was used). Meteorological forecast fields used are total precipitation, evaporation, and 2-metre temperature, which are regridded to the same spatial resolution of the hydrological model (see next section).

A database of observed meteorological fields for Europe was provided by the Joint Research Centre of the European Commission. It consists of maps of spatially interpolated point measurements of precipitation and temperature at the surface level. The database includes daily data from the 1990 to the present, and it is populated by an increasing number of reporting gauges over time, with the latest figures showing on average more than 6000 stations for precipitation and more than 4000 for temperature (see Figure 2 for a recent example of daily data). A subset of the same meteorological station network is used to generate interpolated potential evapotranspiration maps using the Penman-Monteith method.

*Figure 2: stations reporting observed precipitation (left) and average temperature (right) on the 1st October 2013.*

## 2.3 Hydrological modelling

In EFAS, hydrological simulations are performed with Lisflood, a hybrid between a conceptual and a physical rainfall–runoff distributed model, designed to reproduce the main hydrological processes of medium to large river basins (see van der Knijff et al., 2010). The considered model setup for Europe was calibrated at 481 river gauges, using the observed meteorological fields as input and up to 7 years of gauged discharge. A reference hydrological simulation starting in 1990 was run for the European window with the calibrated Lisflood model at 5x5 km resolution, using the observed meteorological fields as input. The operational model is updated daily using the initial states of the previous day and the most recent meteorological observations acquired with about 1 day lag. This simulation, hereafter referred to as EFAS Water Balance (EFAS-WB), represents our best estimate of the hydrological states in the European rivers. The EFAS-WB is used in EFAS with regard to three main aspects (see Figure 1): *I*) deriving climatological features of the runoff in each point of the river network (e.g., average conditions, extremes, alert thresholds, seasonality); *II*) creating initial conditions for daily hydrological runs driven by the latest weather predictions; *III*) providing a reference simulation which is as realistic as possible, to be used as a proxy to evaluate streamflow forecasts in every grid point of the simulation domain. Further details on the EFAS-WB are described by Alfieri et al. (2013b). The same calibrated Lisflood setup is used to perform 10-day EFAS streamflow forecasts updated twice per day, by forcing the hydrological model with initial conditions from the EFAS-WB and with forecast weather fields (described in the previous section) with 1-day temporal resolution.

## 3 Evaluation strategy

EFAS forecasts are run at the ECMWF twice per day since October 2012, using weather predictions initialized at 00 and 12 UTC. This operational dataset of hydrological forecasts was complemented by running 4 years of daily hindcasts with the same model configuration, starting on January 2009. To reduce the computing load, the hindcasts were run only once per day, using forecast runs from 12 UTC. Ensemble streamflow predictions (ESP) are validated against the EFAS-WB for each point of the modelled European river network, comprising 38452 grid points. Such an approach enables a quick spatial overview of skill scores on every region of the computation domain, rather than just at stations where observed discharge is provided. On the other hand it does not account for the potential mismatch between actual river discharge and the simulated EFAS-WB used as reference.

Average scores are calculated over 1-year time windows. This choice proved to be effective as it includes one full hydrological year and dampens the seasonal variability of skill scores. In practice, the verification of dry months leads to higher scores than those of rainy months, as the quantitative forecast of high precipitation amounts is more challenging than forecasting days with zero precipitation. As a result, the evaluation framework was set up to select the first day of each month and calculate the average skill scores of the previous 365 days, starting on the 1st January 2010. The procedure was then semi-automated and skill scores are now updated every month to include results of the latest forecasts.

Skill scores to evaluate the ESP were chosen so that grid points with different upstream area and climatic regime could be compared together in the same graphs and in the same maps. To this end, four different dimensionless skill scores were selected, able to stress different aspects of the forecast performance. These are described in the following sub-sections and summarized in Table 1.

### 3.1 Nash-Sutcliffe efficiency

The Nash-Sutcliffe efficiency (NS, Nash and Sutcliffe, 1970) applied to discharge forecasting can be defined as:

$$\boldsymbol{NS = 1 - \frac{\sum_{t=1}^{N}\left[q_{sim}(t) - q_{fc}(t)\right]^2}{\sum_{t=1}^{N}\left[q_{sim}(t) - \bar{q}_{sim}\right]^2}}\ , \qquad (1)$$

where $q_{sim}$ is the proxy discharge given by the EFAS-WB and $q_{fc}$ is the forecast discharge at the same time step. $t$ is a time index spanning all $N$ forecasts included in the evaluation window, that is $N$=730 in operational forecasts (when two forecasts per day are evaluated) and $N$=365 for hindcasts between 2009 and 2012. In the case of the considered ESP, $q_{fc}$ represents the mean of the 51-member ensemble. The NS values range from -∞ to 1, the latter corresponding to perfect forecasts. NS above 0 means that forecasts perform better than climatological values, in the form of their average discharge $\bar{q}_{sim}$. In the presented work, NS values are calculated for fixed forecast lead times between 1 and 10 days, and the average values over 1 year windows are shown, as described in the previous section.

### 3.2 Forecast bias

Monitoring the bias of ensemble streamflow predictions is of vital importance for a flood awareness system based on a threshold exceedance approach as in EFAS. Flood alerts are detected by comparing EFAS simulations driven by weather forecasts as input, against reference warning thresholds, derived from the EFAS-WB. If weather forecasts were persistently different from observed meteorological values, discharge forecasts would be consequently biased, which may result in statistically significant over- or under-prediction of flood alerts. The main potential source of bias in ESP is the quantitative forecast of precipitation, particularly for high flow events. However, biased forecast values of temperature may induce cyclical drifts of discharge predictions, particularly in hydrological regimes where the snow accumulation and melting processes play a prominent role. In addition, precipitation, temperature and evapo-transpiration are key drivers for the soil moisture state, therefore consistent bias in their forecast values can affect the streamflow potentially over long ranges (i.e., monthly to inter-annual time scales). In the presented evaluation framework, the bias at each grid point is rescaled by the corresponding average discharge for the same period, calculated from the EFAS-WB:

$$\mathbf{Pbias} = \frac{\frac{1}{N}\sum_{t=1}^{N}\left[q_{sim}(t)-q_{fc}(t)\right]}{\bar{q}_{sim}} \qquad (2)$$

Being a linear operator, the sum of the percentage bias (Pbias) of all ensemble members is equal to the percentage bias of the ensemble mean.

### 3.3 Coefficient of variation of the RMSE

The Root Mean Squared Error (RMSE) has long been used to assess the magnitude of the error of deterministic forecasts. It has the advantage that it retains the units of the forecast variable and it includes the effect of both bias and variance of estimation. In addition, the RMSE depends on a quadratic function of the estimation residuals. This lead to some peculiarities, among which: 1) it is highly affected by few large errors and 2) it is often used as an error function to be minimized in a wide range of calibration and optimization processes. On the other hand it is difficult to compare RMSE values among different river stations, as their climatological discharge values may be substantially different. One option to compare the RMSE at different locations is to rescale it by the corresponding average discharge, as shown in Reed et al. (2007), so that resulting values become dimensionless:

$$CV = \frac{\sqrt{\frac{\sum_{t=1}^{N}\left[q_{sim}(t)-q_{fc}(t)\right]^{2}}{N}}}{\bar{q}_{sim}}, \qquad (3)$$

The resulting score is commonly referred to as coefficient of variation (CV) of the RMSE and, as for the RMSE, values close to zero are preferable. Also, when CV values are close to 1 it means that the RMSE of estimation is of the same order as the average discharge. Indeed, it can be associated to an inverse of the signal-to-noise ratio. By definition the CV penalizes river reaches with low average discharge compared to its variability, therefore higher CV values are expected in small or flash-flood prone river basins, such as those along the Mediterranean coast, where the predictability is indeed shorter than in large river basins.

### 3.4 Continuous Ranked Probability Skill Score

To fully exploit and assess the added value of probabilistic predictions, the Continuous Ranked Probability Skill Score (CRPSS) is used to evaluate the quantitative skills of the ESP. The CRPSS (e.g., Hersbach, 2000) is defined as:

$$CRPSS = \frac{\overline{CRPS_{ref}} - \overline{CRPS_{forecast}}}{\overline{CRPS_{ref}}}, \qquad (4)$$

where

$$CRPS = \int_{-\infty}^{\infty} [F(y) - F_0(y)]^2 dy \qquad (5)$$

and

$$F_0(y) = \begin{cases} 0, & y < observed\ value \\ 1, & y \geq observed\ value \end{cases} \qquad (6)$$

while *F(y)* is the stepwise cumulative distribution function (cdf) of the ESP of each considered forecast. The CRPSS is a dimensionless indicator of the skill of ensemble predictions, measured by (*CRPS*$_{forecast}$), compared to that of a reference forecast (*CRPS*$_{ref}$). The CRPSS ranges between 1 (for perfect predictions) to -∞, though ESP are valuable only when CRPSS>0, i.e., when the forecasts perform better than the reference. In this work, we compare and discuss the use of two different *CRPS*$_{ref}$ to evaluate the CRPSS, the first based on the average climatological discharge $\bar{q}_{sim}$ (*CRPS*$_{ref,ad}$), and the second based on a persistence forecast (*CRPS*$_{ref,pf}$), meaning a forecast given by assuming the same value used to initialize the ESP. It is worth noting that both reference CRPS are based on deterministic predictions, hence the *CRPS*$_{ref}$ reduces to the mean absolute error (Hersbach, 2000):

$$CRPS_{ref,ad} = \frac{1}{M}\sum_{t=1}^{M}|q_{sim}(t) - \bar{q}_{sim}| \qquad (7)$$

where *t* is a daily time index going from 1/1/1990 to the present. On the other hand,

$$CRPS_{ref,pf}(LT) = \frac{1}{N}\sum_{t=1}^{N}|q_{sim}(t) - q_{sim}(t - LT)| \qquad (8)$$

where *N* has the same meaning as in Eq.1.

Two significant differences between Eq.7 and 8 can be seen. The *CRPS*$_{ref,ad}$ is a constant value and only depends on the location, though it needs climatological information to be evaluated, in the form of a reference time series of observations or proxy simulations (i.e., the EFAS-WB in this case). On the other hand, the *CRPS*$_{ref,pf}$ depends on the lead time of the forecast (LT). It does not need any prior climatological information on the discharge regime at the point but the discharge value used to initialize the forecast.

## 4 Results

Skill scores of the last available year are now routinely calculated on the 13th day of each month, after all meteorological observations to update the EFAS-WB are received and the hydrological model is run. Simulated proxy discharges need to be computed until the 11th of the same month, so that 10-day ESP starting on the 1st can be evaluated. Scores described in Sect. 3 are shown in **Figure 3**. NS, CV and Pbias are deterministic scores; hence they are calculated on the ensemble mean, while the CRPSS take into account the whole ensemble. A forecast lead time of 5 days is chosen for most figures in the article, being representative of the general behaviour of the ESP and a frequent lead time of EFAS flood alerts. One can see that, for LT=5 days, in the vast majority of grid points the ESP is more skilful than a persistence forecast (i.e., $CRPSS_{pf}>0$). The NS and the CV suggest that higher performance is achieved in large rivers of Central and Northern Europe. Excluding Iceland, lower skills are mostly seen in Southern Europe and can be explained by a) resolution issues in small basins, b) less skilful precipitation forecast in mountainous areas, c) a comparatively lower station density to run the EFAS-WB, and d) the higher proportion of convective precipitation, leading to higher space-time variability of rainfall rates and larger extremes over short (i.e., 1-day or sub-daily) durations. Similarly, the Pbias (on gray background in **Figure 3**) shows a widespread underestimation of discharge over the main mountain ranges (i.e., Pyrenees, Alps and Balkans, among others), mostly in the range 10% to 50% of the corresponding average flow. These findings are in line with previous works by Wittmann et al. (2010) and Pappenberger et al. (2013), who showed increasing underestimation of precipitation and streamflow forecast in the Alpine region during intense precipitation events. The apparently poor performance over Iceland in **Figure 3** is actually imputable to an incorrect reference streamflow. Indeed, the number of reporting stations for this region is very low (see an example in **Figure 2**), particularly for precipitation, thus leading to a considerable under-prediction of the streamflow. In other words, although EFAS streamflow forecasts over Iceland may be skilful, the current availability of meteorological observations prevents from simulating reliable reference discharge to perform forecast evaluation in this area. In the following analyses, summary scores of grid points in Iceland are excluded from all figures, which brings the dataset to a subset of 37588 points.

*Figure 3: $CRPSS_{pf}$, CV, NS and Pbias over Europe for 1 year of daily forecasts ending on the 1st October 2013 (5-day lead time).*

### 4.1 Performance versus forecast range

Skill scores as in **Figure 3** are shown in **Figure 4** for each forecast lead time between 1 and 10 days. A solid line indicates the mean value among all grid points, while grey shades denote the 5%-95% (light grey) and the 25%-75% (dark grey) of their distribution. In the top-left panel, the CRPSS calculated using the average discharge as reference (i.e., $\text{CRPSS}_{\text{ad}}$) is shown with a thick dashed line (mean value) together with the corresponding 25%-75% values (dotted lines). Differences between the two methods are the largest for the first lead time, where in many cases the ESP does not bring substantial differences in comparison to a persistence forecast, due to the large weight of the initial model states. On the other hand, the $\text{CRPSS}_{\text{ad}}$ decreases roughly linearly and suggests the presence of a crossing point for a LT>10 days, when the climatological average discharge seems to become a more skillful benchmark than a persistence forecast. As expected, the CV tends to deteriorate with the lead time, though without a significant increase of the spread of its distribution. Similarly, the mean NS ranges between 0.9 for LT=1 and 0.7 at the end of the forecast range, while in 99% of forecasts NS>0 for LT=10 days. The Pbias shows a rather constant mean under-prediction of 2% to 4%. Its distribution has an increasing spread with the lead time, with 65% to 70% of grid points lying constantly below the zero line.

*Figure 4: CRPSS, CV, NS and Pbias of ESP versus the forecast lead time.*

### 4.2 Performance versus catchment size

**Figure 5** displays the four scores against the upstream area of each grid point, calculated over 1 year ending on 1/10/2013 and for a 5-day lead time. In addition, solid lines indicate the empirical median value (i.e., 50th percentile), in light grey, and the central 90% of the distribution (i.e., 5th to 95th percentiles), in dark grey. Largest values on the x-axis correspond to the lower Danube River, with upstream area up to about 800,000 km$^2$. On the left side of each panel, one can note the model grid resolution as limit, with catchments area being always a multiple of 25 km$^2$. Results in **Figure 5** denote a general positive trend of skill scores with increasing upstream area. Indeed, in large rivers, a) the discharge varies more gradually due to the smoothing and averaging effect of the complex river network and b) the influence of the initial discharge, compared to the forecast precipitation input, is larger than in smaller catchments. In detail, as the basin time of concentration increases and approaches the magnitude of the forecast range, a larger proportion of the forecast discharge at the river outlet is made up by a water volume which is already in the model, (i.e., gauged) at the starting time of the forecast run. Therefore the skill of weather forecasts affects that of streamflow forecasts with an average delay increasing with the upstream area, which can be in the order of some days for large European rivers. On the other hand, **Figure 5** shows a clear deterioration of scores for catchments smaller than 300 km$^2$, that is, for a ratio between upstream area and grid size of the weather forecasts of about 0.3. Results are in agreement with those of Pappenberger et al. (2010), though Bartholmes et al. (2006) suggested a minimum threshold of 4000 km$^2$ if extreme values are considered. Indeed, the latter value is used in EFAS as minimum upstream area for flood alerts to be issued to partner institutes. The median value of the Pbias in **Figure 5** indicates that the deterioration of scores can be partly attributed to the underestimation of the discharge for small catchments, which decreases below 2%, in absolute value, for upstream areas larger than 400 km$^2$. As commented in Sect. 4, such trend is to be attributed to the under-prediction of quantitative precipitation in mountain areas and of extreme values in general, not fully captured by the atmospheric circulation model due to its grid size on average coarser than the observation network.

*Figure 5: $CRPSS_{pf}$, CV, NS and Pbias of ESP versus the upstream area of each river grid point.*

### 4.3 Evolution of 12-month average performance

The evolution of summary scores over the past 5 years is shown in **Figure 6**. Scores are calculated on the 365 days preceding the first day of each month indicated in the x axis. In the top-left panel both $CRPSS_{pf}$ and $CRPSS_{ad}$ are shown, using the same line types as in **Figure 4**. In addition, the average discharge over all grid points of the river network, for each evaluation period, is drawn at the bottom. One can note how the $CRPSS_{ad}$ is largely affected by the magnitude of the observed runoff, so that, in drier years, it gives the impression of increasing forecast performance, and vice-versa. In the $CRPSS_{pf}$, no dependence on the average runoff is visible. The latter shows an improvement of the forecast skills during the year 2013, particularly for the mean of the distribution and for the 75th and 95th quantiles. Such improvement is also pointed out by a reduced mean CV and increased mean NS, where in both cases the central 90% of the distribution becomes narrower since the beginning of 2013, though with a subsequent widening towards the end of the year.

Interestingly, the bottom-right panel denotes a slow but constant increase of a negative bias in forecast streamflow over the last years. This appears consistently on all lead times (not shown), though it is more significant towards the end of the forecast range. On the other hand, no corresponding trend was reported in the forecast input precipitation produced by the ECMWF-ENS (personal communication, see some additional details in http://www.ecmwf.int/products/forecasts/d/charts/medium/verification), nor in temperature (possibly inducing a larger snow fraction). Instead, the main reason for such discrepancy is most likely due to the progressive increase in the number of stations reporting meteorological observations in recent years. Higher station density leads to a more realistic representation of the input maps to run the EFAS-WB, so that small-scale features such as convective cells are more likely to be better observed quantitatively. In this regard, Kann and Haiden (2005) showed that when high density stations networks are used as reference, the mean absolute error of forecast precipitation tend to increase with the reduction of the aggregation area. Further, some of the stations added recently are located in elevated areas, such as in the Alps and the Pyrenees, where the orography enhances annual rainfall totals and consequently the runoff. Indeed, these areas are where the under-prediction of discharges has become clearer in the recent years, as shown in **Figure 3**.

*Figure 6: Trend of 12-month average CRPSS, CV, NS and Pbias of ESP from 2009 onwards.*

## 5 Discussion and conclusions

This article presents the current status of the evaluation framework used to monitor and update regularly the forecast performance of the European Flood Awareness System. Results suggest that streamflow forecasts driven by weather predictions provide significant added value to the monitoring of the main European rivers. As expected, performance decreases with lead time, though it remains skilful for the whole 10-day range, in comparison to the use of climatological or persistence forecasts. In large river basins of Europe, the average time lag between weather forcing and runoff is on the order of some days. Hence the real-time hydrological simulation run with meteorological observations gives a significant proportion of the overall predictability, increasing with the basin time of concentration. In smaller river basins, the effect of initial conditions is less important, therefore the predictability is shorter as it mostly depends on that of the weather forecasts. In river basins of size below 300-400 km$^2$ forecast skill becomes poorer. Their forecasts show large variability, often even for 1-day lead time, and significant underestimation of the runoff in mountain regions.

Being designed on dimensionless scores, the main strength of the proposed verification system is in highlighting relative changes of performance, which can be detected over different regions, forecast lead time, basin size and, most importantly, in time. An evaluation of 12-month average scores over the past 5 years suggests a moderate improvement for all 12-month forecasts ending from the beginning of 2013 onwards. Such improvement occurred notwithstanding an increasing negative forecast bias, especially in mountain regions. This can be attributed to a progressive increase of the meteorological stations used to run the EFAS-WB, which in turn has improved the representation of the runoff dynamics in the presence of pronounced orography. Although the parameterization of the hydrological model was subject to changes and improvements every 1 to 1.5 years on average, the 5 year simulation shown in this study was carried out with a fixed model version, corresponding to the current operational one at the time of writing. Therefore, the positive trend of performance shown in **Figure 6** is likely to underestimate the real improvements which have occurred and rather reflect that of weather forecasts used as input.

## 5.1 The benchmark of skill scores

The four performance scores presented in the article can be classified into two categories, depending on whether the comparison is carried out against a benchmark or not. On the one hand, the CV and the Pbias give a measure of the RMSE and of the bias of forecasts, respectively. RMSE and bias are commonly used in verification because of their physical meaning, as they quantify the error with the same units of the forecast variable. They are rescaled by the average flow to make them comparable over different regions and along the river network. On the other hand, the NS and the CRPSS give a relative performance in comparison to an alternative benchmark forecast. Literature works show a surprising variety of different benchmarks used for comparison (see Pappenberger et al., submitted, for a recent review), sometimes without motivating the choice. Here we argue that, in assessing the predictability of a forecasting system, the benchmark should represent a realistic forecast achievable in case the system was not in place. The use of persistence forecasts is hereby suggested as a suitable benchmark, in that it does not require climatological information of the runoff at the river point, nor additional model runs. In comparison to a benchmark based on the average discharge, persistence acknowledges the role of initial conditions, indicating that the highest value of forecasts corresponds to a balance between the ability to provide accurate forecasts and the ability to detect deviations from an initial state (see *CRPSS<sub>ad</sub>* versus *CRPSS<sub>pf</sub>* in **Figure 4**). Further, persistence is independent of seasonal variations or trends in the mean value of the forecast variable, as discussed in Sect. 4.3.

It is worth noting that the same principle can be applied to the Nash-Sutcliffe efficiency, as suggested by Plate and Lindenmaier (2008), leading to a modified formulation which uses a persistence forecast as reference value:

$$\boldsymbol{NS(LT) = 1 - \frac{\sum_{t=1}^{N}\left[q_{sim}(t) - q_{fc}(t)\right]^2}{\sum_{t=1}^{N}\left[q_{sim}(t) - q_{sim}(t-LT)\right]^2}} \,. \qquad (9)$$

This formulation was not tested in the present framework, though it may be a valid alternative to the NS for large river basins (see e.g., Pagano, 2013). Its application will be considered for future system developments.

## 5.2 The EFAS-WB as reference simulations

The main assumption of the presented approach is that the EFAS-WB can be used as a realistic representation of the actual runoff. On the other hand the use of the output of a distributed model as the EFAS-WB allows a performance evaluation over the full computation domain. Moreover, the continuous increase in the number of reporting stations, both for meteorological and hydrological data, is progressively pushing the EFAS-WB closer to the real streamflow conditions in the European rivers. This occurs thanks to a better reproduction of the meteorological input data and to the increase of the number of river stations where the parameters of the hydrological model can be calibrated. Recent advances in the meteorological dataset include the addition of more than 10 high density national networks and an improved approach to interpolating point values into spatial maps (see Ntegeka et al., 2013). This is currently being tested and will be used in the next version of EFAS, together with additional historical observed streamflow at a number of river gauges to improve the model calibration. Similarly, resulting simulated discharges of the EFAS-WB can potentially become a dataset to validate and benchmark a wide range of hydrological models, particularly on large scales. Current main limitations of simulated discharges are at the lower end of the range of the space-time scale of simulated catchments. In fact, the current daily time aggregation of input data induces a smoothing of output discharges, so that simulated extreme values have reported under-estimation issues, relatively to observed values. In addition, the presented scores are not able to capture potential errors in the hydrological model, because both ESP and the EFAS-WB used for validation are generated by the same model. However, this is evaluated separately at those stations where the model parameters are calibrated (see Feyen et al., 2007). Also, an assessment of the total predictive uncertainty is performed at river gauges (currently about 40) where discharge values are received in real-time. The methodology and results are described by Bogner and Pappenberger (2011).

### 5.3 Concluding remarks

In its current state, the evaluation framework has proved its usefulness in spotting strengths and weaknesses of ensemble forecasts used in EFAS, including trends of performance in time and size limits of river basins under monitoring. In addition, it has pointed out a number of key developments to focus on to improve the evaluation and the diagnostic of the forecasting system:

- Implementation of the evaluation framework to streamflow predictions derived from all the different numerical weather predictions used as input in EFAS, including DWD, COSMO-LEPS and products which are foreseen to be tested in the future.

- Enlarging the collection of near real time observed discharges for continuous monitoring of the skill scores of both the EFAS-WB and streamflow predictions against observed values.

- Comparison of performance scores for updated model versions. A new EFAS version was implemented in January 2014, which includes a more extensive calibration of the hydrological model and an enhanced dataset of meteorological observations.

- Complementing the current approach with skill scores targeted to evaluate the performance in forecasting extreme events, including threshold exceedance analyses.

- Set up a visualization platform on the web where performance can be monitored by developers, analysts on duty and users, to aid the monitoring of forecasts and the diagnostic of issues.

## References

Alfieri, L., Burek, P., Dutra, E., Krzeminski, B., Muraro, D., Thielen, J. and Pappenberger, F., 2013a. GloFAS – global ensemble streamflow forecasting and flood early warning, Hydrol. Earth Syst. Sci., 17, 1161–1175, doi:10.5194/hess-17-1161-2013,.

Alfieri, L., Salamon, P., Bianchi, A., Neal, J., Bates, P. and Feyen, L., 2013b. Advances in pan-European flood hazard mapping, Hydrol. Process., doi:10.1002/hyp.9947.

Alfieri, L., Salamon, P., Pappenberger, F., Wetterhall, F. and Thielen, J., 2012. Operational early warning systems for water-related hazards in Europe, Environ. Sci. Policy, 21, 35–49.

Arheimer, B., Lindström, G. and Olsson, J., 2011. A systematic review of sensitivities in the Swedish flood-forecasting system, Atmos. Res., 100, 275–284.

Bartholmes, J. C., Thielen, J., Ramos, M. H. and Gentilini, S., 2009. The European flood alert system EFAS - Part 2: Statistical skill assessment of probabilistic and deterministic operational forecasts, Hydrol. Earth Syst. Sc., 13, 141–153.

Bartholmes, J., Thielen, J. and Ramos, M. H., 2006. Quantitative analyses of EFAS forecasts using different verification (skill) scores, The benefit of probabilistic flood forecasting on European scale, edited by: Thielen, J., EUR, 22560, 58–79.

Bechtold, P., Semane, N., Lopez, P., Chaboureau, J.-P., Beljaars, A. and Bormann, N., 2014. Representing equilibrium and nonequilibrium convection in large-scale models, J. Atmos. Sci., 71, 734–753, doi:10.1175/JAS-D-13-0163.1.

Bogner, K. and Pappenberger, F., 2011. Multiscale error analysis, correction, and predictive uncertainty estimation in a flood forecasting system, Water Resour. Res., 47, W07524, doi:10.1029/2010WR009137.

Brown, J. D., Demargne, J., Seo, D.-J. and Liu, Y., 2010. The Ensemble Verification System (EVS): A software tool for verifying ensemble forecasts of hydrometeorological and hydrologic variables at discrete locations, Environ. Modell. Softw., 25, 854–872, doi:10.1016/j.envsoft.2010.01.009.

Cloke, H. L. and Pappenberger, F., 2009. Ensemble flood forecasting: A review, J. Hydrol., 375, 613–626.

Demargne, J., Brown, J., Liu, Y., Seo, D.-J., Wu, L., Toth, Z. and Zhu, Y., 2010. Diagnostic verification of hydrometeorological and hydrologic ensembles, Atmos. Sci. Lett., 11, 114–122.

Feyen, L., Vrugt, J. A., Nualláin, B., van der Knijff, J. and De Roo, A., 2007. Parameter optimisation and uncertainty assessment for large-scale streamflow simulation with the LISFLOOD model, J. Hydrol., 332, 276–289, doi:10.1016/j.jhydrol.2006.07.004.

Gourley, J. J., Erlingis, J. M., Hong, Y. and Wells, E. B., 2012. Evaluation of Tools Used for Monitoring and Forecasting Flash Floods in the United States, Weather Forecast., 27, 158–173, doi:10.1175/WAF-D-10-05043.1.

Hersbach, H., 2000. Decomposition of the continuous ranked probability score for ensemble prediction systems, Weather Forecast., 15, 559–570.

Kann, A. and Haiden, T., 2005. The August 2002 flood in Austria: sensitivity of precipitation forecast skill to areal and temporal averaging, Meteorol. Z., 14, 369–377, doi:10.1127/0941-2948/2005/0042.

Van der Knijff, J. M., Younis, J. and de Roo, A. P. J., 2010. LISFLOOD: A GIS-based distributed model for river basin scale water balance and flood simulation, Int. J. Geogr. Inf. Sci., 24, 189–212.

Legates, D. R. and McCabe, G. J., 1999. Evaluating the use of "goodness-of-fit" Measures in hydrologic and hydroclimatic model validation, Water Resour. Res., 35, 233–241, doi:10.1029/1998WR900018.

Majewski, D., Liermann, D., Prohl, P., Ritter, B., Buchhold, M., Hanisch, T., Paul, G., Wergen, W. and Baumgardner, J., 2002. The operational global icosahedral-hexagonal gridpoint model GME: Description and high-resolution tests, Mon. Weather Rev., 130, 319–338.

Marsigli, C., Boccanera, F., Montani, A. and Paccagnella, T., 2005. The COSMO-LEPS mesoscale ensemble system: Validation of the methodology and verification, Nonlinear Proc. Geoph., 12, 527–536.

Miller, M., Buizza, R., Haseler, J., Hortal, M., Janssen, P. and Untch, A., 2010. Increased resolution in the ECMWF deterministic and ensemble prediction systems, ECMWF Newsletter, 124, 10–16.

Nash, J. E. and Sutcliffe, J. V., 1970. River flow forecasting through conceptual models part I–A discussion of principles, J. Hydrol., 10, 282–290.

Ntegeka, V., Salamon, P., Gomes, G., Sint, H., Lorini, V. and Thielen, J., 2013. A high-resolution European dataset for hydrologic modeling, in EGU General Assembly Conference Abstracts, 15, 7460.

Pagano, T. C., 2013. Evaluation of Mekong River Commission operational flood forecasts, 2000–2012, Hydrol. Earth Syst. Sci. Discuss., 10, 14433–14461, doi:10.5194/hessd-10-14433-2013.

Pappenberger, F., Ramos, M. H., Cloke, H. L., Wetterhall, F., Alfieri, L., Bogner, K., Mueller, A., and Salamon, P., submitted. How do I know if my forecasts are better? Using benchmarks in Hydrological Ensemble Predictions, J. Hydrol.

Pappenberger, F., Thielen, J. and Del Medico, M., 2010. The impact of weather forecast improvements on large scale hydrology: analysing a decade of forecasts of the European Flood Alert System, Hydrol. Process., 25, 1091–1113, doi:10.1002/hyp.7772.

Pappenberger, F., Wetterhall, F., Albergel, C., Alfieri, L., Balsamo, G., Bogner, K., Haiden, T., Hewson, T., Magnusson, L., de Rosnay, P., Muñoz Sabater, J. and Tsonevsky, I., 2013. Floods in Central Europe in June 2013, ECMWF Newsletter, 136, 9–11.

Plate, E. J. and Lindenmaier, F., 2008. Quality assessment of forecasts, in Mekong River Commission: Sixth Annual Flood Forum, Phnom Penh, May, 10 pp.

Randrianasolo, A., Ramos, M. H., Thirel, G., Andreassian, V. and Martin, E., 2010. Comparing the scores of hydrological ensemble forecasts issued by two different hydrological models, Atmos. Sci. Lett., 11, 100–107, doi:10.1002/asl.259.

Reed, S., Schaake, J. and Zhang, Z., 2007. A distributed hydrologic model and threshold frequency-based method for flash flood forecasting at ungauged locations, J. Hydrol., 337, 402–420.

Renner, M., Werner, M. G. F., Rademacher, S. and Sprokkereef, E., 2009. Verification of ensemble flow forecasts for the River Rhine, J. Hydrol., 376, 463–475.

Steppeler, J., Doms, G., Scḧattler, U., Bitzer, H. W., Gassmann, A., Damrath, U. and Gregoric, G., 2003. Meso-gamma scale forecasts using the nonhydrostatic model LM, Meteorol. Atmos. Phys., 82, 75–96.

Thielen, J., Bartholmes, J., Ramos, M.-H. and De Roo, A., 2009. The European flood alert system - part 1: Concept and development, Hydrol. Earth Syst. Sc., 13, 125–140.

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

Trinh, B. N., Thielen-del Pozo, J. and Thirel, G., 2013. The reduction continuous rank probability score for evaluating discharge forecasts from hydrological ensemble prediction systems, Atmos. Sci. Lett., 14, 61–65, doi:10.1002/asl2.417.

Verkade, J. S., Brown, J. D., Reggiani, P. and Weerts, A. H., 2013. Post-processing ECMWF precipitation and temperature ensemble reforecasts for operational hydrologic forecasting at various spatial scales, J. Hydrol., 501, 73–91, doi:10.1016/j.jhydrol.2013.07.039.

Wetterhall, F., Pappenberger, F., Alfieri, L., Cloke, H. L., Thielen-del Pozo, J., Balabanova, S., Daňhelka, J., Vogelbacher, A., Salamon, P., Carrasco, I., Cabrera-Tordera, A. J., Corzo-Toscano, M., Garcia-Padilla, M., Garcia-Sanchez, R. J., Ardilouze, C., Jurela, S., Terek, B., Csik, A., Casey, J., Stankūnavičius, G., Ceres, V., Sprokkereef, E., Stam, J., Anghel, E., Vladikovic, D., Alionte Eklund, C., Hjerdt, N., Djerv, H., Holmberg, F., Nilsson, J., Nyström, K., Sušnik, M., Hazlinger, M. and Holubecka, M., 2013. HESS Opinions "Forecaster priorities for improving probabilistic flood forecasts," Hydrol. Earth Syst. Sci., 17, 4389–4399, doi:10.5194/hess-17-4389-2013.

Wilks, D. S., 2006. Statistical Methods in the Atmospheric Sciences: An Introduction, electronic version, Elsevier, San Diego, CA.

Wittmann, C., Haiden, T. and Kann, A., 2010. Evaluating multi-scale precipitation forecasts using high resolution analysis, Adv. Sci. Res., 4, 89–98.

Wood, A. W., Kumar, A. and Lettenmaier, D. P., 2005. A retrospective assessment of National Centers for Environmental Prediction climate model–based ensemble hydrologic forecasting in the western United States, J. Geophys. Res-Atmos., 110, doi:10.1029/2004JD004508.

## Tables

Table 1: Summary of performance scores and their information content.

| Score | Short name | Use |
|---|---|---|
| Nash-Sutcliffe efficiency | NS | Normalized measure of the mean squared error of the ensemble mean in comparison to a constant climatological mean |
| Percent bias | Pbias | Dimensionless measure of the forecast bias |
| Coefficient of variation of the Root Mean Squared Error | CV | Dimensionless measure of the Root Mean Squared Error of the ensemble mean |
| Continuous Ranked Probability Skill Score (average discharge as reference) | $\mathrm{CRPSS}_{\mathrm{ad}}$ | Skill score to compare the distribution of ensemble forecasts around observations, as opposed to using the climatological average discharge |
| Continuous Ranked Probability Skill Score (persistence forecast as reference) | $\mathrm{CRPSS}_{\mathrm{pf}}$ | Skill score to compare the distribution of ensemble forecasts around observations, as opposed to using the persistence of the initial discharge |

## Figure captions

Figure 1: Schematic view of the EFAS hydro-meteorological forecasting system.

Figure 2: stations reporting observed precipitation (left) and average temperature (right) on the 1st October 2013.

Figure 3: $\mathrm{CRPSS}_{pf}$, CV, NS and Pbias over Europe for 1 year of daily forecasts ending on the 1st October 2013 (5-day lead time).

Figure 4: CRPSS, CV, NS and Pbias of ESP versus the forecast lead time.

Figure 5: $\mathrm{CRPSS}_{pf}$, CV, NS and Pbias of ESP versus the upstream area of each river grid point.

Figure 6: Trend of 12-month average CRPSS, CV, NS and Pbias of ESP from 2009 onwards.

Figure 1

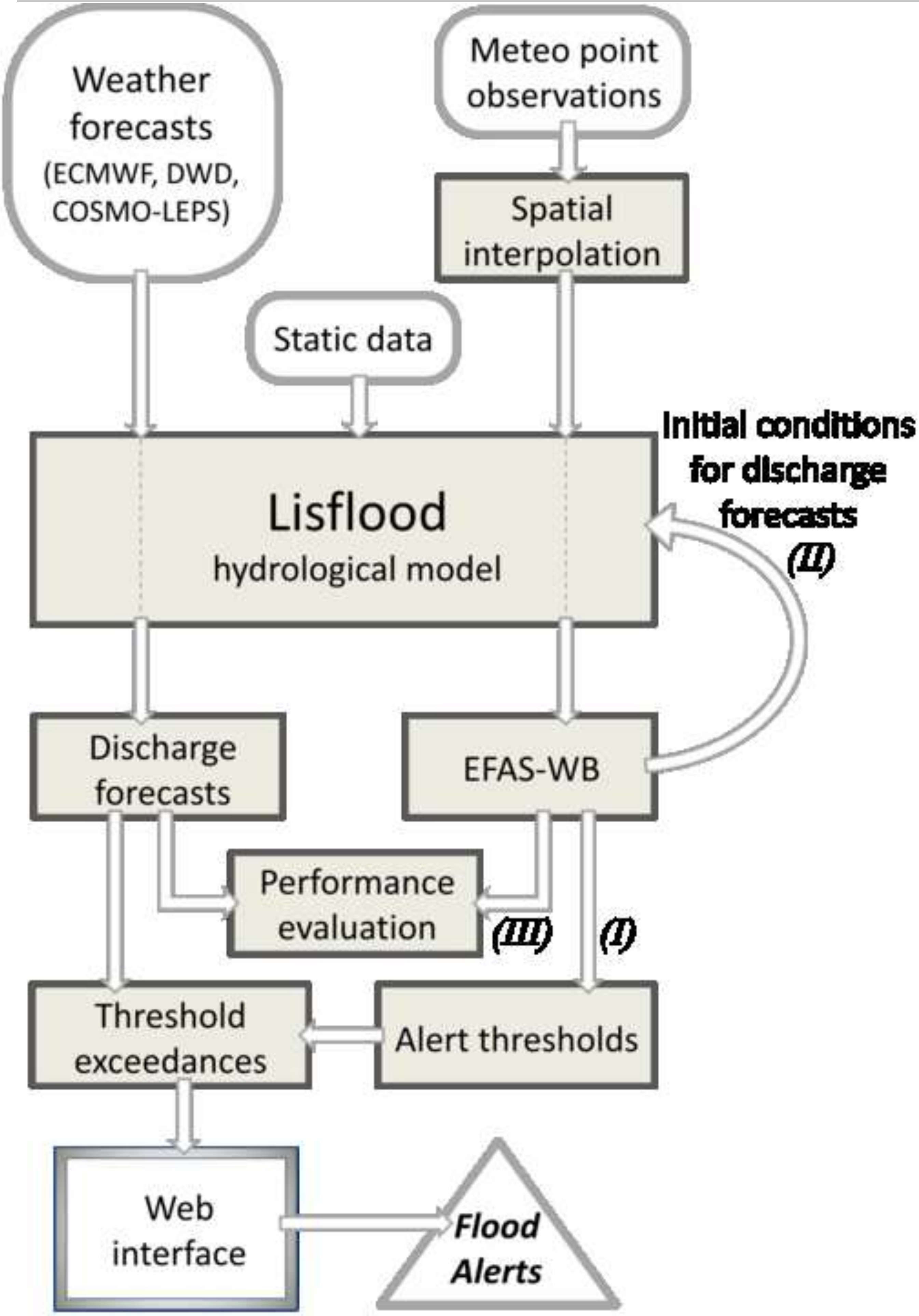

Weather forecasts (ECMWF, DWD, COSMO-LEPS)
Meteo point observations
Spatial interpolation
Static data
Lisflood
hydrological model
Initial conditions for discharge forecasts
(II)
Discharge forecasts
EFAS-WB
Performance evaluation
(III)
(I)
Threshold exceedances
Alert thresholds
Web interface
Flood Alerts

Figure 2

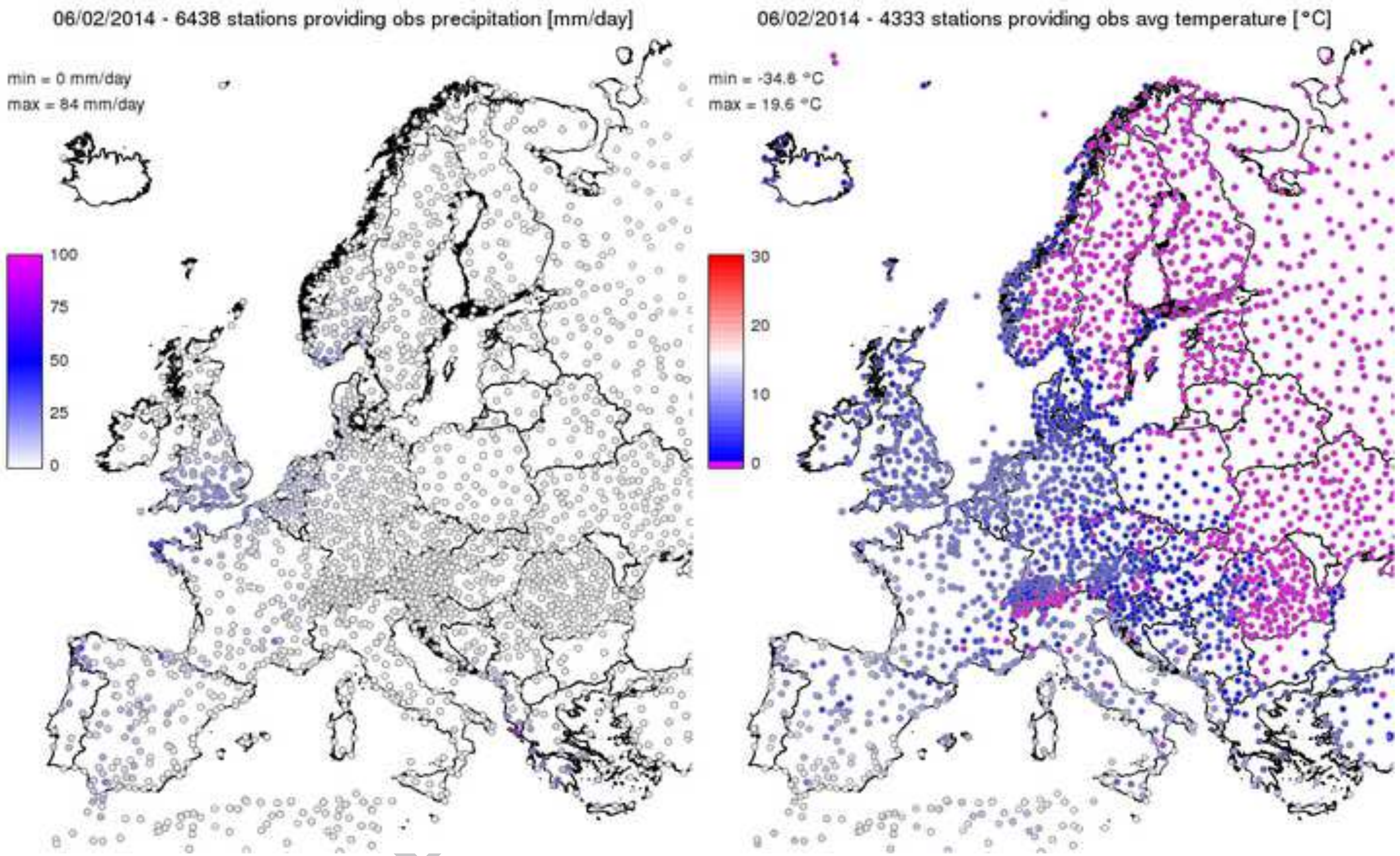

06/02/2014 - 6438 stations providing obs precipitation [mm/day]
min = 0 mm/day
max = 84 mm/day
100
75
50
25
0
06/02/2014 - 4333 stations providing obs avg temperature [°C]
min = -34.8 °C
max = 19.6 °C
30
20
10
0

Figure 3

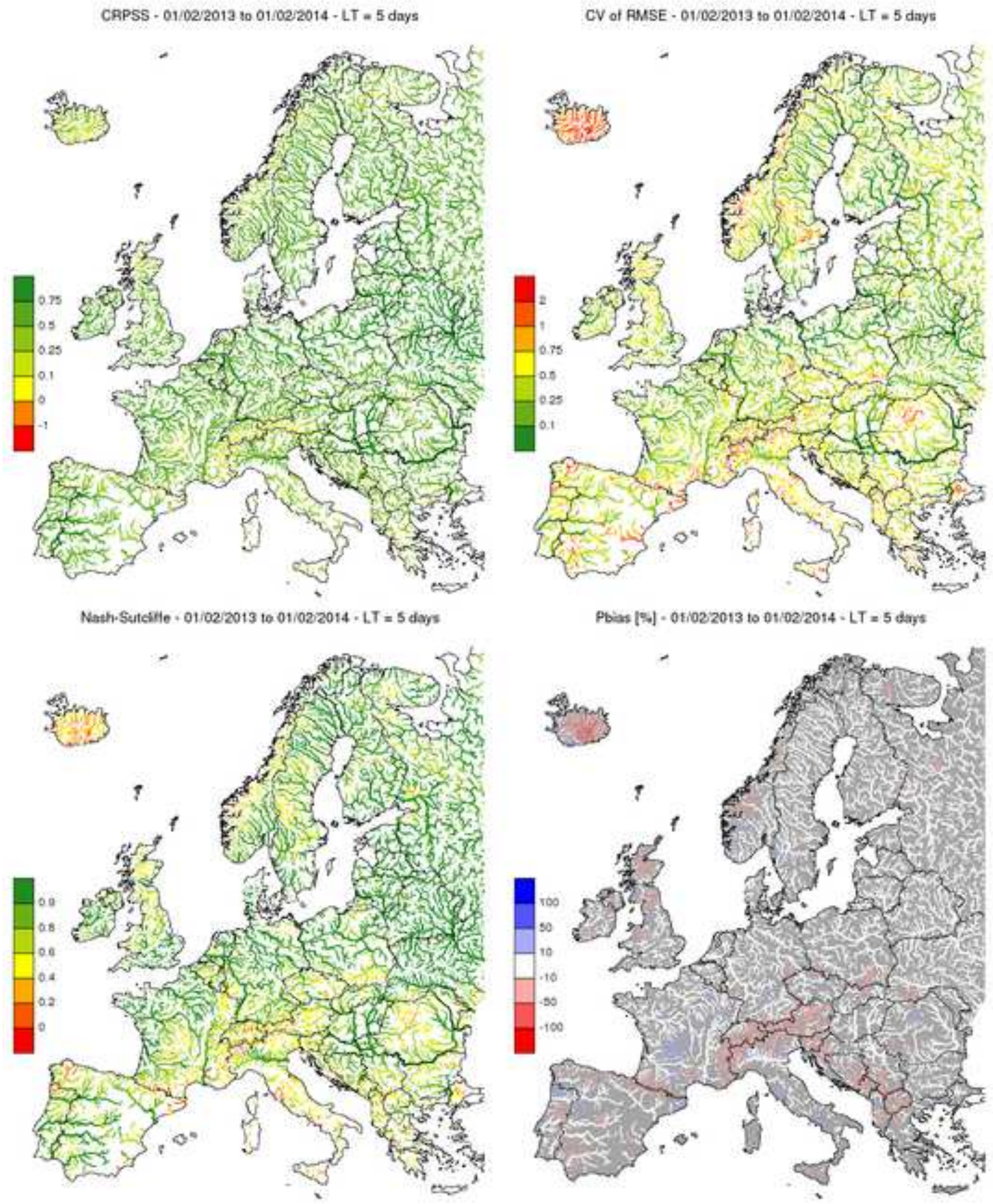

CRPSS - 01/02/2013 to 01/02/2014 - LT = 5 days
0.75
0.5
0.25
0.1
0
-1
CV of RMSE - 01/02/2013 to 01/02/2014 - LT = 5 days
2
1
0.75
0.5
0.25
0.1
Nash-Sutcliffe - 01/02/2013 to 01/02/2014 - LT = 5 days
0.9
0.8
0.6
0.4
0.2
0
Pbias [%] - 01/02/2013 to 01/02/2014 - LT = 5 days
100
50
10
-10
-50
-100

Figure 4

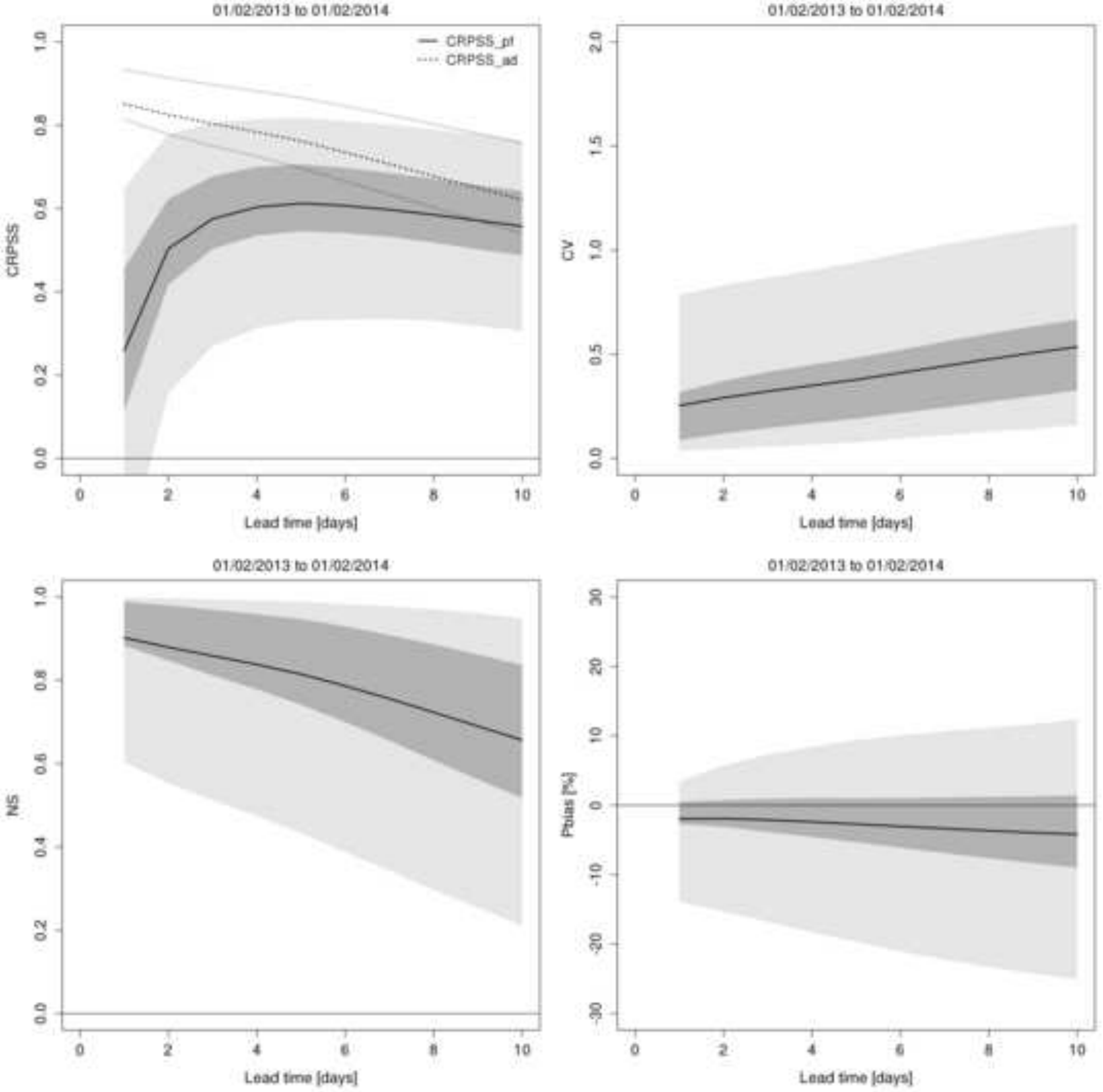

01/02/2013 to 01/02/2014
CRPSS_pf
CRPSS_ad
CRPSS
0.0
0.2
0.4
0.6
0.8
1.0
Lead time [days]
0
2
4
6
8
10
01/02/2013 to 01/02/2014
CV
0.0
0.5
1.0
1.5
2.0
Lead time [days]
01/02/2013 to 01/02/2014
NS
Lead time [days]
01/02/2013 to 01/02/2014
Pbias [%]
-30
-20
-10
0
10
20
30
Lead time [days]

Figure 5

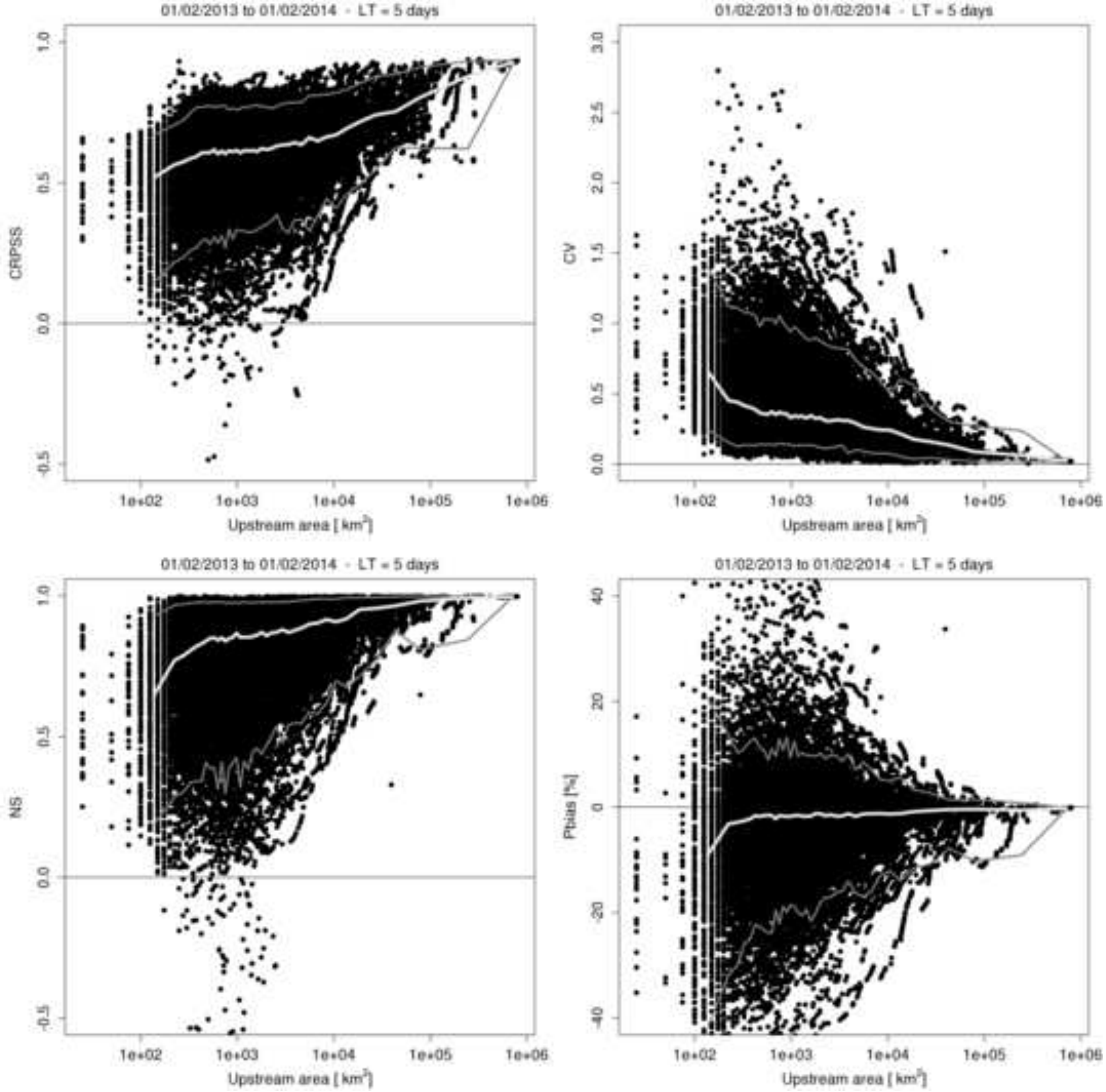
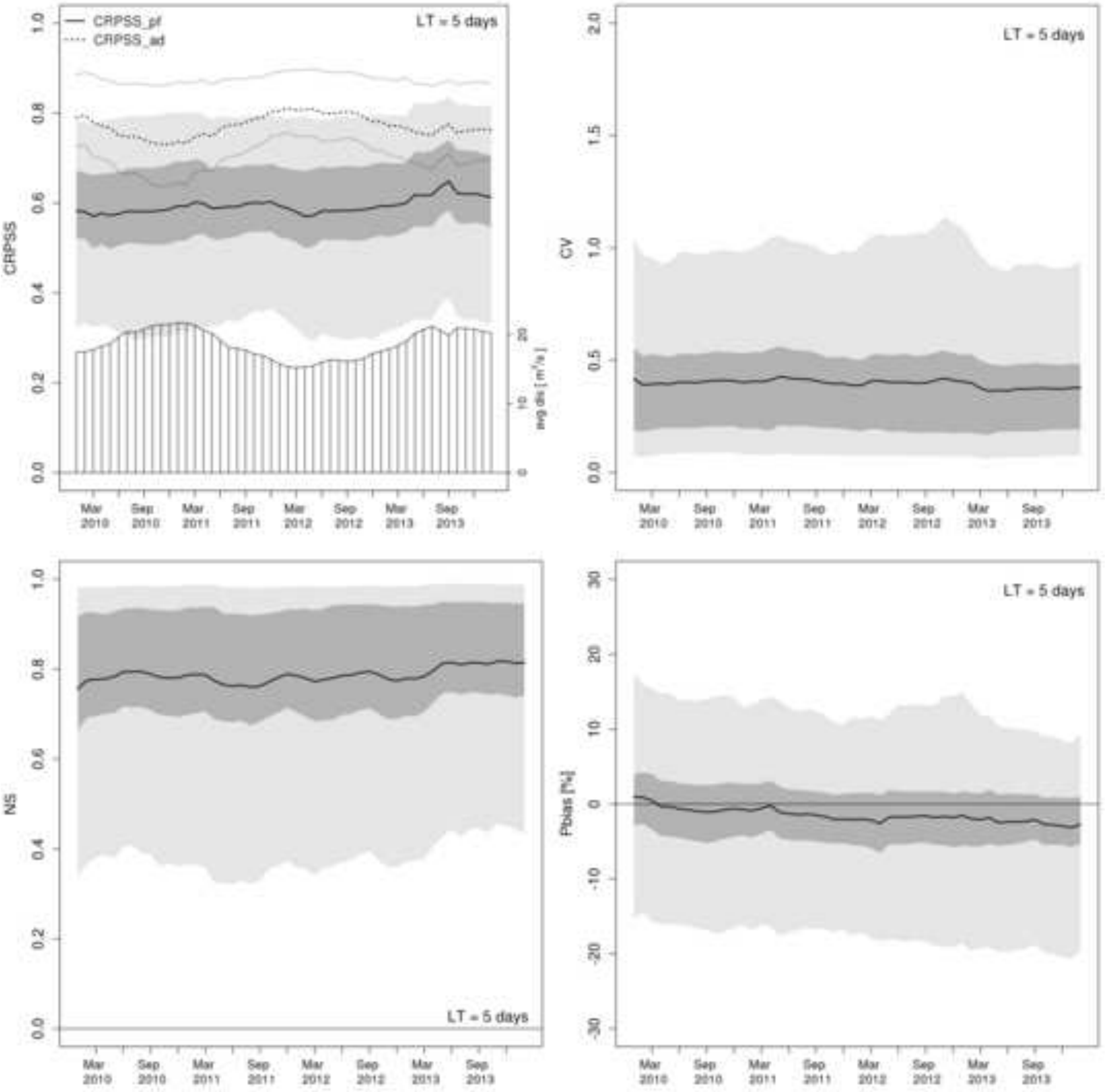

01/02/2013 to 01/02/2014 - LT = 5 days
CRPSS
1.0
0.5
0.0
-0.5
1e+02
1e+03
1e+04
1e+05
1e+06
Upstream area [ km$^2$]
01/02/2013 to 01/02/2014 - LT = 5 days
CV
3.0
2.5
2.0
1.5
1.0
0.5
0.0
1e+02
1e+03
1e+04
1e+05
1e+06
Upstream area [ km$^2$]
01/02/2013 to 01/02/2014 - LT = 5 days
NS
1.0
0.5
0.0
-0.5
1e+02
1e+03
1e+04
1e+05
1e+06
Upstream area [ km$^2$]
01/02/2013 to 01/02/2014 - LT = 5 days
Pbias [%]
40
20
0
-20
-40
1e+02
1e+03
1e+04
1e+05
1e+06
Upstream area [ km$^2$]

Figure 6


CRPSS_pf
CRPSS_ad
LT = 5 days
CRPSS
1.0
0.8
0.6
0.4
0.2
0.0
avg dis [ $m^3/s$ ]
20
10
0
CV
2.0
1.5
1.0
0.5
0.0
NS
Pbias [%]
30
20
10
0
-10
-20
-30
Mar 2010
Sep 2010
Mar 2011
Sep 2011
Mar 2012
Sep 2012
Mar 2013
Sep 2013

## Research Highlights

- The evaluation framework of the European Flood Awareness System is presented

- Skill scores of  ensemble streamflow predictions over Europe are updated regularly

- Predictions are skillful in river basins larger than 300 km$^2$ over the 10-day range

- The use of the CRPSS based on two different references is discussed