Journal of Hydrology 522 (2015) 697–713

Contents lists available at ScienceDirect

Journal of Hydrology

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

How do I know if my forecasts are better? Using benchmarks in hydrological ensemble prediction

F. Pappenberger [a,f,g,*], M.H. Ramos [b], H.L. Cloke [d,e], F. Wetterhall [a], L. Alfieri [a,c], K. Bogner [a], A. Mueller [a,d], P. Salamon [c]

[a] European Centre For Medium Range Weather Forecasts, Reading, UK
[b] IRSTEA, Hydrology Group, UR HBAN, Antony, France
[c] IES, Joint Research Centre of the European Commission, Ispra, Italy
[d] Department of Geography & Environmental Science, University of Reading, Reading, UK
[e] Department of Meteorology, University of Reading, Reading, UK
[f] School of Geographical Sciences, University of Bristol, Bristol, UK
[g] College of Hydrology and Water Resources, Hohai University, Nanjing, China

ARTICLE INFO

Article history:
Received 5 February 2014
Received in revised form 10 January 2015
Accepted 10 January 2015
Available online 20 January 2015
This manuscript was handled by Konstantine P. Georgakakos, Editor-in-Chief, with the assistance of Yu Zhang, Associate Editor

Keywords:
Hydrological ensemble prediction
Forecast performance
Evaluation
Verification
Benchmark
Probabilistic forecasts

SUMMARY

The skill of a forecast can be assessed by comparing the relative proximity of both the forecast and a *benchmark* to the observations. Example benchmarks include climatology or a naïve forecast. Hydrological ensemble prediction systems (HEPS) are currently transforming the hydrological forecasting environment but in this new field there is little information to guide researchers and operational forecasters on how benchmarks can be best used to evaluate their probabilistic forecasts. In this study, it is identified that the forecast skill calculated can vary depending on the benchmark selected and that the selection of a benchmark for determining forecasting system skill is sensitive to a number of hydrological and system factors. A benchmark intercomparison experiment is then undertaken using the continuous ranked probability score (CRPS), a reference forecasting system and a suite of 23 different methods to derive benchmarks. The benchmarks are assessed within the operational set-up of the European Flood Awareness System (EFAS) to determine those that are 'toughest to beat' and so give the most robust discrimination of forecast skill, particularly for the spatial average fields that EFAS relies upon.

Evaluating against an observed discharge proxy the benchmark that has most utility for EFAS and avoids the most naïve skill across different hydrological situations is found to be meteorological persistency. This benchmark uses the latest meteorological observations of precipitation and temperature to drive the hydrological model. Hydrological long term average benchmarks, which are currently used in EFAS, are very easily beaten by the forecasting system and the use of these produces much naïve skill. When decomposed into seasons, the advanced meteorological benchmarks, which make use of meteorological observations from the past 20 years at the same calendar date, have the most skill discrimination. They are also good at discriminating skill in low flows and for all catchment sizes. Simpler meteorological benchmarks are particularly useful for high flows. Recommendations for EFAS are to move to routine use of meteorological persistency, an advanced meteorological benchmark and a simple meteorological benchmark in order to provide a robust evaluation of forecast skill. This work provides the first comprehensive evidence on how benchmarks can be used in evaluation of skill in probabilistic hydrological forecasts and which benchmarks are most useful for skill discrimination and avoidance of naïve skill in a large scale HEPS. It is recommended that all HEPS use the evidence and methodology provided here to evaluate which benchmarks to employ; so forecasters can have trust in their skill evaluation and will have confidence that their forecasts are indeed better.

1. Introduction

River flow forecasts are used to make decisions on upcoming floods and low flows/droughts by hydro-meteorological agencies around the world (Pagano et al., 2013; Wetterhall et al., 2013).

* Corresponding author at: European Centre for Medium-Range Weather Forecasts, Reading, UK.
E-mail address: florian.pappenberger@ecmwf.int (F. Pappenberger).

http://dx.doi.org/10.1016/j.jhydrol.2015.01.024
0022-1694/

The forecasts from these operational systems are evaluated in terms of the degree of similarity between some verification data, such as observations of river discharge, and the forecast (Demargne et al., 2009). However, another important component of the forecast evaluation is whether the forecasts add value or have *skill* compared to climatology or another simple 'best guess' (Luo et al., 2012; Perrin et al., 2006; Fewtrell et al., 2011). This is particularly important for computationally expensive forecasts which need an automated quality check, for understanding components of the forecast that may be underperforming or when new research-intensive developments are to be introduced into the forecasting system. The skill of a forecast can be assessed by how close it was to the observations compared to how close a *benchmark* was, such as a climatology or a naïve forecast (Demargne and Brown, 2013; Ewen, 2011; Garrick et al., 1978; Jolliffe and Stephenson, 2011; Kachroo, 1992; Seibert, 2001).

The relationship between skill, forecast performance and a benchmark can be generalized as:

$$\text{Skill} \sim \frac{f(\text{forecast, observations})}{f(\text{benchmark, observations})} \quad (1)$$

and such skill analysis is often integrated into an automatic forecast evaluation system. f denotes here a function (i.e. verification metric) which expresses the difference between quantities, the forecast or benchmark discharge and the observed discharge. In this paper the selection of meaningful benchmarks for evaluating skill in the hydrological ensemble prediction systems (HEPS) is considered.

1.1. Which benchmark?

The choice of the benchmark influences the resulting measure of skill (for a given verification function or metric). Differences found between the skill (and thus the quality) of different model predictions may simply be explained through variation in the underlying benchmark (Hamill and Juras, 2006; Węglarczyk, 1998). Assuming that some information is present in the forecast, benchmarks that are too naïve can easily result in a high skill being calculated. Thus the importance of using benchmarks that are known and understood is essential in assessing how 'good' forecasts are (Seibert, 2001; Garrick et al., 1978; Martinec and Rango, 1989; Murphy and Winkler, 1987; Schaefli and Gupta, 2007). There is a wealth of literature on comparing models or forecasts, developing techniques to evaluate skill and on the use of benchmarks in hydro-meteorological forecasting (Brown et al., 2010; Dawson et al., 2007; Ewen, 2011; Gordon et al., 2000; Nicolle et al., 2013; Pappenberger and Beven, 2004; Pappenberger et al., 2011a; Rodwell et al., 2010; Rykiel, 1996). Although there is surprisingly little consensus on which benchmarks are most suited for which application, benchmark suitability has been found to depend on the model structure used in the forecasting system, the season, catchment characteristics, river regime and flow conditions. What is clear however is that the choice of a benchmark is a critical issue when evaluating forecast skill.

Benchmarks can be classified by their ability to represent potential attributes of improvement of the forecasts under evaluation. Three broad classes of benchmarks are summarised in Table 1. The analysis in this paper is done only for discharge forecasts. However HEPS evaluation may also include the verification of the atmospheric forecasts (e.g. precipitation and temperature) to support the hydrologic forecast evaluation. First, there are *climatological* approaches, which use seasonal or other spatio-temporal averages of previous observed river discharges. Another type of approach considers whether there is a *change-signal*, such as when using persistency of the last observation. Benchmarking with simpler models can be viewed as a *gain-based* approach. It is useful, for instance, when evaluating the gain in performance when additional procedures or new developments are introduced into the forecasting system, such as data assimilation or post-processing techniques.

1.2. Benchmarks for hydrological ensemble predictions

This paper focuses on the use of benchmarks in the evaluation of skill of ensemble or probabilistic hydrological forecasts made by HEPS. These systems may use ensembles of meteorological forecasts, hydrological models and model parameterisations, observational uncertainties and past model errors to provide a set of forecasts which can be used to determine the likelihood of river flows, i.e., a predictive distribution (Cloke and Pappenberger, 2009; Cloke et al., 2013a,b). HEPS produce probabilistic forecasts of a future state (such as river discharge) and these probabilities also need to be evaluated when assessing the skill of the forecasts. In addition evaluation of HEPS forecasts should involve both a measure-oriented and a distribution-oriented approach (Murphy and Winkler, 1987) to fully describe the relationship between forecasts and observations based on their joint distribution.

Current practice in employing benchmarks in HEPS has been characterised through a review and assessment of the scientific literature[1] (Table 2). In general catchment size, time step or hydro-climatology does not seem to guide the choice of benchmarks, although there are a few exceptions for individual studies. However, a connection to lead time is evident in current practice: most seasonal forecasting systems use climatology as a benchmark, whereas for shorter range forecasts (several hours to several days) the variety of benchmarks used shows lack of a consensus. Only seamless predictions systems employ a single benchmark across all temporal scales (Demargne et al., 2014). One clear finding from this review is that HEPS evaluations most often use one arbitrarily chosen benchmark, and there is a lack of an extensive analysis of the impact of the choice of a benchmark of forecast performance. What is required is an evaluation of the different benchmarks within a single reference forecasting system in order to understand the impact of the choice of a benchmark to characterise forecast skill.

1.3. Aim and scope of the paper

The objective of this paper is to investigate the role of the choice of a benchmark in the assessment of the skill of hydrological ensemble forecasts through an inter-comparison of benchmarks within a reference operational forecasting system and for a given verification metric. No other aspect than forecast skill will be presented in this paper, therefore no direct comparison between forecasts and observations will be included, only a comparison between the accuracy of the different benchmarks. First the study aims to demonstrate how the calculated forecasting system skill can vary according to the underlying benchmark used. The study thus seeks to highlight the importance of a thorough assessment of benchmark selection for forecasting systems. Next the study aims to demonstrate how the skill discrimination of a benchmark is also sensitive to a number of hydrological and system factors. Lastly, this study aims to demonstrate how a benchmark intercomparison exercise can be undertaken for a large scale operational forecasting system leading to insights about how best to use benchmarks to discriminate skill in these flood forecasts. The study is set within the framework of the continental scale EFAS.

[1] Search of literature in Web of Knowledge (*wok.mimas.ac.uk/*) on the 01/10/2013 using the search terms forecasting, ensemble, hydrology and discharge. Papers were screened individually, which resulted in a total of 120 papers in the peer reviewed literature. Papers were analysed to categorise which type of benchmark was being applied (if any) and the rationale.

Table 1
A classification of benchmarks for river flow modelling and forecasting.

Class	Name	Example	Why useful	Example reference
Climatology	Conditional climatology	Seasonal average of observed discharges or other averages based on historic data	For seasonal forecasting where forecast signal is dominated by the seasonality of the flow	Pagano (2013), Randrianasolo et al. (2010)
Change-signal	Persistency	Last observed discharge	For short range forecasting where forecast signal is dominated by the auto-regression of flow	Alfieri et al. (2014), Berthet et al. (2009)
Gain-based	Simplified model	Simple lumped model which is easier to set-up and calibrated in comparison to more complex models or systems without data assimilation or post-processing techniques	For testing whether complicated models are worth their mettle, but challenging to implement over large domains	Romanowicz et al. (2008), Zalachori et al. (2012)

Table 2
Analysis of benchmarks used in survey of 120 HEPS articles.

% of articles (from 120)	Criteria met	Example references
19	Only comparison of scientific methods of HEPS and no benchmark explicitly considered	Bogner and Pappenberger (2011) compare different post-processing methods without using any benchmarks
31	Climatology (although in some references, classes not always clearly described)	Fundel et al. (2013) compare the properties of a probabilistic drought forecast to climatology and Jaun and Ahrens (2009) create a probabilistic benchmark from climatology
8	Change-signal: persistency	Hopson and Webster (2010) use flow persistence
14	Gain-based: 6% compared against current systems and 8% to simplified models	Brown and Seo (2010) compare an improved post-processing method with an existing forecasting system
28	Visual comparison	Thielen et al. (2009b) which analysed the performance of forecasts from different lead times (10 days, monthly, seasonal) for a particular flood event
<1	Impossible to determine benchmark	n/a

It is important for the reader to note that most evaluation in this study is undertaken using an observed river discharge proxy, which is calculated by running the distributed hydrological model with observed meteorological data. This is current operational practice in EFAS. The reader should also bear in mind that the quality of the hydrological model only has a minor role in most of the benchmarks tested. The evaluation seeks to find the benchmark that most closely represents (proxy) observations and thus is toughest to beat over the whole European domain (i.e. a spatial average of all grid cells). Evaluation is undertaken for forecast lead times ranging from 1 to 10 days. Additional evaluation is undertaken for the features identified in the above analysis to affect skill determination, namely: full hydrograph, the decomposed hydrograph, high and low flows, catchment response time and size. In addition to the evaluation for ungauged basins with simulated flow, the evaluation of forecasts with observations at gauged locations is needed to assess the impact of both the meteorological and hydrologic uncertainties (and is required when evaluating the EFAS post-processing component), thus here selected river gauging stations are also used to indicate how benchmarks influence point scale skill.

In the next sections, EFAS is introduced, the suite of benchmarks selected and the evaluation methods are presented. Results of the benchmark intercomparison exercise are presented and discussed within the context of utility for EFAS and the wider implications for all HEPS.

2. Methods

2.1. Forecasting system: the European Flood Awareness System and evaluation data

This study uses as its reference forecasting system the European Flood Awareness System (EFAS), which has been developed at the Joint Research Centre of the European Commission (EC-JRC) since 2002, in close collaboration with national hydrological and meteorological services and other European research institutes. The system is designed to give a European overview of ongoing floods and to forecast floods with the aim of early warning for national and trans-national river basins at imminent risk of extreme runoff conditions (Alfieri et al., 2014, 2013; Pappenberger et al., 2011b, 2013; Thielen et al., 2009a; van der Knijff et al., 2010). Fig. 1 shows the domain which EFAS covers.

The uncertainty of weather forecasts is accounted for using a multi-model approach, i.e. predictions come from different atmospheric circulation models, including deterministic weather predictions and the ensemble prediction system of the European Centre for Medium-Range Weather Forecasts. Weather predictions are the input to the LISFLOOD hydrological model (Thielen et al., 2009a). Model outputs are daily forecasts of discharge up to 15 days in advance. These are translated into probabilistic exceedances of critical thresholds and communicated to end-users (Demeritt et al., 2013; Ramos et al., 2007, 2010). EFAS results are post-processed at observed stations to derive predictive probabilities (Bogner and Pappenberger, 2010). Further information including reporting of performance can be found in the EFAS bi-monthly bulletins on http://www.efas.eu.

In this study, benchmarks are created using the latest operational set-up of the EFAS system. The study area comprises all the catchments surveyed by EFAS in Europe (mainly transnational catchments), which are spatially discretized in grid cells of 5×5 km. The evaluation period is based on daily forecasts issued from January 2009 to December 2012. Benchmarks are evaluated against observations, however, for EFAS an observed river discharge proxy is used which is calculated by running the LISFLOOD hydrological model with observed meteorological forcings. This type of proxy is very useful in distributed, continental scale modelling for evaluating the spatial predictions at all grid points, including those that are ungauged (Pappenberger et al., 2008). It offers a homogeneous verification data set at all points in space and for longer time series, usually not available when dealing with measurement datasets. The drawback of using proxy observations is that the benchmark evaluation assessment will be focused on the impact of the meteorological uncertainty only. Observed flows are needed if hydrological modelling uncertainties are to be included. Since the selection of the most discriminating benchmarks can be different under these two configurations, evaluation has also been performed against observed river discharges. This

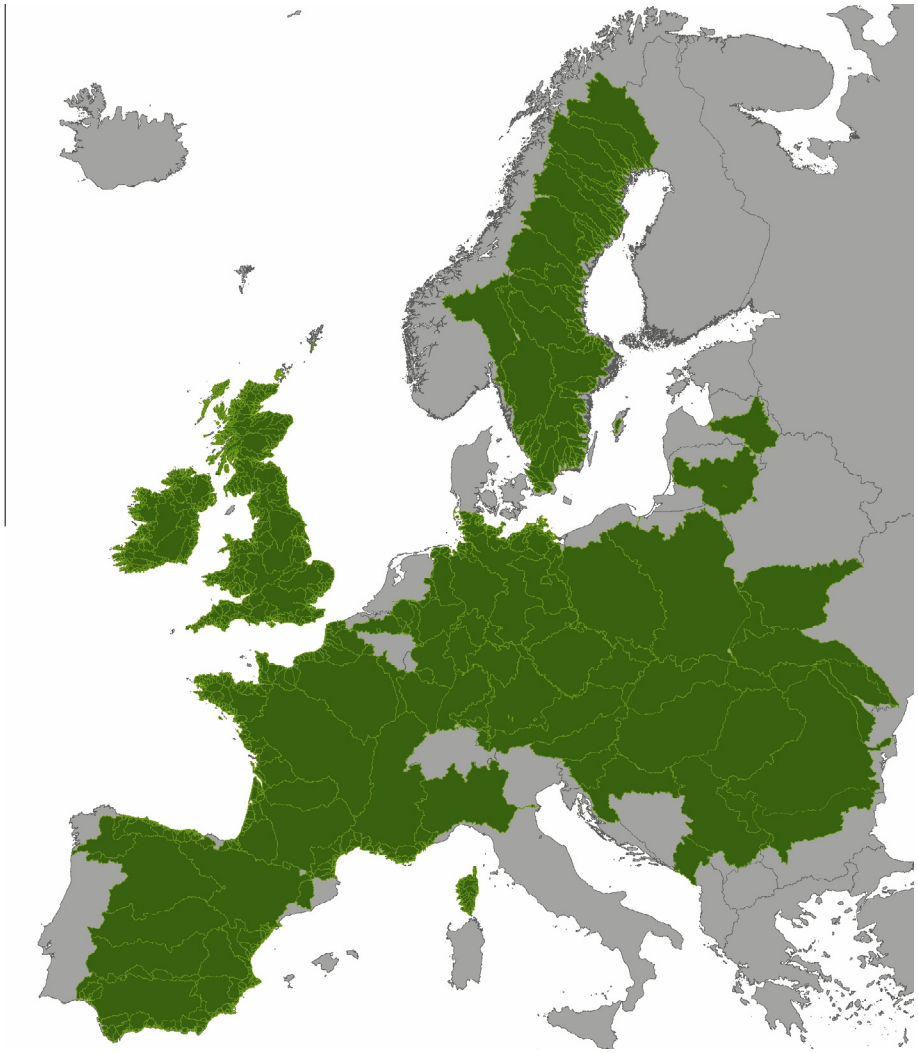

Fig. 1. The European domain that EFAS covers. Forecasts are run for all catchments shown within the domain. The green catchments indicate areas which are covered by Memoranda of Understanding for which EFAS produces flood watches and alerts. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

was however only possible for selected sites where river gauging station data were available. These points correspond to the sites where supervised post-processing is applied in EFAS to correct for the systematic and non-systematic errors in the forecast (Bogner and Pappenberger, 2010). Although the limited number of these post-processed stations in the EFAS domain prevents full operational evaluation of the system, their evaluation can provide insights into the strengths and limitations of hydrological forecasting systems and show pathways for further developments.

Finally, observed weather data were obtained by combining point observations from the Monitoring Agricultural Resources (MARS) agro-meteorological database (Baruth et al., 2007), the World Meteorological Organizations (WMO) synoptic observations (http://www.wmo.int/pages/prog/www/), the German Weather Service (http://www.dwd.de/) network and other national/regional providers. The database includes an increasing number of gauging stations over time, with figures for 2013 showing on average more than 6000 stations for precipitation and more than 4000 stations for temperature. These are used to generate daily gridded values through the inverse distance interpolation technique. A subset of the same meteorological station network was used to derive interpolated potential evapotranspiration maps using the Penman–Monteith approach.

2.2. Selection of benchmarks

Twenty-three benchmarks were designed and used in this study. They belong to the 'climatological' and 'change-signal' benchmark classes (Table 1) and were selected based on the review of current practice in HEPS (Table 2). They are subdivided into 5 main groups based on how they are constructed: simple meteorological-driven benchmarks; advanced meteorological benchmarks; simple hydrological benchmarks; climatological hydrological benchmarks; and hydrological analogues. A detailed description is given in Table 3.

In this study, no gain-based simple models were tested because of the complexity and computational expense of running any model at the continental scale, making these benchmarks infeasible in such an operational environment. The benchmarks selected also had the following attributes, essential for an operational HEPS: they can be calculated in near real-time (requires fast calculation and data processing, and data availability) in order to identify unusually poor performance in individual forecasts (known in operational forecasting as 'forecast busts'); they provide or can be calculated as probabilistic density functions of river discharge values (i.e., there is no need to reproduce a continuous time series); and, if possible, they are spatially consistent, so that not only the temporal but also the spatial correlations and covariance are main-

Table 3
Benchmarks used in this study.

Group	Name	ID	River discharge data	Meteorological input data	Class	Probabilistic
Simple Meteorological Benchmarks (SMB)	M: last obs	**a.1**	Calculated by hydrological model	The last meteorological observation of temperature and precipitation as weather forecast over all lead times (meteorological persistence)	Change signal	No
	M: average 7 days	**a.2**	Calculated by hydrological model	The average of meteorological observations of temperature and precipitation over the last 7 days as weather forecast over all lead times	Climatology/ change signal	No
	M: average 15 days	**a.3**	Calculated by hydrological model	The average of meteorological observations of temperature and precipitation over the last 15 days as weather forecast over all lead times	Climatology/ change signal	No
	M: average 30 days	**a.4**	Calculated by hydrological model	The average of meteorological observations of temperature and precipitation over the last 30 days as weather forecast over all lead times	Climatology/ change signal	No
	M: average 60 days	**a.5**	Calculated by hydrological model	The average of meteorological observations of temperature and precipitation over the last 60 days as weather forecast over all lead times	Climatology/ change signal	No
Simple Meteorological Benchmarks with zero precipitation (SMB0)	M: last obs with 0 precip	**b.1**	Calculated by hydrological model	The last meteorological observation of temperature as weather forecast over all lead times. Observations of precipitation set to 0	Change signal	No
	M: average 7 days (0 precip)	**b.2**	Calculated by hydrological model	The average of meteorological observations of temperature over the last 7 days as weather forecast over all lead times. Observations of precipitation set to 0	Climatology/ change signal	No
	M: average 15 days (0 precip)	**b.3**	Calculated by hydrological model	The average of meteorological observations of temperature over the last 15 days as weather forecast over all lead times. Observations of precipitation set to 0	Climatology/ change signal	No
	M: average 30 days (0 precip)	**b.4**	Calculated by hydrological model	The average of meteorological observations of temperature over the last 30 days as weather forecast over all lead times. Observations of precipitation set to 0	Climatology/ change signal	No
	M: average 60 days (0 precip)	**b.5**	Calculated by hydrological model	The average of meteorological observations of temperature over the last 60 days as weather forecast over all lead times. Observations of precipitation set to 0	Climatology/ change signal	No
Advanced Meteorological Benchmarks (AMB)	M: 20 years	**c.1**	Calculated by hydrological model	Observations of temperature and precipitation from the past 20 years at the same calendar day as the forecast	Climatology	Yes
	M: 20 years (analogues)	**c.2**	Calculated by hydrological model	First observations of temperature and precipitation from the past 20 years from the same calendar day as the forecast are selected. Then a subset of these are selected using the 10 years which have the smallest mean absolute error between the preceding 10 days of each past year and the current forecast day.	Climatology	Yes
Advanced Meteorological Benchmarks with zero precipitation (AMB0)	M: 20 years (0 precip)	**d.1**	Calculated by hydrological model	Observations of temperature from the past 20 years at the same calendar day with observations of precipitation set to 0	Climatology/ change signal	Yes
	M: 20 years (analogues) (0 precip)	**d.2**	Calculated by hydrological model	First observations of temperature from the past 20 years from the same calendar day as the forecast are selected. Then a subset of these are selected using the 10 years which have the smallest mean absolute error between the preceding 10 days of each past year and the current forecast day. Observations of precipitation set to 0		Yes
Simple Hydrological Benchmarks (SHB)	H: last obs	**e.1**	Last hydrological observation as streamflow forecast over all lead times (persistency forecast)	None	Change signal	No
	H: average 7 days	**e.2**	Average of discharge observations over the last 7 days as streamflow forecast over all lead times	None	Climatology/ change signal	No
	H: average 30 days	**e.3**	Average of discharge observations over the last 30 days as streamflow forecast over all lead times	None	Climatology/ change signal	No
	H: Average 60 days	**e.4**	Average of discharge observations over the last 60 days as streamflow forecast over all lead times	None	Climatology/ change signal	no

(*continued on next page*)

Table 3 (*continued*)

Group	Name	ID	River discharge data	Meteorological input data	Class	Probabilistic
	H: prob persist	**e.5**	An ensemble composed of the last 10 discharge observations as streamflow forecasts over all lead times	None	Change signal	Yes
Climatic Hydrological benchmarks (CHB)	H: climate	**f.1**	Average discharge over all available observations, previous to the forecast period, as streamflow forecast over all lead times	None	Climatology	No
	H: climate (1 month)	**f.2**	Average discharge from previous years, over all available observations, but considering only the same month as the forecast, used as streamflow forecast over all lead times	None	Climatology	No
	H: climate (3 months)	**f.3**	Average discharge from all 3-month periods corresponding to the same month as the forecast ±1 adjacent month computed over all available observations in the past years, used as streamflow forecast over all lead times	None	Climatology	No
	H: Analogue (temporal)	**f.4**	The mean absolute error between the preceding 10 days and observation period (20 years) was computed. The ten closest periods were used. The benchmark composed of the historical values from the 10 selected years	None	Climatology	No

tained (i.e., they look like physically realistic forecasts, without presenting any significant jumps). Although this last aspect of consistency is not a firm requirement for benchmark evaluation, it is a desirable feature, since they allow easier understanding of the use of benchmarks as 'reference forecasts' by end-users. It is also important to note that benchmarks can change over time: for instance, they can improve through higher resolution observations or better meteorological forecasts (e.g., in the case of the analogue-based benchmarks).

Current practice in EFAS is to use the hydrological persistency benchmark *H:last obs* (**e.1**, Table 3) and in particular the hydrological climatology benchmark, *H:climate* (**f.1**) for selected evaluation activities, and no benchmark is currently used in EFAS operational forecasting.

2.3. Evaluation of benchmarks

In this study, benchmark evaluation is undertaken with the continuous ranked probability score (CRPS, Hersbach, 2000), which is a well-known *headline score:* the score most often published in official reports and, in the case of the EFAS forecasting system, the score used to track performance in an operational mode over the past years. It is for most cases the recommended evaluation method for HEPS forecasts. The CRPS compares the distribution of the forecasts (or here benchmark forecasts) with the distribution of the observations (represented as the Heaviside function). It ranges from 0 to infinity with lower values representing a better score. It collapses to the mean absolute error for deterministic forecasts (important here as several single-valued benchmarks have been selected in the evaluation as well as probabilistic ones).

Due to the large computational burden involved in the intercomparison of such a large number of benchmarks, only the CRPS has been used for evaluation. As different evaluation criteria may lead to different results (Cloke and Pappenberger, 2008) other scores were also tested e.g. Brier Score, Logarithmic scores, RMSE on a subset of the results but found no significant differences (not shown for brevity). Although for calibration and post-processing studies a logarithmic type score should be applied for this type of analysis the CRPS is considered to be suitable (Weijs et al., 2010).

In order to make spatial locations comparable, river discharges were normalized by the mean discharge and standard deviation at each location (Trinh et al., 2013), computed over the evaluation period. Specific characteristics of the flow time series are also analysed: the falling and rising limbs of the hydrographs, as well as discharge not exceeding the 20th percentile of observed discharge (low flows) and discharge exceeding the 80th percentile (high flows). This allows analysis of different aspects of the forecasting system. In most figures the CRPS is averaged over all grid cells apart from the section where a comparison to discharge stations is shown.

When focusing on specific sub-groups of observations, the threshold-weighted Continuous Rank Probability Score ($CRPS^t$) (Gneiting and Ranjan, 2011; Lerch, 2012) can be an useful criterion as it can be conditioned on different discharge signatures (similar to weather regimes see Lerch and Thorarinsdottir, 2013). It is defined by:

$$\mathrm{CRPS}^t(f,y) = \int (F(z) - 1_A(y \leqslant z))^2 u(z)dz \qquad (2)$$

$$\text{with } 1_A(y \leqslant z) := \begin{cases} 1 & \text{if} \quad y \leqslant z \in A \\ 0 & \text{if} \quad y \leqslant z \notin A \end{cases}$$

where F is the predictive cumulative distribution function (cdf) corresponding to the probability density function (pdf) f of the forecast and y is the observation. u is a nonnegative weight function of the forecast z, with $(z) = 1_A(z \geqslant r)$, which is equal to 1 for z values of the observation that are larger than or equal to a threshold $r \in \mathbb{R}$ (otherwise the function is 0). In our case, the thresholds are chosen according to the flow gradient (rising and falling limbs) and to the 20th and 80th flow percentiles. In the case of the full hydrograph evaluation the weight function u is set to 1 for all forecast discharge values and the $CRPS^t$ becomes equivalent to the traditional CRPS. For simplicity, the $CRPS^t$ is written as CRPS throughout this paper.

In the following analysis, each benchmark in Table 3 is evaluated against an observed discharge proxy. This is done separately for each day of lead time, from 1 to 10 days, and considers the following:

– *Full hydrograph:* all daily time steps of discharge forecasts are considered and the CRPS is averaged over (i) the whole time series (4 years, 2009–2012), and (ii) the seasons of the year

(3-month verification periods). In both cases, evaluation is carried out against simulated discharges (observed meteorological data used as input to the forecasting system);
- *Decomposed hydrograph:* daily time steps of discharge forecasts are separated according to the discharge gradient into rising limbs (time steps at which discharges are increasing in time: $Q(t)) > Q(t+1)$) and falling limbs (time steps at which discharges are decreasing in time: $Q(t) \leqslant Q(t-1)$). The CRPS is averaged over the respective time steps in each group, spanning the whole time series (4 years, 2009–2012). Evaluation is carried out against simulated discharges;
- *High and low flows:* from the daily time steps of discharge forecasts, two flow groups are considered. One focuses on high flows and considers only discharges that exceed the 80th percentile of observed discharge. Another focusses on low flows and includes only time steps at which discharges are lower than the percentile 20th of observed discharge. Percentiles are estimated from a climatological record of 30 years of model runs. The CRPS is averaged over the respective time steps in each group, spanning the whole time series (4 years, 2009–2012). Evaluation is carried out against simulated discharges;
- *Catchment characteristics:* all daily time steps of discharge forecasts are considered and the CRPS is averaged over the whole time series (4 years, 2009–2012). The results are analysed according to the catchment response time and size. To establish the flashiness of a catchment, the correlation between a time series with a lag of one and a lag of four is compared. The larger the difference, the more flashy a catchment can be classified as. This has been computed for all EFAS grid points over Europe and the catchments classified into high, low and medium flashiness by using the upper, medium and lower third of the resulting differences. Evaluation is carried out against simulated discharges;

An evaluation is also performed for observed discharge at the river gauging stations which are currently setup in EFAS to use supervised postprocessing. All daily time steps of discharge forecasts are considered and the CRPS is averaged over the whole time series (4 years, 2009–2012). Note that this type of evaluation with observed discharge on a limited set of points is included as an indication of using gauged observations in this analysis. It cannot be directly compared with the spatial average taken over all grid cells using an observed discharge proxy. As further supervised postprocessing stations become available in the operational EFAS, benchmark evaluation will be part of the routine reanalysis of the system. Thus a full evaluation for EFAS should assess the hierarchy of benchmarks in terms of skill discrimination for evaluations with both simulated discharge (proxy observations) and observed discharge at stations.

3. Spatial results using proxy observations of discharge

The overall performance is assessed with the CRPS. Lower values of the CRPS indicate a better representation of observations, and thus also indicate a benchmark which is 'tougher to beat' and which can better discriminate skill.

3.1. Rising and falling limbs of the hydrographs

In this section hydrographs are decomposed into their rising and falling limbs. The shape of these hydrograph components are governed by the meteorological forcing and the transient and permanent catchment characteristics (saturation and land cover, for example). The rising limb is usually steeper and associated with higher precipitation inputs, whereas the falling limb is marked by a more gradual decrease in discharge, and this is more governed by the time transfer of water within the catchment to its outlet than the meteorological forcing. The rising limb is particularly important to analyse in any flood forecasting system.

To focus on rising and falling limbs, CRPS values were calculated using a weighted function based on a positive (or negative) slope (see Eq. (2)). Fig. 2 shows the results of benchmarks performance when the score is computed over rising and falling limbs, respectively.

In Fig. 2a it can be seen that for the rising limb the meteorological persistence (*M: last obs* ***a.1***), and the hydrological persistence (*H: last obs* **e.1**) have the lowest CRPS at first and thus the most skill discrimination at short lead times, when river discharge increases quickly from one time step to the other. The *Advanced Meteorological Benchmarks* (*M: 20 years* ***c.1*** *and M: 20 years analogues* ***c.2***) have the lowest CRPS from day 3 of lead time onwards, and thus the most skill discrimination at longer lead times. Fig. 2b shows the difference between the CRPS values of the rising limb to the full hydrograph evaluation. This is useful in understanding how the ability of individual benchmarks to discriminate skill changes as a hydrograph is decomposed, essential if one is using a particular benchmark in a forecasting system. Numbers below zero in Fig. 2b indicate that the full hydrograph has lower CRPS values than the rising limb. This is true for all but the hydrological persistence (*H: last obs* **e.1**), which has a relative performance equivalent to zero due to the fact that this benchmark, because of its nature, has a non-changing error structure. All of the benchmarks are able to discriminate less skill for the rising limb than for the full hydrograph.

From Fig. 2c, one can see that for the falling limb, the meteorological persistence (*M: last obs* ***a.1***) and the Advanced Meteorological Benchmarks (*M: 20 years* ***c.1*** *and M: 20 years analogues* ***c.2***) have the lowest CRPS values at first, and these all have the best skill discrimination at short lead times (<2 days), with the Advanced Meteorological Benchmarks remaining the most discriminatory at longer lead times (2 days<). The hydrological persistency benchmark does not do so well in comparison to the rising limb. This is also confirmed by the comparison of the skill discrimination of the falling limb and full hydrograph (Fig. 2d). It is interesting to note from this figure that for the falling limb, the CRPS difference is generally higher than zero. This is because the falling limb is dominated by non-precipitation and therefore easier to forecast than the full hydrograph. The *H: last obs* **e.1** benchmark, as a climatic benchmark, and as noted for the rising limb, is not sensitive to the separation into hydrograph components.

3.2. Full hydrograph

In Fig. 3, the CRPS for the different benchmarks evaluated for the full time series is shown. Fig. 3 groups the benchmarks according to Table 3. A number of the benchmarks perform nearly equally well at the lead time of one day. The lowest (best) CRPS is initially given by the *m:last obs* benchmark (**a.1** in Table 3) based on the last observed meteorological observations. This benchmark however quickly deteriorates and from lead time 3 days onwards the *m: 20 years* (**c.1**) benchmark dominates, which is given by the ensemble of meteorological forecasts based on the last 20 years of observations. There is no significant difference between *m:20 years* and the analogue version of this benchmark *m:20 years* (*analogue*) (***c.2***), and both of these Advanced Meteorological Benchmarks perform equally well. This possibly indicates that the method for selecting the analogue is probably not sophisticated enough as others studies report greater success using this approach (Radanovics et al., 2013). Most of the other meteorological benchmarks behave very similarly to each other with no significant impact of the averaging window size. This is likely caused by the smoothing effect of the hydrological model (Fig. 3a, b, and d).

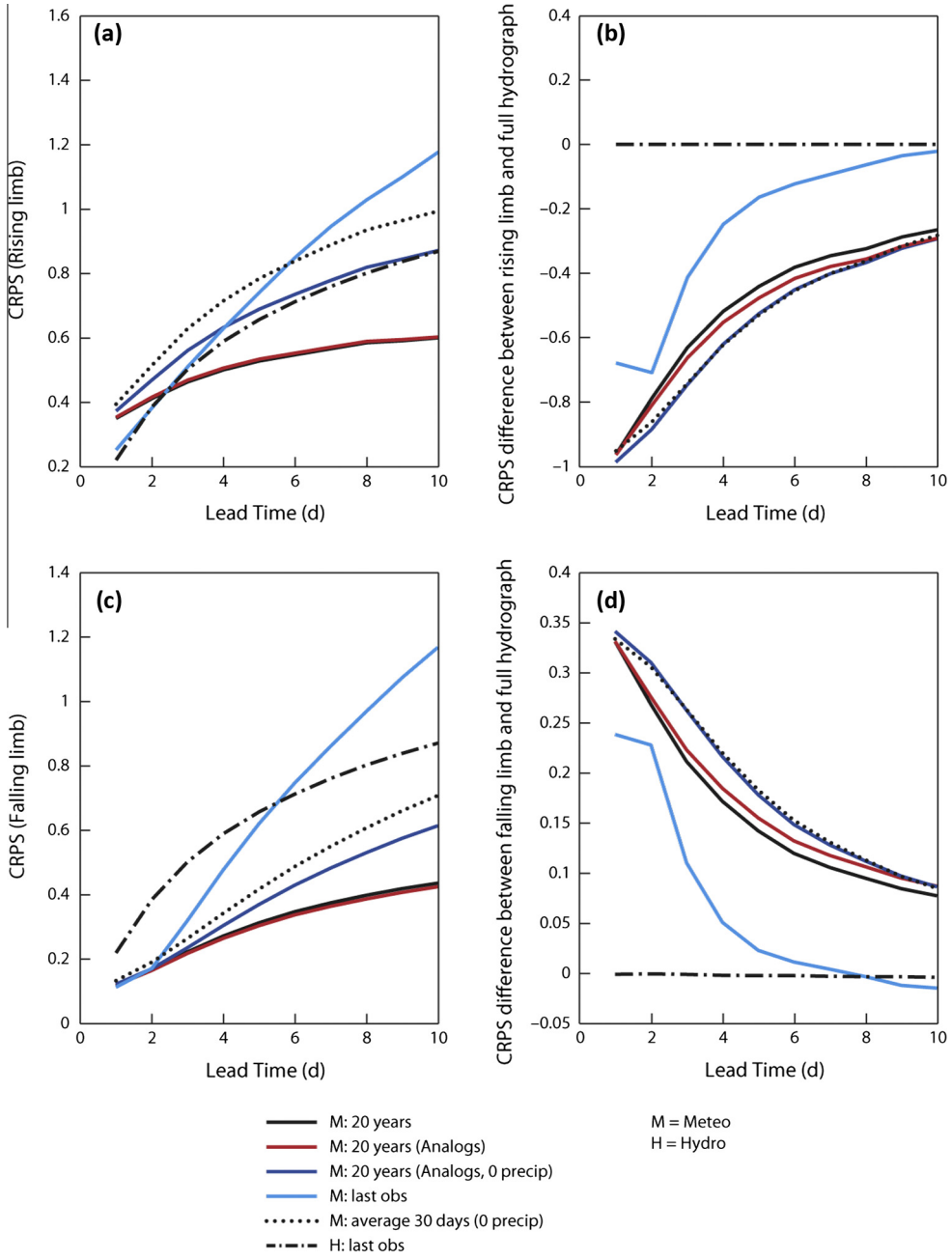

Fig. 2. (*a, top left hand side*) CRPS of the rising limb of the hydrograph; (*b top right hand side*) CRPS difference of the rising limb of the hydrograph in comparison to the full hydrograph. (*c, bottom left hand side*) CRPS of the falling limb of the hydrograph; (*d, bottom right hand side*) CRPS difference of the falling limb of the hydrograph in comparison to the full hydrograph. Values below zero indicate that the CRPS for the rising/falling limb is higher than for the full hydrograph, and thus the benchmark has less ability to discriminate skill for the rising/falling limb.

All hydrological benchmarks (Fig. 3e and f) have higher (worse) CRPS values than any of the meteorological benchmarks, with greater differences in particular at the shorter lead time of one day. The best hydrological benchmarks are those based on persistency (Fig. 3e), with the *H:last obs* (**e.1**), the last observed discharge performing best up to a lead time of 3 days and the probabilistic hydrologic persistency benchmark, *H: prob. persist* (**e.5**) being the best at longer lead times. The worst performances are obtained for climatology-based benchmarks based on hydrological long term averages (Fig. 3e and f). These benchmarks exhibit a marked flat behaviour with very similar CRPS values for all lead times, showing that average values of flows are not skilful forecasts at any of the lead times tested. These hydrological long term average benchmarks are very easily beaten by a forecasting system and would produce much naïve skill.

For both meteorological and hydrological benchmarks, the closer a benchmark reflects climatology, the less it is influenced by lead time, which is to be expected as the mean error will be constant. For hydrological average benchmarks (Fig. 3e and f) the lower performance may be because these may exhibit a distinct jump in comparison to the current observed flow. For example, if the climatological average flow is 400 m^3/s and the last observed discharge is 200 m^3/s, then the forecast will have a discontinuity jump at the first lead time.

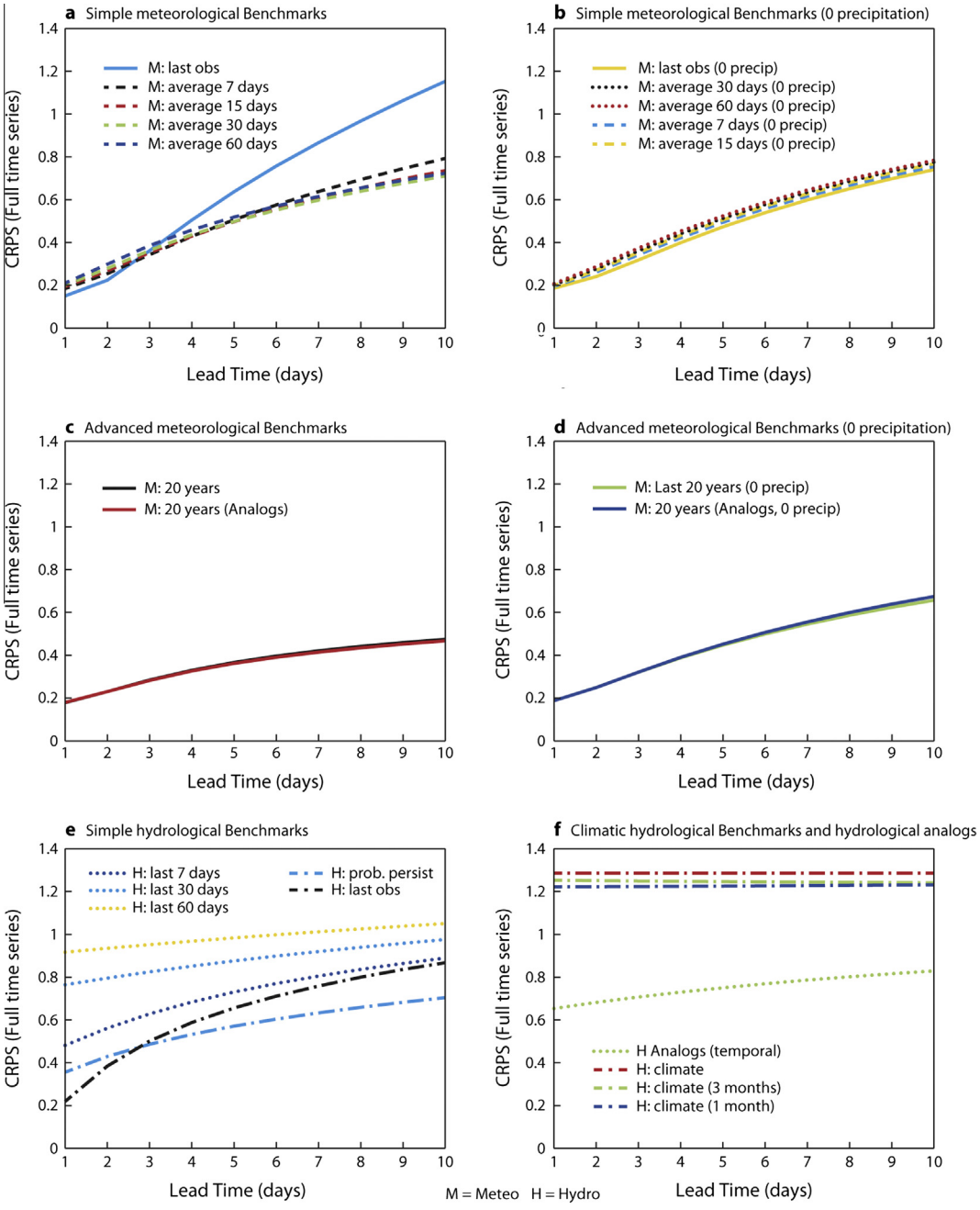

Fig. 3. CRPS of the full hydrograph for lead times from 1 to 10 days.

3.3. Seasonality

In the previous analysis, the average of CRPS values over 4 years of forecasts was computed. However, a clear seasonal dependency of skill can be expected in Europe for the European Flood Awareness System (Kiese et al., 2010; Pappenberger and Buizza, 2009; Pappenberger et al., 2011b). In Fig. 4, the CRPS (of the full hydrograph) is analysed separately according to three seasons for a lead time of 5 days (similar results are found for other lead times) based on standard division of the water year in Europe (October to September). The first season is the period from October to January, the second one from February to May and the third one from June to September.

In this and the following sections only a subset of benchmarks has been presented from the full results for brevity. Selection is based on ability to discriminate in the previous section supplemented with those benchmarks that exhibit lowest CRPS values in these sections.

It is clear that in most cases the February to May season performs worse than the other seasons as would be expected because of the complex contributions of snowmelt to runoff, with errors in both precipitation and temperature affecting results as well as the lower predictability of precipitation. The only case where this pattern is significantly different is the meteorological persistency benchmark (*M: last obs:* ***a.1***). Overall the Jun-Sep season has lower CRPS values, which may be due to the fact that it is typically a drier period, with lower values of average runoff. However, at the storm event scale, errors can be bigger, due to the higher proportion of convection, flash floods, and space–time variability of events in general. This might be the cause for larger error (in summer) for

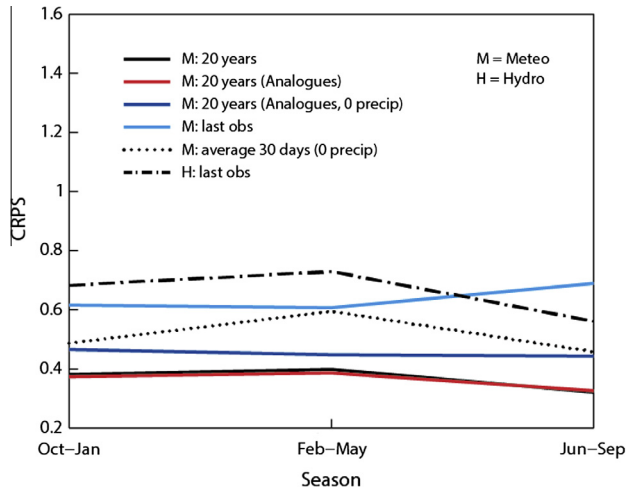

Fig. 4. CRPS for the leadtime 5 days and 3 different seasons.

the "Meteo: last obs" family. Indeed, extreme storms are often short, but if such rain rates are fed as input in the benchmark simulation for the whole forecast horizon (10 days) you get unrealistic values of discharge with a large associated error. Overall the benchmarks with the most skill discrimination in all seasons (i.e. the lowest CRPS values) are the Advanced Meteorological Benchmarks (*M: 20 years c.1 and M: 20 years analogues c.2*).

3.4. High and low flows

In this section evaluation is performed on flows separated by their percentile, i.e. high and low flow. Evaluating high flow is important for flood forecasting systems, and low flow, which tends to have a longer predictability is useful in hydrological drought forecasting systems, and is less often investigated than high flows (Fundel et al., 2012; Pushpalatha et al., 2012). Both of these functions are performed by EFAS and so are investigated here. Fig. 5 shows the results of the evaluation of the benchmarks studied when the CRPS values are evaluated based only on flows lower than the 20th percentile of the reference discharges (low flows) and on flows exceeding the 80th percentile (high flows), respectively.

For low flows, the *M: 20 years* (*analogues*) (*0 precip*) (**d.2**) benchmark has the lowest CRPS. This benchmark uses observations of temperature from the past 20 years from the same calendar date followed by a selection based on the MAE and with the observations of precipitation set to zero. It is thus similar to the Advanced Meteorological Benchmark **c.2**, but just uses zero precipitation. The probabilistic hydrological persistency (*H: prob persist* ***e.5***) and the hydrological temporal analogues (H: Analogues – temporal, **f.4**) can also discriminate skill well and have low CRPS values (Fig. 5a). Comparatively to the CRPS for the full hydrograph (Fig. 5b), it is very noticeable that these benchmarks can discriminate more skill in low flow regimes (the relative values nearly reach a value of 1).

For high flows, the reverse is seen, with *the simple meteorological benchmarks: M: average 30 days, 0 precip* (***b.4***) *and M: last obs* (***a.1***) being able to discriminate the most skill. Although this at first may be counter-intuitive, it is likely to be because of the dominance of hydrograph recession at high flows. All other meteorological benchmarks do not discriminate skill well and hydrological benchmarks based on discharge are also not suitable for discriminating skill at high flows. These findings are reinforced by the comparison of high flow CRPS with the full hydrograph in Fig. 5c – the opposite findings compared to Fig. 5d.

3.5. Dependency on catchment size and response time

The dependency of the skill discrimination of the benchmarks on catchment size was evaluated for a fixed lead time of 3 days, and results are shown in Fig. 6. All hydrological benchmarks, as exemplified by *H: last obs*, indicate a "V" shape, meaning that they have the lowest CRPS (e.g. higher skill discrimination) at catchments with sizes between 3000 and 6000 km^2. In contrast, all meteorological benchmarks show a clear decline of CRPS values as catchment size increases: the smoothing effect of meteorological inputs at larger catchments reflects positively in ability to discriminate skill. For all catchment sizes the Advanced Meteorological Benchmarks (**c.1** and **c.2**) again show the lowest CRPS and the best ability to discriminate skill.

Faster responding, flashy, catchments should be more difficult to model with this type of system, and thus one would expect the CRPS to be higher for these in all our benchmarks. To establish the flashiness of a catchment the auto correlation (lag −3) is computed; the larger the difference the more flashy a catchment. This has been computed for all EFAS grid points over Europe and the catchments classified into high, low and medium flashiness by using the upper, medium and lower third of the resulting differences. Fig. 7 displays the results for each benchmark analysed according to its classification in terms of response time. For all benchmarks, there is a general tendency to observe higher CRPS values (i.e., worse skill discrimination) at more fast-responding catchments. Here again meteorological benchmarks show better skill discrimination.

4. Station results using gauged observations of discharge

In this section the benchmarks are evaluated against point observations of river flow at several gauging stations where real-time data has been collected for the daily operational EFAS forecast runs. Fig. 8 illustrates the results obtained at a subset of five of these gauging stations, which are representative of the CRPS performance obtained at all investigated stations. Since the CRPS for ensemble prediction systems is quite sensitive with respect to observed values falling out of the range of all the predicted model outcomes, the CRPS values reach an order of magnitude higher than the previous results (which was based on an observation proxy, a reference discharge that is not observed but simulated by the hydrological model using observed meteorological inputs). This is particularly observed for two of the advanced meteorological benchmarks analysed. They notably show the worst results for short lead-times. This can be explained by the fact, that there has been no data assimilation method included for updating the model output with past observed discharge values. However without adjusting the differences between the model simulation and the observation at the forecast initialization, the error for the first lead-times, where the forecast ensembles show only very little spread, can be quite big and the observed value lies most often far outside of the ensemble range.

When only single model outputs are compared with the observations, the CRPS reduces to the MAE. The lowest CRPS (MAE) values are for the Simple Hydrological Benchmarks: *H: average 7 days* (**e.2**), *H: average 30 days* (**e.3**) and with *H: last obs* (**e.1**) also having a relatively low CRPS (MAE) value.

The results in Section 4 from the spatial analysis suggested that the various meteorological benchmarks are the most valuable for discriminating skill routinely against spatial proxies of discharge observations, for spatial warning systems such as EFAS. This analysis, of course, only considered the meteorological uncertainty in the forecasting system. The results here in Section 5 show that

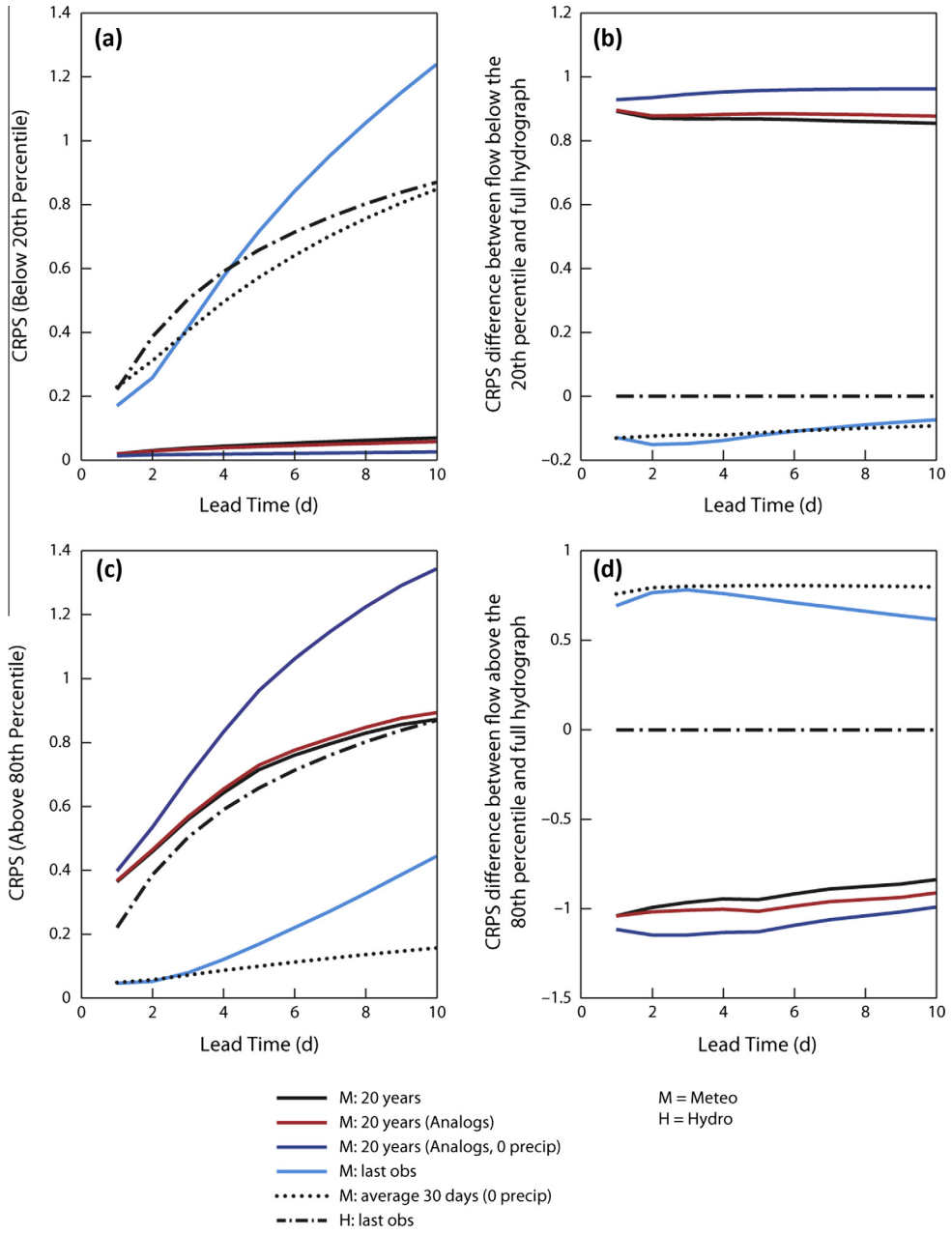

Fig. 5. (*a, top left hand side*) CRPS of the 20th percentile of flows; (*b, top right hand side*) CRPS difference of the 20th percentile of the hydrograph in comparison to the full hydrograph. (*c, bottom left hand side*) CRPS of the 80th percentile of flows; (*d, bottom right hand side*) CRPS difference of the 80th percentile of the hydrograph in comparison to the full hydrograph. Positive values mean that the below 20th percentile predictions can discriminate more skill than the full hydrograph predictions and *vice versa*.

for point analysis it is likely that the hydrological benchmarks can be more useful. This is likely to be because the evaluation with observed discharge takes into account both the meteorological and hydrological uncertainties. When evaluating EFAS for its flood warnings over the whole gridded domain, the evaluation with simulated discharge is meaningful. However the analysis of EFAS performance should also include a comparison with observed discharges on gauged basins and a comparison of warnings with flooding/no-flooding observed events over the spatial domain to account for both the meteorological and hydrological uncertainties. A full station analysis of benchmark skill discrimination considering both meteorological and hydrological uncertainties will be implemented routinely as further postprocessed observation stations become operational within EFAS.

5. Results summary

In order to provide a summary of the results of all the evaluations undertaken a matrix of the CRPS values calculated over the previous 6 results sections are presented in Fig. 9. The benchmarks with the most skill discrimination are those with bluer colours. These are summarised in Table 4 for key hydrograph features. The effect of lead time can be clearly seen in the matrix, with better skill discrimination at shorter lead times, and variations between benchmarks and between evaluations. The difference between the station analysis (observed discharge) and the spatial analysis (observed discharge proxy) is also demonstrated in this figure, with the CRPS values in general much higher for the station analysis and the better discrimination of the hydrological

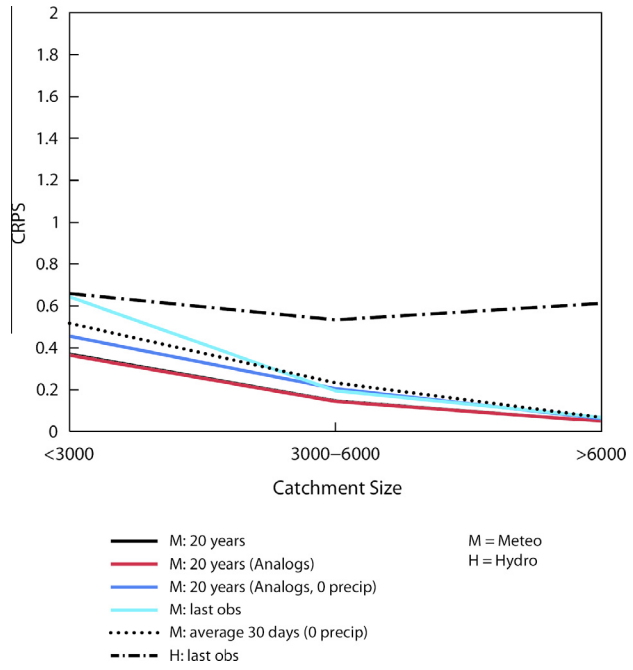

Fig. 6. Evaluation of skill differentiation measured with CRPS for different catchment sizes (lead time 3 days as an example).

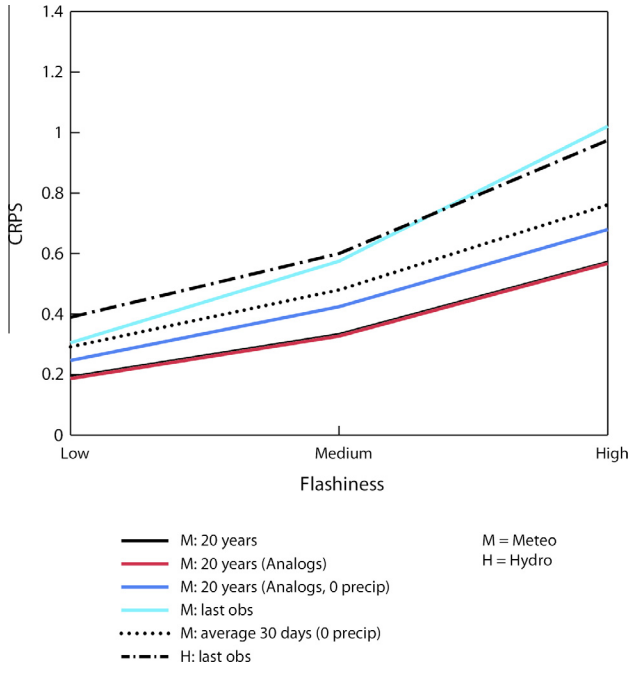

Fig. 7. Evaluation of skill differentiation measured with CRPS for catchments with differing flashiness in their hydrographs (lead time 3 days as an example).

benchmarks evident (as would be expected when using actual observed discharge rather than proxy).

6. Discussion

6.1. Avoiding naïve skill

The most useful and honest benchmark for use in forecast evaluation is one that is tough to beat. The previous analysis has demonstrated that different benchmarks evaluated with the CRPS have a variable discrimination of skill when evaluated against both proxy and gauged observations, and this often changes with lead time (Fig. 9). Those that have lower CRPS values can be used to best discriminate the skill of the forecasts. A summary of these discriminating benchmarks is provided in Table 4.

The European Flood Awareness System is designed to predict floods spatially across the whole of Europe. The rising limb and the top 20% of the discharge are the most important features to evaluate. In addition, spatial fields of observation proxies are the most useful 'observation' to evaluate against, although this inevitably means that the majority of the evaluation is of the meteorological forcing. In addition EFAS provides information on the medium range and therefore benchmarks that provide more skill discrimination at the longer lead times, i.e. those driven by the last meteorological observation or by an ensemble of observations over the last year needs to be considered.

Benchmarks are important to understand the value of a forecasting system, judge any additional developments and allow for a communication on skill to any stakeholders and end-users, hopefully leading to better decisions (Ramos et al., 2010, 2013). Such a skill analysis allows a system, model or forecast to be classified as having:

- *No skill:* The Hydrological Ensemble System is consistently worse than a set benchmark.
- *Naïve skill:* The forecast system is skilful against a too simplistic benchmark. More challenging (difficult to beat) benchmarks could be designed.
- *Real skill:* No benchmark which can be implemented at a lower cost than the operational system can beat my forecast system.

In the last column of Table 4 a comparison of the CRPS values between the benchmarks *H: last obs* and *H: climate*, which are those currently used in EFAS, and the most discriminatory benchmarks found in the analysis presented in this paper is given. These values highlight the substantial overestimation (Eq. (3)) of forecast skill in the current benchmarks, and this naïve skill can be avoided by implementing the most discriminating benchmarks in EFAS.

$$\text{naïve skill} = \frac{|f(\text{most discriminating benchmark}) - f(\text{current benchmark})|}{f(\text{current benchmark})} \quad (3)$$

6.2. Towards implementing benchmark analysis as an operational norm

The results presented here are evidence of the variable skill discrimination of many commonly used benchmarks, and the dangers of naïve skill attribution when using too 'easy' a benchmark. There is a distinct opportunity in the HEPS community to routinely implement this benchmark analysis, which will objectively select benchmarks with the most skill discrimination to avoid naïve skill when reporting HEPS performance with skill scores. This will make sure that knowing exactly how much better our forecasts are is the operational norm.

One important aspect to note about all of the benchmarks described so far is that they have distinctly varying computational costs (Fig. 10) and this is often the main barrier to implementing a benchmark operationally. In particular the meteorological benchmarks identified are more computationally expensive than the hydrological benchmarks as they involve running the hydrological model in its current operational setup. However it is important to note that although easier to construct, observation based benchmarks are not without their problems and are highly dependent on observed discharge data. Observed discharge data captured during floods usually contains errors from when measurement devices

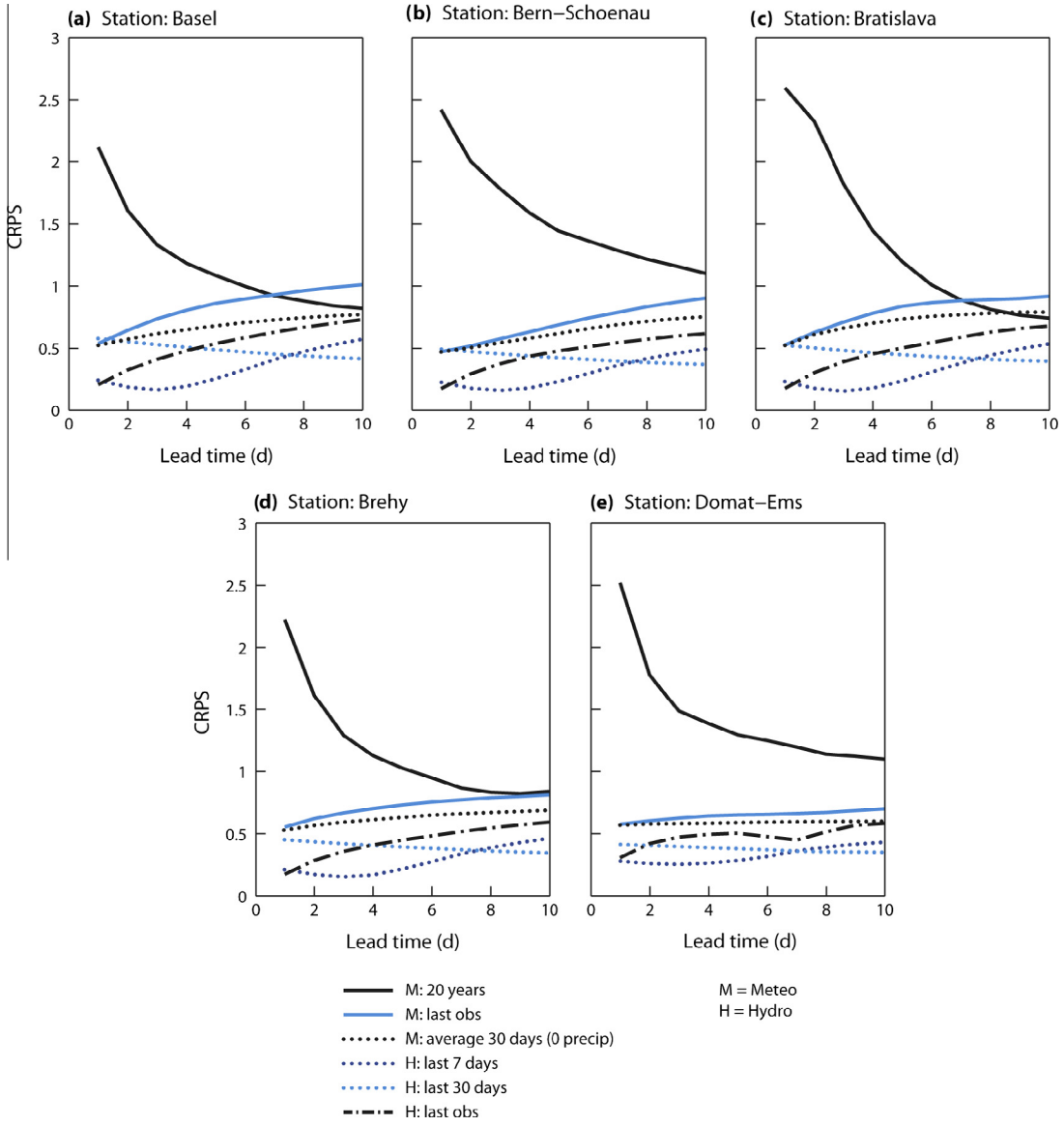

Fig. 8. Evaluation of skill differentiation measured with CRPS for observed discharge data for five stations.

stop functioning or are destroyed or bypassed by flood waters. Another difficulty with observation-based benchmarks is that historic data series are often too short to build a robust ensemble forecast as benchmark.

Although a whole suite of benchmarks was tested within this experimental study, the choice of benchmarks was determined by the exact nature of the pan-European, distributed forecasting system used and within the framework of an operational HEPS. When EFAS runs operationally it produces an ensemble of river discharge forecasts on each EFAS grid cell (259,024 cells across Europe). Therefore, benchmarks which require a high degree of human supervision or optimization had to be excluded. For example, many post-processing methods have a component of forecasting errors, which could be used to formulate a benchmark, but the implementation of such benchmarks would require often supervised fitting and is not easy to automate (Bogner and Pappenberger, 2011).

In addition, since the ultimate aim is to implement the results of this study operationally, each benchmark that was tested had to be easy to maintain and computationally inexpensive. Therefore, even if adding benchmarks operationally based on simplified hydrological models would be desirable, currently for EFAS the cost in terms of maintenance and development remains too high.

The reader should note again that it is recommended that a range of performance measures/verification scores are used to test conclusions from any study such as this (Cloke and Pappenberger, 2008), and that the selection of a benchmark will always be specific to the problem. Benchmark models cannot necessarily be easily and straightforwardly transferred to a different catchment, a different type of forecast system, application or modelling environment (Krause et al., 2005; Pushpalatha et al., 2012; Ritter and Munoz-Carpena, 2013; Schaefli and Gupta, 2007). Although establishing a single consensus regarding which benchmark to use for hydrological modelling and forecasting is not realistic (Perrin et al., 2006; Węglarczyk, 1998), this study aims to clarify the advantages and disadvantages of the various options and demonstrating best practice in the evaluation of hydrological predictions. It is hoped that the methods and evaluations presented in this paper themselves become a 'benchmark' for forecasters when implementing a benchmark evaluation in HEPS.

6.3. Future research on benchmark selection

Although many benchmarks were evaluated in this experiment, there remain several aspects that could be further analysed. For

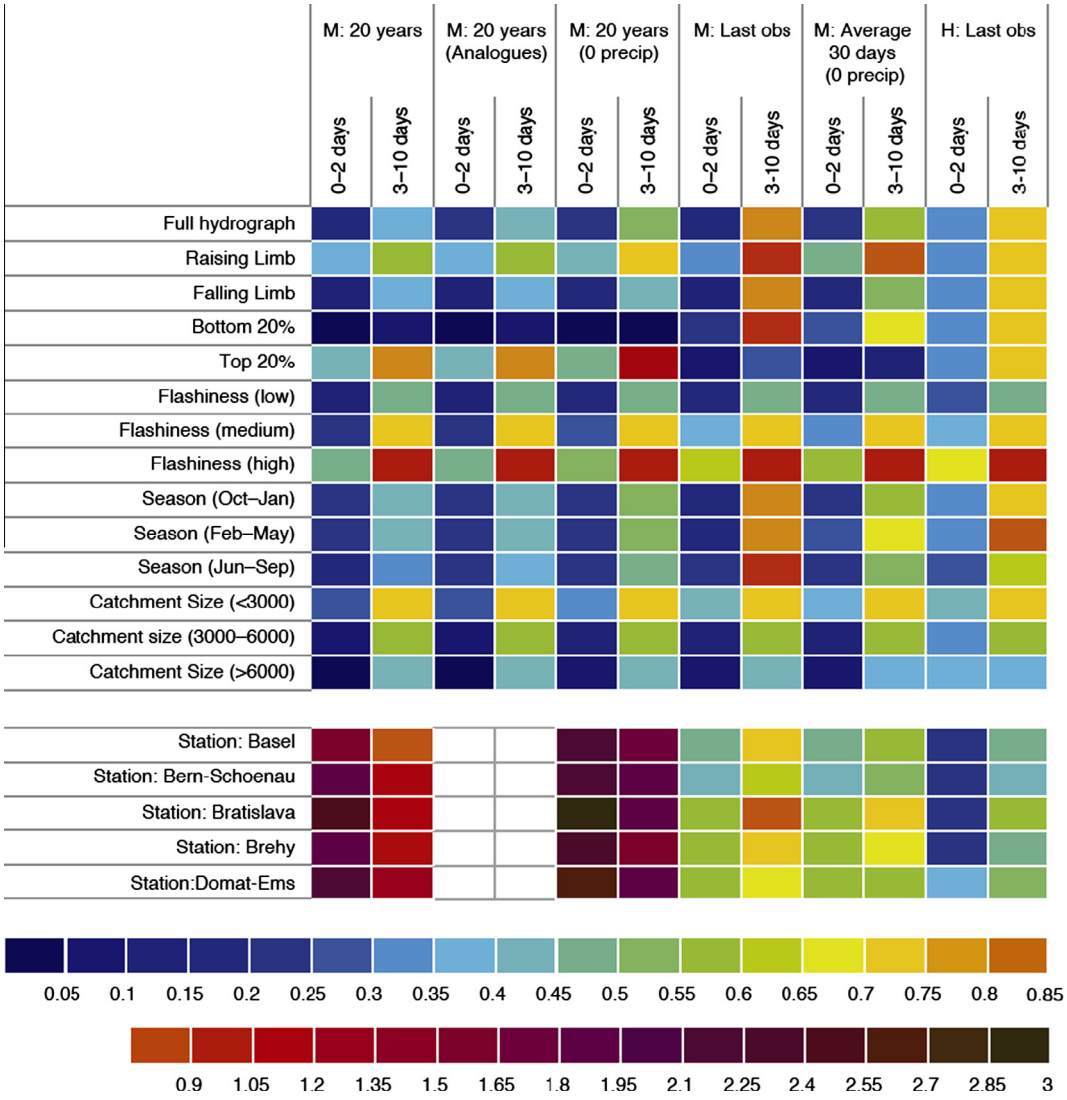

Fig. 9. Summary matrix of the CRPS results for all experiments.

Table 4
The most discriminating benchmark methods from those tested presented according to different hydrograph features (verified with simulated flows) and the skill over estimation avoided in comparison to current EFAS benchmarks. The benchmarks which rely on the last observations are deterministic all others are probabilistic.

Hydrograph feature		Most discriminating method	Naïve skill avoided	
			H: last obs (%)	H: climate (%)
Full hydrograph	0–2 days	M: last obs; (probabilistic alternative M: 20 years)	34	81
	3–10 days	M: 20 years	45	68
Rising limb	0–2 days	H: last obs[a] (probabilistic alternative M: 20 years)	0	71
	3–10 days	M: 20 years	45	68
Falling limb	0–2 days	M: last obs	45	85
	3–10 days	M: 20 years	51	72
Bottom 20%	0–2 days	M: 20 years (zero precip)[b]	95	99
	3–10 days	M: 20 years (zero precip)	96	98
Top 20%	0–2 days	M: last obs[c] (probabilistic alternative M: 20 years)	84	95
	3–10 days	M: last obs (probabilistic alternative M: 20 years)	63	78

[a] Not significantly different from: M: last obs.
[b] Not significantly different from: M: 20 years.
[c] Not significantly different from: M: average 7 days.

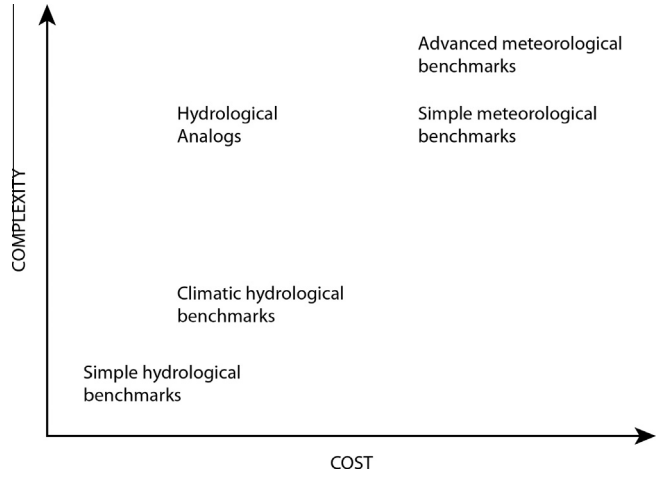

Fig. 10. Illustrative scheme of cost and complexity of the five benchmark groups used in this study.

example the results presented showed that *H: prob. persist* the probabilistic hydrological benchmark (**c.5**) (which is an amalgamation of several deterministic benchmarks) does not discriminate more skill than the deterministic hydrological persistence (**c.1**), indicating that a different approach to build this benchmark may have to be developed (Brown, 2013). The analogue benchmarks could also be improved by for example conditioning them on large scale atmospheric patterns. This could potentially improve skill discrimination further. As EFAS is a pan-European system, the analogues would have to be optimised differently for different catchments, which would further complicate the selection. However, if successful, such a tailored analogue ensemble could prove useful, especially if the ensemble forecasts would for some reason not be available. More complex analogue methods are difficult to envisage at the European Scale, however they have been extensively tested for use with probabilistic hydrological forecasts in France, where the meteorological analogues are selected by analysing atmospheric patterns form historical observations (Marty et al., 2012; Obled et al., 2002; Radanovics et al., 2013).

The benchmarks in this paper could be improved for example by lagging the benchmarks and weighting them according to past performance similar to the methods presented by Stedinger and Kim (2010) and Weijs and Van de Giesen (2013), which they applied for forecasts. Future work could also look at the supervised post-processing of further stations in order to build up the point scale analysis, this would also help determine whether there were any spatial patterns within the benchmarks that might be important. The paper mainly concentrates on a pan-European spatial average. The optimal length and ensemble size of the hydrological model driven by past observations also needs to be investigated further. One other issue that requires further study is that of verifying a seamless prediction system (from short to long range, cf. Ebert et al., 2013), which potentially requires the use of the same benchmark for all lead times. As the EFAS system focuses on flood warnings, further analysis on the full range of high flow conditions would be a useful future research exercise. Here only the results of a threshold of 80th percentile were used to represent high flows. It may be that the higher discharge percentiles would lead to results that would lead to the selection of different benchmarks although conclusions may be difficult to draw due to the increasing the sampling uncertainty of the computed verification metrics.

Further studies could consider other aspects of the forecast/benchmark accuracy by using different verification metrics since the CRPS is only an overall score, as a single verification metric cannot describe in depth the quality of any forecast system. This is especially important since other verification criteria might be more important to some forecast users than the CRPS (e.g. discrimination between different flood warning levels, reliability). One further study of value would be a more detailed sensitivity study of the total uncertainty when evaluating with observed discharge versus measuring the impact of the meteorological uncertainty only when evaluating with simulated discharge.

7. Conclusions

This paper considers how best to use benchmarks for forecast evaluation in HEPS, which is particularly useful for automatic quality checking of large scale forecasts and for when forecasting system upgrades are made. A suite of 23 benchmarks were evaluated within the framework of the European Flood Awareness System (EFAS), which produces probabilistic forecasts for the European continent on a 5 × 5 km grid with a lead time of 10 days. The Continuous Rank Probability Score (CRPS) was used to compare the full discharge hydrograph, the falling and rising limb as well as the 20th and 80th discharge percentiles. Seasonality and catchment characteristics were also considered. Up to 99% overestimation of skill was detected in using the current benchmarks of Hydrological persistence (*H: last obs*) and climatology (*H: clim*), with this naïve skill likely resulting in overconfidence in forecasts.

The recommended practice for EFAS, which evaluates forecast skill on pan-European spatial averages of simulated discharge, is to implement Advanced Meteorological Benchmarks (*M: 20 years*), as these discriminate skill better than those benchmarks currently applied for longer lead times. However, the implementation of an ensemble of benchmarks would also be useful, with the inclusion of meteorological persistency and a simple meteorological benchmark providing the most robust evaluation of forecast skill.

This paper has provided an analysis on how benchmarks intercompare in this reference forecasting system, but these results are useful for all operational HEPS forecasters. The magnitudes of the naïve skill that can be avoided by altering the benchmark used for the skill calculation has been demonstrated. It has also been demonstrated that using a benchmark for large scale HEPS although computationally expensive remains feasible within the forecasting timescale. Given the impact of the choice of benchmarks on the forecast performance results, it should be a community requirement for any HEPS evaluation study (and journal article) to clearly define the benchmark(s) being used and the reasons leading to this selection.

Acknowledgements

Hannah Cloke and Anna Mueller are funded by NERC's Flooding from Intense Rainfall Programme, Project SINATRA: Susceptibility to Intense Rainfall and Flooding grant number NE/K00896X/1.

References

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

Alfieri, L., Pappenberger, F., Wetterhall, F., Haiden, T., Richardson, D., Salamon, P., 2014. Evaluation of ensemble streamflow predictions in Europe. J. Hydrol. 517, 913–922. http://dx.doi.org/10.1016/j.jhydrol.2014.06.035.

Baruth, B., Genovese, G., Leo, O., Boogard, H., te Roller, J.A., van Diepen, K., 2007. CGMS version 9.2 User Manual and Technical Documentation, JRC Scientific and Technical Report, ISBN 978-92-79-06995-6, OPOCE, Luxembourg.

Berthet, L., Andréassian, V., Perrin, C., Javelle, P., 2009. How crucial is it to account for the antecedent moisture conditions in flood forecasting? Comparison of event-based and continuous approaches on 178 catchments. Hydrol. Earth Syst. Sci. 13 (6), 819–831.

Bogner, K., Pappenberger, F., 2010. Multiscale error analysis, correction, and predictive uncertainty estimation in a flood forecasting system. Water Resour. Res. 47 (7), W07524. http://dx.doi.org/10.1029/2010WR009137.
Bogner, K., Pappenberger, F., 2011. Multiscale error analysis, correction, and predictive uncertainty estimation in a flood forecasting system. Water Resour. Res., 47
Brown, J., 2013. Comments to HEPEX Science and Challenges: Verification of Ensemble Forecasts. <http://hepex.irstea.fr/hepex-science-and-challenges-verification-of-ensemble-forecasts-44/>.
Brown, J.D., Seo, D.-J., 2010. A nonparametric postprocessor for bias correction of hydrometeorological and hydrologic ensemble forecasts. J. Hydrometeorol. 11 (3), 642–665.
Brown, J.D., Demargne, J., Seo, D.-J., Liu, Y., 2010. The ensemble verification system (EVS): a software tool for verifying ensemble forecasts of hydrometeorological and hydrologic variables at discrete locations. Environ. Model. Softw. 25 (7), 854–872.
Cloke, H.L., Pappenberger, F., 2008. Evaluating forecasts of extreme events for hydrological applications: an approach for screening unfamiliar performance measures. Meterol. Appl. 15 (1), 181–197.
Cloke, H.L., Pappenberger, F., 2009. Ensemble flood forecasting: a review. J. Hydrol. 375 (3–4), 613–626.
Cloke, H.L., Pappenberger, F., van Andel, S.J., Schaake, J., Thielen, J., Ramos, M.-H., 2013a. Hydrological ensemble prediction systems preface. Hydrol. Process. 27 (1), 1–4.
Cloke, H.L., Wetterhall, F., He, Y., Freer, J.E., Pappenberger, F., 2013b. Modelling climate impact on floods with ensemble climate projections. Q. J. R. Meteorol. Soc. 139 (671), 282–297.
Dawson, C.W., Abrahart, R.J., See, L.M., 2007. HydroTest: a web-based toolbox of evaluation metrics for the standardised assessment of hydrological forecasts. Environ. Modell. Softw. 22, 1034–1052.
Demargne, J., Brown, J., 2013. HEPEX Science and Challenges: Verification of Ensemble Forecasts. <http://hepex.irstea.fr/hepex-science-and-challenges-verification-of-ensemble-forecasts/>.
Demargne, J., Mullusky, M., Werner, K., Adams, T., Lindsey, S., Schwein, N., Marosi, W., Welles, E., 2009. Application of forecast verification science to operational river forecasting in the U.S. national weather service. Bull. Am. Meteorol. Soc. 90 (6), 779–784.
Demargne, J., Wu, L., Regonda, S.K., Brown, J.D., Lee, H., He, M., Seo, D.-J., Hartman, R., Herr, H.D., Fresch, M., Schaake, J., Zhu, Y., 2014. The science of NOAA's operational hydrologic ensemble forecast service. Bull. Amer. Meteor. Soc. 95, 79–98.
Demeritt, D., Nobert, S., Cloke, H.L., Pappenberger, F., 2013. The European flood alert system and the communication, perception, and use of ensemble predictions for operational flood risk management. Hydrol. Process. 27 (1), 147–157.
Ebert, E., Wilson, L., Weigel, A., Mittermaier, M., Nurmi, P., Gill, P., Göber, M., Joslyn, S., Brown, B., Fowler, T., Watkins, A., 2013. Progress and challenges in forecast verification. Met. Apps. 20, 130–139. http://dx.doi.org/10.1002/met.1392.
Ewen, J., 2011. Hydrograph matching method for measuring model performance. J. Hydrol. 408 (1–2), 178–187.
Fewtrell, T.J., Duncan, A., Sampson, C.C., Neal, J.C., Bates, P.D., 2011. Benchmarking urban flood models of varying complexity and scale using high resolution terrestrial LiDAR data. Phys. Chem. Earth 36 (7–8), 281–291.
Fundel, F., Jörg-Hess, S., Zappa, M., 2012. Long-range hydrometeorological ensemble predictions of drought parameters. Hydrol. Earth Syst. Sci. Discuss. 9, 6857–6887.
Fundel, F., Jörg-Hess, S., Zappa, M., 2013. Monthly hydrometeorological ensemble prediction of streamflow droughts and corresponding drought indices. Hydrol. Earth Syst. Sci. 17 (1), 395–407.
Garrick, M., Cunnane, C., Nash, J.E., 1978. Criterion of efficiency for rainfall–runoff models. J. Hydrol. 36 (3–4), 375–381.
Gneiting, T., Ranjan, R., 2011. Comparing density forecasts using threshold- and quantile-weighted scoring rules. J. Business Econ. Statist. 29 (3), 411–422.
Gordon, N., Shaykewich, J., World Meteorological Organization, 2000. Guidelines on Performance Assessment of Public Weather Services. WMO.
Hamill, T.M., Juras, J., 2006. Measuring forecast skill: is it real skill or is it the varying climatology? Q. J. R. Meteorol. Soc. 132 (621C), 2905–2923.
Hersbach, H., 2000. Decomposition of the continuous ranked probability score for ensemble prediction systems. Weather Forecasting 15 (5), 559–570.
Hopson, T.M., Webster, P.J., 2010. A 1–10-day ensemble forecasting scheme for the major river basins of Bangladesh: forecasting severe floods of 2003–07. J. Hydrometeor 11, 618–641. http://dx.doi.org/10.1175/2009JHM1006.1.
Jaun, S., Ahrens, B., 2009. Evaluation of a probabilistic hydrometeorological forecast system. Hydrol. Earth Syst. Sci. 13 (7), 1031–1043.
Jolliffe, J.T., Stephenson, D.B. (Eds.), 2011. Forecast Verification: A Practioner's Guide in Atmospheric Science. Wiley.
Kachroo, R.K., 1992. River flow forecasting: 1 a discussion of the principles. J. Hydrol. 133 (1–2), 1–15.
Kiese, S., Pappenberger, A., Friess, W., Mahler, H.C., 2010. Equilibrium studies of protein aggregates and homogeneous nucleation in protein formulation. J. Pharm. Sci. 99 (2), 632–644.
Krause, P., Boyle, D.P., Bäse, F., 2005. Comparison of different efficiency criteria for hydrological model assessment. Adv. Geosci. 5, 89–97.
Lerch, S., 2012. Verification of probabilistic forecasts for rare and extreme events. Ruprecht-Karls-Universitaet Heidelberg, Heidelberg.
Lerch, S., Thorarinsdottir, T.L., 2013. Comparison of nonhomogeneous regression models for probabilistic wind speed forecasting. Tellus A 65, 21206.
Luo, Y.Q., Randerson, J.T., Abramowitz, G., Bacour, C., Blyth, E., Carvalhais, N., Ciais, P., Dalmonech, D., Fisher, J.B., Fisher, R., Friedlingstein, P., Hibbard, K., Hoffman, F., Huntzinger, D., Jones, C.D., Koven, C., Lawrence, D., Li, D.J., Mahecha, M., Niu, S.L., Norby, R., Piao, S.L., Qi, X., Peylin, P., Prentice, I.C., Riley, W., Reichstein, M., Schwalm, C., Wang, Y.P., Xia, J.Y., Zaehle, S., Zhou, X.H., 2012. A framework for benchmarking land models. Biogeosciences 9 (10), 3857–3874.
Martinec, J., Rango, A., 1989. Merits of statistical criteria for the performance of hydrological models. Water Resour. Bull. 25 (2), 421–432.
Marty, R., Zin, I., Obled, C., Bontron, G., Djerboua, A., 2012. Toward real-time daily PQPF by an analogue sorting approach: application to flash-flood catchments. J. Appl. Meteorol. Clim. 51 (3), 505–520.
Murphy, A.H., Winkler, R.L., 1987. A general framework for forecasts verification. Mon. Weather Rev. 115 (7), 1330–1338.
Nicolle, P., Pushpalatha, R., Perrin, C., François, D., Thiéry, D., Mathevet, T., Le Lay, M., Besson, F., Soubeyroux, J.M., Viel, C., Regimbeau, F., Andréassian, V., Maugis, P., Augeard, B., Morice, E., 2013. Benchmarking hydrological models for low-flow simulation and forecasting on French catchments. Hydrol. Earth Syst. Sci. Discuss. 10 (11), 13979–14040.
Obled, C., Bontron, G., Garçon, R., 2002. Quantitative precipitation forecasts: a statistical adaptation of model outputs through an analogues sorting approach. Atmos. Res. 63, 303–324.
Pagano, T.C., 2013. Evaluation of mekong river commission operational flood forecasts, 2000–2012. Hydrol. Earth Syst. Sci. Discuss. 10 (11), 14433–14461.
Pagano, T.C., Shrestha, D.L., Wang, Q.J., Robertson, D.E., Hapuarachchi, P., 2013. Ensemble dressing for hydrological applications. Hydrol. Process. 27 (1), 106–116.
Pappenberger, F., Beven, K., 2004. Functional classification and evaluation of hydrographs based on multicomponent mapping. J. River Basin Manage. 2 (2), 89–100.
Pappenberger, F., Buizza, R., 2009. The skill of ECMWF precipitation and temperature predictions in the Danube basin as forcings of hydrological models. Weather Forecasting 24 (3), 749–766.
Pappenberger, F., Bartholmes, J., Thielen, J., Cloke, H.L., Buizza, R., de Roo, A., 2008. New dimensions in early flood warning across the globe using grand-ensemble weather predictions. Geophys. Res. Lett. 35 (10).
Pappenberger, F., Bogner, K., Wetterhall, F., He, Y., Cloke, H.L., Thielen, J., 2011a. Forecast convergence score: a forecaster's approach to analysing hydro-meteorological forecast systems. Adv. Geosci. 29, 27–32.
Pappenberger, F., Thielen, J., Del Medico, M., 2011b. The impact of weather forecast improvements on large scale hydrology: analysing a decade of forecasts of the European flood alert system. Hydrol. Process. 25 (7), 1091–1113.
Pappenberger, F., Wetterhall, F., Dutra, E., Di Giuseppe, F., Bogner, K., Alfieri, L., Cloke, H.L., 2013. Seamless forecasting of extreme events on a global scale. In: Boegh, E. et al. (Eds.), Climate and Land Surface Changes in Hydrology. IAHS Publication, pp. 3–10.
Perrin, C., Andreassian, V., Michel, C., 2006. Simple benchmark models as a basis for model efficiency criteria. Arch. Fuer Hydrobiol. Suppl. 161 (1–2), 221–244.
Pushpalatha, R., Perrin, C., Le Moine, N., Andreassian, V., 2012. A review of efficiency criteria suitable for evaluating low-flow simulations. J. Hydrol. 420, 171–182.
Radanovics, S., Vidal, J.P., Sauquet, E., Ben Daoud, A., Bontron, G., 2013. Optimising predictor domains for spatially coherent precipitation downscaling. Hydrol. Earth Syst. Sci. Discuss. 10 (4), 4015–4061.
Ramos, M.-H., Bartholmes, J., Thielen-del Pozo, J., 2007. Development of decision support products based on ensemble forecasts in the European flood alert system. Atmos. Sci. Lett. 8 (4), 113–119.
Ramos, M.H., Mathevet, T., Thielen, J., Pappenberger, F., 2010. Communicating uncertainty in hydro-meteorological forecasts: mission impossible? Meteorol. Appl. 17 (2), 223–2258.
Ramos, M.H., van Andel, S.J., Pappenberger, F., 2013. Do probabilistic forecasts lead to better decisions? Hydrol. Earth Syst. Sci. 17, 2219–2232. http://dx.doi.org/10.5194/hess-17-2219-2013.
Randrianasolo, A., Ramos, M.H., Thirel, G., Andréassian, V., Martin, E., 2010. Comparing the scores of hydrological ensemble forecasts issued by two different hydrological models. Atmos. Sci. Lett. 11 (2), 100–107.
Ritter, A., Munoz-Carpena, R., 2013. Performance evaluation of hydrological models: statistical significance for reducing subjectivity in goodness-of-fit assessments. J. Hydrol. 480, 33–45.
Rodwell, M.J., Richardson, D.S., Hewson, T.D., Haiden, T., 2010. A new equitable score suitable for verifying precipitation in numerical weather prediction. Q. J. R. Meteorol. Soc. 136 (650), 1344–1363.
Romanowicz, R.J., Young, P.C., Beven, K.J., Pappenberger, F., 2008. A data based mechanistic approach to nonlinear flood routing and adaptive flood level forecasting. Adv. Water Resour. 31 (8), 1048–1056.
Rykiel, E.J., 1996. Testing ecological models: the meaning of validation. Ecol. Model. 90 (3), 229–244.
Schaefli, B., Gupta, H.V., 2007. Do Nash values have value? Hydrol. Process. 21 (15), 2075–2080.
Seibert, J., 2001. On the need for benchmarks in hydrological modelling. Hydrol. Process. 15 (6), 1063–1064.
Stedinger, J.R., Kim, Y.-O., 2010. Probabilities for ensemble forecasts reflecting climate information. J. Hydrol. 391, 9–23.
Thielen, J., Bartholmes, J., Ramos, M.H., de Roo, A., 2009a. The European flood alert system – Part 1: concept and development. Hydrol. Earth Syst. Sci. 13 (2), 125–140.
Thielen, J., Bogner, K., Pappenberger, F., Kalas, M., del Medico, M., de Roo, A., 2009b. Monthly-, medium-, and short-range flood warning: testing the limits of predictability. Meteorol. Appl. 16 (1), 77–90.

Trinh, B.N., Thielen-del Pozo, J., Thirel, G., 2013. The reduction continuous rank probability score for evaluating discharge forecasts from hydrological ensemble prediction systems. Atmos. Sci. Lett. 14 (2), 61–65.
van der Knijff, J.M., Younis, J., 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.
Węglarczyk, S., 1998. The interdependence and applicability of some statistical quality measures for hydrological models. J. Hydrol. 206 (1–2), 98–103.
Weijs, S.V., van de Giesen, N., 2013. An information-theoretical perspective on weighted ensemble forecasts. J. Hydrol. 498, 177–190.
Weijs, S.V., Schoups, G., van de Giesen, N., 2010. Why hydrological predictions should be evaluated using information theory. Hydrol. Earth Syst. Sci. 14, 2545–2558. http://dx.doi.org/10.5194/hess-14-2545-2010.
Wetterhall, F., Pappenberger, F., 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., Holubecka, M., 2013. Forecasters priorities for improving probabilistic flood forecasts. Hydrol. Earth Syst. Sci. Discuss. 10 (2), 2215–2242.
Zalachori, I., Ramos, M.H., Garçon, R., Mathevet, T., Gailhard, J., 2012. Statistical processing of forecasts for hydrological ensemble prediction: a comparative study of different bias correction strategies. Adv. Sci. Res. 8, 135–141.