Hydrol. Earth Syst. Sci., 13, 141–153, 2009
www.hydrol-earth-syst-sci.net/13/141/2009/

Hydrology and Earth System Sciences

The european flood alert system EFAS – Part 2: Statistical skill assessment of probabilistic and deterministic operational forecasts

J. C. Bartholmes[1,*], J. Thielen[1], M. H. Ramos[1,], and S. Gentilini[1]**

[1]EC, Joint Research Centre, Institute for Environment and Sustainability, Via Fermi 2749, 21027 Ispra (VA), Italy
[*]now at: EC, Joint Research Centre, Scientific EC Work Programme Unit, Square de Meeus 8, 1050 Brussels, Belgium
[**]now at: CEMAGREF, Parc de Tourvoie 44, 92163 Antony Cedex, France

Received: 14 December 2007 – Published in Hydrol. Earth Syst. Sci. Discuss.: 6 February 2008
Revised: 12 January 2009 – Accepted: 12 January 2009 – Published: 5 February 2009

Abstract. Since 2005 the European Flood Alert System (EFAS) has been producing probabilistic hydrological forecasts in pre-operational mode at the Joint Research Centre (JRC) of the European Commission. EFAS aims at increasing preparedness for floods in trans-national European river basins by providing medium-range deterministic and probabilistic flood forecasting information, from 3 to 10 days in advance, to national hydro-meteorological services.

This paper is Part 2 of a study presenting the development and skill assessment of EFAS. In Part 1, the scientific approach adopted in the development of the system has been presented, as well as its basic principles and forecast products. In the present article, two years of existing operational EFAS forecasts are statistically assessed and the skill of EFAS forecasts is analysed with several skill scores. The analysis is based on the comparison of threshold exceedances between proxy-observed and forecasted discharges. Skill is assessed both with and without taking into account the persistence of the forecasted signal during consecutive forecasts.

Skill assessment approaches are mostly adopted from meteorology and the analysis also compares probabilistic and deterministic aspects of EFAS. Furthermore, the utility of different skill scores is discussed and their strengths and shortcomings illustrated. The analysis shows the benefit of incorporating past forecasts in the probability analysis, for medium-range forecasts, which effectively increases the skill of the forecasts.

1 Introduction

The increasing awareness that fluvial floods in Europe constitute a non-negligible threat to the well-being of the population, prompted the European Commission to trigger the development of a European Flood Alert System (EFAS) in 2003. After several severe flood events of trans-national dimensions which struck Europe (EEA, 2003), the European Commission initiated the development of a system (i.e. EFAS) that could provide medium-range pre-alerts for the trans-national river basins in Europe, and could thus raise preparedness prior to a possible upcoming flood event.

The objective of this work is to assess the skill of this specific operational hydrological forecasting system. Skill studies regarding the use of meteorological ensemble forecasts for producing stream-flow predictions are reported in the literature (Franz et al., 2003; Clark and Hay, 2004; Roulin and Vannitsem, 2005; Bartholmes and Todini, 2005; Roulin, 2007), but none to this extent (at least to the knowledge of the authors), for systems of similar size. This makes the comparison of skill scores obtained from the analysis of EFAS forecasts with skill scores from other studies difficult.

Firstly, a rigorous comparison of EFAS results with results obtained from regional hydrological forecasting systems is not straightforward, mainly due to the large differences between the way inputs (weather forecasts) and outputs (discharge forecasts and alert level exceedances) are processed. The differences in space and time scales, as well as in the setup of the hydrological model, need also to be assessed before any statistical comparative analysis. Currently, no scientific project has been dedicated to such an assessment, which implies the set up of a rigorous protocol for a sounded based comparison (i.e., same input data, same

Correspondence to: J. C. Bartholmes
(jens.bartholmes@ec.europa.eu)

time period of analysis, same verification skill scores, etc.). Within the EFAS project, preliminary comparative analysis of the alert levels issued during a flood event has only been performed for individual case studies and been recently published (Kalas et al., 2008; Younis et al., 2008).

Secondly, the statistical skill assessment of EFAS forecasts here presented is unique in the sense that the scores have been calculated for the entire Europe, for each river pixel, over a period of 2 years, with a minimum of input data, and with exclusively probabilistic skill scores. Since, typically, skill scores, such as the Brier Skill Score (BSS), depend to a large degree on the chosen climatology used as reference, without having comparable input data, the comparison of skill scores from different studies remains difficult.

Part 1 (Thielen et al., 2009) of this publication focussed on the larger context within which EFAS has been developed. It presents the general and technical set-up and describes the methodologies, input data and products of EFAS. The present paper (Part 2) concerns the assessment of EFAS overall forecasting skill over a full two-year period of existing EFAS operational hydrological forecasts.

Literature on skill scores dates back more than 120 years (Peirce, 1884; Gilbert, 1884) and often "the wheel has been reinvented" leading also to confusing double-naming of the same scores (Baldwin, 2004; Stephenson, 2000). Not all skill scores are equally suited for the skill assessment of a given forecasting system and there is no single skill score that can convey all necessary information. Therefore sets of skill scores are normally used to cover a wider spectrum of properties (Baldwin, 2004). However, care should be taken to avoid being "engulfed in an avalanche of numbers that are never used" (Gordon and Shaykewich, 2000). For the present study on EFAS skill assessment, a first set of skill measures was selected. This set was subsequently reduced further taking into account the shortcomings of certain measures that arose during the analysis.

Finally, the scores used for this study are mainly described in the literature in the context of meteorological applications, but, as they deal with continuous variables that are transformed – using certain thresholds – into dichotomous events, it was assumed that they are as well applicable to discharge threshold exceedances as used in EFAS. The differences in terminology between hydrology and meteorology are a well known problem (see Pappenberger et al., 2008) and, where deemed appropriate, additional explanations are given in this paper.

The main reasons leading to the choice of certain skill scores are outlined in the following. The first essential aspect was that the chosen skill measures had to be equally representative for different climatic regimes that can be found over Europe. Regarding this aspect, McBride and Ebert (2000) promote the Hanssen-Kuipers skill score[1] (Hanssen and Kuipers, 1965) as being independent of the particular distribution of observed precipitation. Similar characteristics are attributed by Stephenson (2000) to the "odds ratio". Göber (2004) goes even further claiming that only the odds ratio "enables a fair comparison of categorical forecasts for different years, regions, events". Other scores like the Critical Success Index or CSI (also Thread Score / TS and Gilbert Skill Score/GSS) (Gilbert, 1884, Schaefer, 1990) or the Heidke (1926) skill score strongly depend on the frequency of certain (precipitation) events (Ebert and McBride, 1997). Likewise, Wilson (2001) stated that the Equitable Threat Score or ETS (Schaefer, 1990) is a "reasonable" score but that it is not independent from the observed event distribution.

The second important factor considered here in the choice of skill measures was that a skill score should be as little as possible influenced by a bias in the forecast, i.e. that it should be insensitive to over- or under-forecasting. For example, Mason (1989) showed that the CSI is highly sensitive to such forecasting biases, whereas the odds ratio was found to be quite insensitive to them (Baldwin, 2004). In this context, Gandin and Murphy (1992) defined a score as *equitable* if random and constant forecasts result in the same score value, and if they do not encourage over- or under-forecasting as the score is maximised for unbiased forecasts (bias=1). Scores like percentage correct (PC) were found to be not equitable as they could be easily "improved" by over-forecasting. It was, however, stated by Marzban (1998) that there is no such a thing as a strictly equitable skill score when it comes to forecasting extreme events.

Some of these properties and the final choice of skill scores for this study will be discussed in Sect. 3. An extensive review on skill scores can be found in Stanski et al. (1989), as well as in the works of Murphy (1996, 1997).

This study looks at the past performance of a forecasting system, but also aims to improve the future performance of such a system by making it possible to incorporate the past experience into the current forecasting procedure, thereby allowing the forecaster to better estimate the probability of a current forecast. This is particularly important for a flood alert system that covers a heterogeneous area with several river basins and for which local expertise is not always at hand.

This paper is structured in the following way: in Sect. 2, a short description of the data used is given, followed by the description of the methodology in Sect. 3. Results are presented in Sect. 4 and discussed in Section 5, followed by the final conclusions in Sect. 6.

2 Data

The deterministic forecasts of the German National Weather Service (hereafter, DWD) and the deterministic (hereafter, EUD) as well as the probabilistic (51 members, hereafter,

[1] Rediscovered Peirce (1884) skill score also referred to as True Skill Statistic TSS (Flueck, 1987).

EUE or EPS for Ensemble Prediction System) forecasts of the European Centre for Medium-Range Weather Forecasts (ECMWF) are the input data for EFAS. The forecast ranges are 10 days for the ECMWF forecasts and 7 days for the DWD forecasts. In the absence of discharge measurements to set up the initial conditions at the beginning of the forecasts, a proxy for observed discharge is calculated using observed meteorological station data that is provided by the JRC MARSSTAT Unit (internet: http://agrifish.jrc.it). The data used in EFAS are described in more detail in Part 1 (Thielen et al., 2009).

The present analysis is based on operational EFAS discharge forecast data simulated using the 12:00 UTC weather forecast for the full 25-month period of January 2005 to February 2007. Grid cells with upstream area of less than 4000 km^2 are not included in the analysis as previous works (Bartholmes et al., 2006) showed that EFAS skill computed as a function of upstream areas remains similar for areas greater than 4000 km^2. EFAS skill only deteriorates when going below this threshold, which corresponds to the spatial resolution of the meteorological data that were used for the calibration of the forecasting system.

In the pre-operational system, only leadtimes of 3 to 10 days are considered, due to the medium-range pre-alert orientation of EFAS. It is considered that for short leadtimes discharges can be forecasted much better by the national hydrometeorological services. From the technical point of view, it should also be mentioned that for leadtimes shorter than 3 days, the influence of deterministic initial conditions is still important on the probabilistic forecasts, and spread (i.e. uncertainty information) is limited. However, in this study, for completeness of the analysis and research purposes, discharge forecast data for all leadtimes, from 1 to 10 days, were considered.

We also note that in our first efforts to perform the skill assessment of EFAS forecasts (Bartholmes et al., 2006) serious shortcomings due to a lesser quantity of data available (i.e. too few forecasted events) appeared. We consider that in the present study, based on 2 years of data, the results are statistically more significant.

3 Methodology

Skill assessment in deterministic stream-flow hydrology normally compares observed point measurements (discharges at gauging stations) with simulated discharges from hydrological model output. The Nash-Sutcliffe coefficient (Nash and Sutcliffe, 1970) is one of the most used skill scores in hydrologic simulation and gives a measure of the discrepancy between observed and simulated discharges. In order to give meaningful results, this method needs good observed data for every point of interest and a single model hydrograph to compare it against. When it comes to probabilistic skill assessment of hydrologic forecasts this kind of skill score is less useful. Even if it can be used with the mean of a hydrological probabilistic forecast (Mullusky, 2004), in this study it was decided not to adopt the Nash–Sutcliffe coefficient because it does not take into account all the probabilistic aspects of the EFAS forecasts, and thus omits valuable information. Specifically, probabilistic forecast performance cannot be verified in the same direct way as the deterministic one (i.e. comparison of one observed/forecasted value pair per time step). Instead, observed event frequencies have to be compared to forecasted probabilities, thus necessitating long enough time-series for a statistically meaningful evaluation.

In this context, methods adopted from meteorology – where probabilistic forecasts are already more commonly used – are applied here to assess the skill of EFAS operational forecasts. The variables and the tools used in this analysis are explained in the next paragraphs.

3.1 Variable analysed: the threshold exceedance

Analysing the performance of a hydrological forecasting system at European scale by performing the analysis of the full probabilistic distribution of discharges or a cost-benefit analysis of forecasts issued by the system is not possible, as not all the required data is available at this scale. Therefore, in this analysis, EFAS forecasted discharges are not processed as continuous variables but are reduced to binary events of exceeding or not exceeding a threshold. In other words, the maps describing EFAS forecasts analysed here contain only information if a forecasted discharge in a pixel is above or below a certain threshold. To get to these thresholds, a long term simulation (1990–2004) with the EFAS hydrological model LISFLOOD (de Roo, 2000; van der Knijff and de Roo, 2006) was performed (using the same set-up as the operational system) and statistical thresholds were deduced from the simulated discharges for every grid cell. This approach has the advantage that any systematic over- or underprediction of the model is levelled out (see Part 1, Thielen et al., 2009, for more details). The two thresholds (and definitions) that indicate a probability of flooding and thus are most important for EFAS forecasts are:

Severe EFAS alert level (SAL) threshold:

- *very high possibility of flooding, potentially severe flooding expected*

High EFAS alert level (HAL) threshold:

- *high possibility of flooding, bankful conditions or higher expected*

Due to the small number of events with discharges greater than the EFAS SAL during the study period, the results obtained for this threshold were not statistically significant and are thus not shown in this paper, which presents only the results for the HAL events.

Table 1. Contingency table for the forecast verification of given dichotomous events.

		Observed (Proxy)		
		YES	NO	
Forecasted	YES	a (HIT)	b (FA)	a+b
	NO	c (MISS)	d (CR)	c+d
		a+c	b+d	a+b+c+d

3.2 Criterion of persistence

In a medium-range forecasting system like EFAS, a flood event is forecasted, in most cases, several days ahead of the event when it is flagged for the first time by the system. The forecaster has the possibility to adopt a "*wait and see*" attitude until the next forecast(s) and only take an event into closer consideration if it is confirmed, i.e. if it is forecasted persistently. In this study, we tested if the use of a persistence criterion leads also to quantifiable improvements of the forecast quality.

The following definitions for persistence were used:

A forecast is considered as persistent only if the alert threshold exceedance in a river pixel is forecasted continuously on 2 consecutive dates.

In the probabilistic EPS[2]-based forecasts (EUE), the persistence criterion is also linked to a 2nd threshold: a minimum number of EPS members has to be persistently above the alert threshold.

3.3 Contingency tables

Assuming that the position in time of pairs of forecast (f) and observation $(x)[(f_i, x_i),\ i=1, n]$ is negligible, the empirical relative frequency distribution of the analysed sample "captures all relevant information" (Murphy and Winkler, 1987) and the joint distribution of f and x can be presented by a 2×2 contingency table (see Table 1).

For the construction of contingency tables, the forecasted discharges are transformed into dichotomous events (Atger, 2001; Bradley et al., 2004) regarding the following criteria: does a discharge exceed the EFAS high alert threshold? Do at least 5, 10, 20 etc. EPS members (out of 51 members) forecast a discharge exceeding the EFAS high alert threshold? Is the forecast persistent?

For every combination, a contingency table was calculated for each river pixel with an upstream area larger than 4000 km^2. The four fields of the contingency table are illustrated in Table 1. When the persistence criterion is applied, the event is considered as "Forecasted" if and only if it is forecasted continuously on 2 consecutive dates. For example, an event with persistence=20 EPS would be only classified as "Forecasted" if the previous and the present forecasts had at least 20 EPS forecasting a discharge greater than the EFAS high alert threshold.

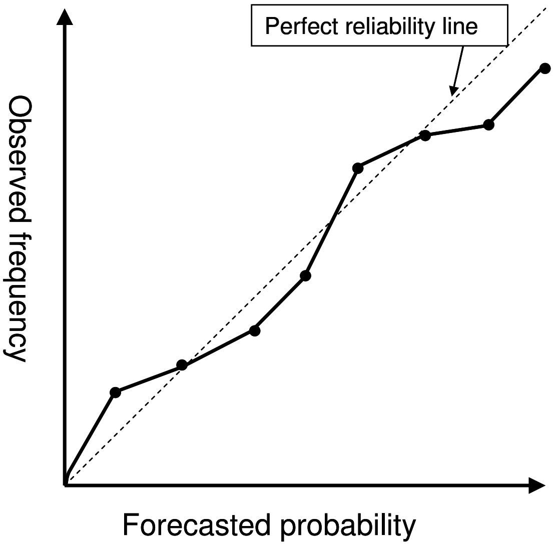

Fig. 1. Schematic reliability diagram.

3.4 Reliability diagram

The reliability diagram (Wilks, 1995) is used to assess the "reliability" of a probabilistic forecast of binary events, i.e. it analyses how forecasted and observed event frequencies compare (see Fig. 1). The forecasted event probabilities of the EFAS EPS forecasts are plotted on the x-axis against the observed event frequencies of the proxy on the y-axis. The closer the data points in this plot are to the 1:1 diagonal line, the more reliable is the forecast. In the case of perfect reliability, a forecast predicting an event with $X\%$ of the ensemble members would have $X\%$ of probability to happen.

3.5 Choice of appropriate skill scores

Taking into account the findings in the literature (see Introduction), the "odds ratio" and the Hanssen-Kuipers score (HK) (both *equitable*) were chosen as skill measures that do not make explicit use of a comparison to a benchmark forecast (Eqs. 1 and 2). As advocated in Stephenson (2000), these two measures are complemented with the Frequency Bias (FB) (Eq. 3).

$$\text{Odds} = \frac{ad}{bc} \text{ range } [0, \infty],\ \text{best} : \infty \tag{1}$$

$$\text{HK} = \frac{a}{a+c} - \frac{b}{b+d} \text{ range } [-1, 1], \text{best} : 1 \tag{2}$$

$$\text{FB} = \frac{a+b}{a+c} \text{ range } [0, \infty],\ \text{best} : 1 \tag{3}$$

with *a, b, c, d* defined as in the contingency table in Table 1.

[2] Ensemble Prediction System.

The Brier skill score (BSS) (Brier, 1950) is widely used in the skill analysis of meteorological probabilistic forecasts. In this study, it was chosen as an inherently probabilistic and *strictly proper* score – a score is "proper" when it is optimized for forecasts that correspond to the best judgement of the forecaster, and it is "strictly proper" if there is only one unique maxima (Murphy and Epstein, 1967). Furthermore, the Brier skill score is (*a*) a highly compressed score, as it directly accounts for the forecast probabilities without necessitating a contingency table for each probability threshold, and (*b*) uses a user-defined benchmark forecast (here, the climatology[3]) (Eqs. 4 and 5):

$$\mathrm{BSS} = 1 - \frac{\mathrm{BS}_f}{\mathrm{BS}_{\mathrm{clim}}} \text{ range } [-\infty, 1], \text{ best} : 1 \tag{4}$$

with

$$\mathrm{BS}_{=} \frac{1}{N} \sum_{1}^{N} (p - o)^2 \text{ range } [0, 1], \text{ best} : 0 \tag{5}$$

where p refers to the probability with which an event is forecasted and o is the binary value of the observation (o=1.0 if the event is observed and o=0.0 if it is not observed). N is the total number of forecast dates. The underscore f denotes the forecast that is analysed, while clim stands for climatology.

As the most intuitive scores, also the probability of detection (POD, Eq. 6) as well as the frequency of hits (FOH) and frequency of misses (FOM) (Eqs. 7, 8) were chosen:

$$\mathrm{POD} = \frac{a}{a + c} \text{ range}[0, 1], \text{ best} : 1 \tag{6}$$

$$\mathrm{FOH} = \frac{a}{a + b} \text{ range}[0, 1], \text{ best} : 1 \tag{7}$$

$$\mathrm{FOM} = \frac{c}{a + c} \text{ range}[0, 1], \text{ best} : 0 \tag{8}$$

4 Results

The results obtained from the two-year skill assessment exercise (2005–2007) are presented first as absolute numbers of the three contingency table fields: "*hits*"(h), "*false alerts*"(f) and "*misses*"(m). As the analysed events (exceedances of the EFAS HAL threshold) can be regarded as rare[4], the fourth field (d in Table 1) – i.e. "positive reject" – is not shown, as it contains numbers roughly two magnitudes bigger than the other three fields, and is of far less interest to this analysis. Additionally, due to the very high number of combinations (and thus contingency tables) only the most representative results are shown.

4.1 Positive effect of persistence

Figure 2 (top left) shows the number of *false alerts* (f) as a function of leadtime for the deterministic (DWD and EUD) and probabilistic (EUE) forecasts (>5EPS, i.e. more than 5 EPS-based simulations giving discharges above EFAS HAL). The effect of using the persistence criterion (full lines) is clearly visible as the number of *false alerts* is reduced by up to 70%. On the contrary, *hits (h)* (Fig. 2 top right) and *misses* (m) (Fig. 2 bottom) are far less influenced by persistence. Actually, the numbers of *hits* and *misses* for the deterministic EFAS forecasts are hardly changed at all and just for EFAS EPS (EUE) the *hits* are reduced while the *misses* are proportionally increased. However, these changes become much less significant for higher numbers of EPS and are insignificant for forecasts with more than 15 (out of 51) EPS-based simulations above EFAS HAL.

When looking at *hits*, *false alerts* and *misses* for a given leadtime, i.e., varying only the minimum number of EPS that have to exceed HAL for the event to be regarded as "Forecasted", the general behaviour of curves is similar to the example in Fig. 3, where the number of occurrences is plotted for a leadtime of 4 days. By increasing the minimum number of EPS-based simulations forecasting discharges above the EFAS high threshold, the number of *false alerts* becomes drastically smaller, the number of *hits* become smaller to a much lesser extent, while the number of *misses* increase proportionally to the decrease in the number of *hits*.

4.1.1 Frequency of hits (FOH) and frequency of misses (FOM)

Figure 4 shows the frequency of hits (FOH) and frequency of misses (FOM) (see Eqs. 7 and 8) for persistent deterministic forecasts and for 15 persistent EPS members (top) as well as for 5 and 35 persistent EPS members (bottom) over the HAL threshold. It can be seen that the FOM for 15 persistent EPS members is lower (higher skill) than the FOM of the deterministic forecasts and FOH of EPS is higher (higher skill) than the FOH of the deterministic forecasts. The FOM for the deterministic forecasts is always higher than 0.5 (dotted line where number of *misses*=number of *hits, 1:1*), which means that in this case there are always more *misses* than *hits*. When waiting for at least 15 EPS members to forecast an HAL threshold exceedance, the FOM of EPS is also always higher than 0.5 (Fig. 4, top). For lower numbers of EPS members, the FOM becomes lower (higher skill), while the FOH for 5 persistent EPS (Fig. 4, bottom) is as low as for the lowest value obtained with the deterministic forecasts. The FOM values of the probabilistic (EUE) forecasts become similar to the deterministic ones (DWD and EUD) when waiting for at least 35 EPS members to exceed the HAL threshold (Fig. 4, bottom).

[3] Climatology: Sample mean frequency of the event computed using long-term statistics. Here, the historical frequency of an event to exceed the HAL threshold.

[4] EFAS HAL threshold is defined – for each grid cell separately – as the discharge above the 99th percentile of ranked long term daily discharges, see also Part 1 (Thielen et al., 2009).

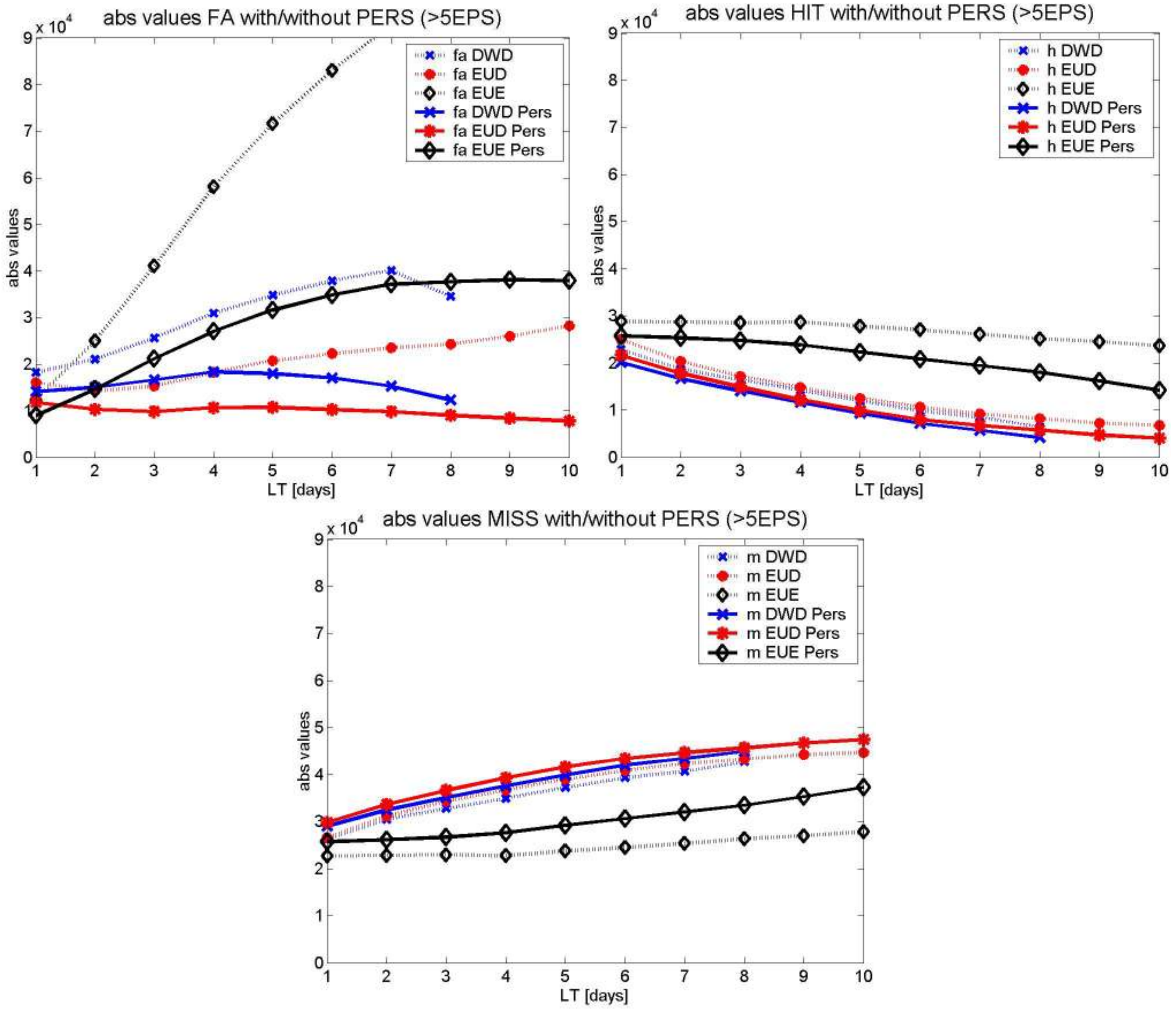

Fig. 2. Absolute numbers of "false alerts f" (top left),"hits h" (top right) and "misses m" (bottom) for deterministic forecasts (DWD and EUD) and at least 5 EPS>HAL (EUE) over leadtime, with (full line) and without (dottetd line) persistence.

4.1.2 Frequency bias

The frequency bias (FB) (Eq. 3) of EFAS forecasts is reported in Fig. 5. The FB values for the deterministic DWD-based forecasts are higher than the FB values for the EUD-based forecasts (ECMWF). The lowest FB results were for the probabilistic EUE-based forecasts. In general, the FB values calculated with the persistence criterion (full lines) are significantly lower than the ones calculated without taking persistence into consideration (dotted lines). Furthermore, Fig. 5 shows that there is a general tendency to over-forecast (i.e. forecast more events than the proxy-observed ones) as the bias value is greater than one. For 20 persistent EPS members, the bias is around 1 (i.e., no bias, indicated by the "bias=1" line in Fig. 5) for the first forecast days of leadtime, but it drops below 1 after day 5. For 5 persistent EPS members, the bias stays between 1.4 and 1.1. For the deterministic forecasts, DWD and EUD-based forecasts with persistence have bias values between 1.3 and 2.5.

4.1.3 Brier skill score

When using the Brier skill score (BSS) it has to be kept in mind that the interpretation of BSS values can be very sensitive to the choice of the reference climatology (Hamill and Jura, 2006). To check this influence on the results of the present analysis, two reference climatologies were compared. On the one hand, the probability of having discharges greater than EFAS high alert threshold (HAL) was set to P_{clim}= 0.00, following the suggestions of Legg and Mylne (2004), who proposed this climatology for rare events. On the other hand, a climatology of P_{clim}= 0.01 was used

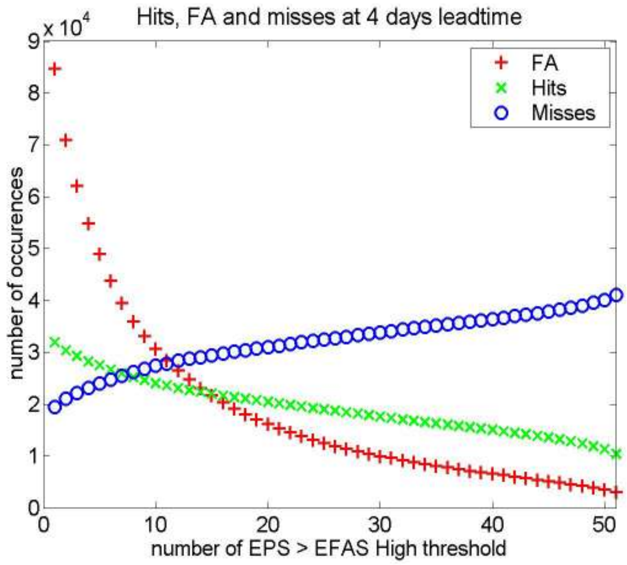

Fig. 3. Absolute numbers reporting the three contingency table fields "hits"(h [x]), "false alerts"(f [+]) and "misses"(m [o]) for at least 5 EPS in the previous forecasts at leadtime 4 days .

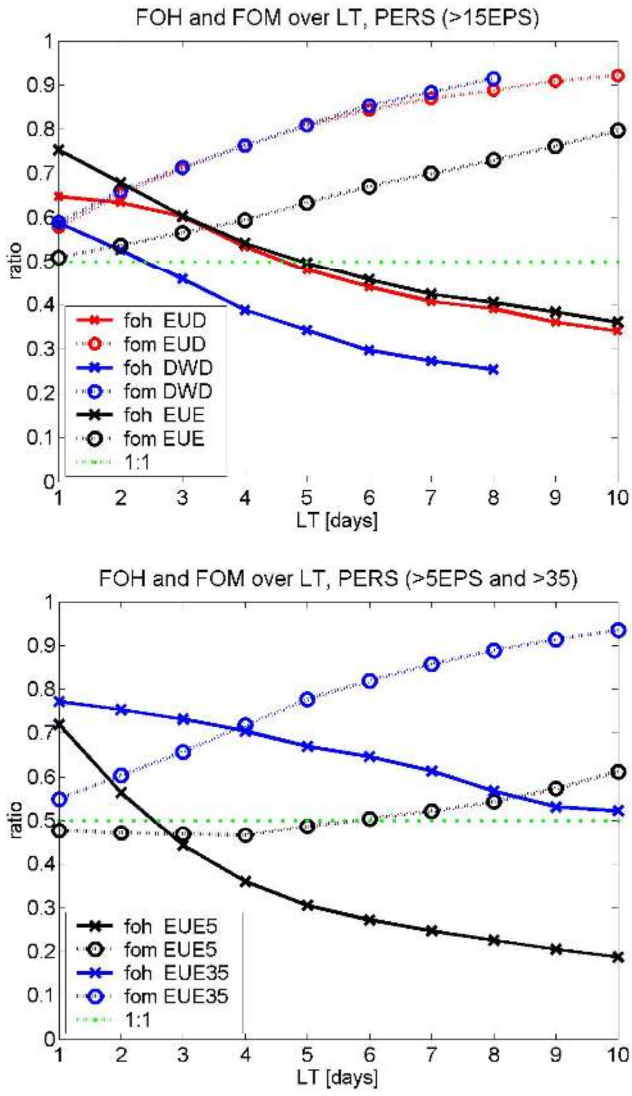

Fig. 4. Frequency of hits (FOH) [x] and frequency of misses (FOM) [o] for deterministic persistence and 15 EPS persistent (top) and 5 and 35 EPS persistent (bottom). 1:1 line means number of misses=number of hits (for FOM) and number of false alerts=number of hits (for FOH).

which corresponds to the empirical event frequency with which the EFAS HAL thresholds (threshold discharge not exceeded in 99% of the cases) were calculated. The results obtained for these two reference climatologies were not significantly different and hence all the diagrams presented hereafter show BSS using P_{clim}= 0.01.

Results for the BSS median values are reported in Fig. 6. The deterministic forecasts without the persistence criterion show no skill compared to the reference (climatology), i.e. the BSS is zero. Also, when considering persistence, they show only very little skill in the first days of leadtime. From Fig. 6, one can see that, with persistence, the Brier skill score stays around 0.0, and, without persistence, it drops steeply after some days of leadtime. The BSS values for the deterministic DWD- and EUD-based forecasts become negative after leadtimes of 3 days and 5 days, respectively. The BSS median for the EPS-based forecasts (EUE) shows the highest skill when persistence is not considered and, for all leadtimes, it shows higher skill than the deterministic forecasts.

Figure 7 shows the relative frequency of EPS BSS values for leadtimes 3, 6 and 10 days in bins with size 0.2: on the top, with no persistence and, on the bottom, with persistence of 20 EPS. When looking at these relative frequencies of the EPS BSS (Fig. 7), one can see that persistence increases the relative frequency of the positive BSS values in the lower skill range (BSS values smaller than 0.4). The increase in relative frequencies of the higher BSS values (0.8–1.0) in the case without persistence (Fig. 7, top) can be regarded as an artefact of the small event probability: if in a pixel nothing is observed in 2 years, but the exceedance of HAL with 1 EPS is (wrongly) forecasted 5 times in the same period, the BSS value is >0.95 for this pixel. This kind of noise is eliminated by the persistence criterion as can be clearly seen in Fig. 7 (bottom).

The spatial distribution of BSS values with persistence (>5EPS) is shown in Fig. 8. For better visibility, distinction is only made between BSS values below zero (red) and above zero (green, higher skill than the climatology). The results obtained when considering no persistence (not shown) are very similar. Basically, there is no recognizable spatial pattern and the skill differs largely from river stretch to river stretch. For example, for the Elbe River, the BSS is very high (with values above 0.5) mostly in its downstream reaches, while for the Danube River, BSS values are between 0.2 and

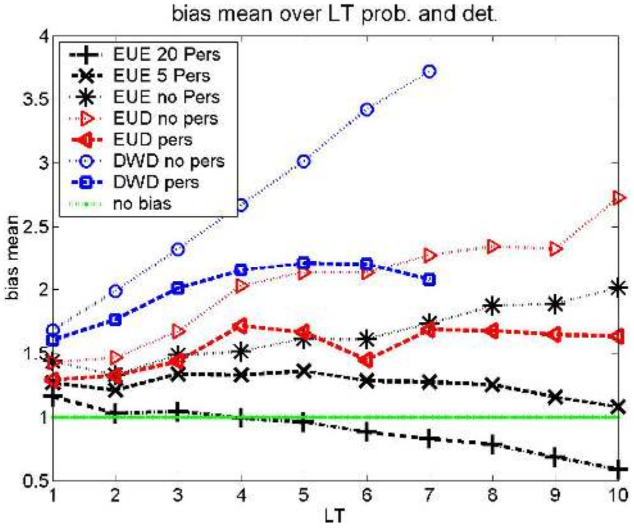

Fig. 5. Frequency bias over leadtime for EUD, DWD and EUE with no, 5 and 20 persistent EPS.

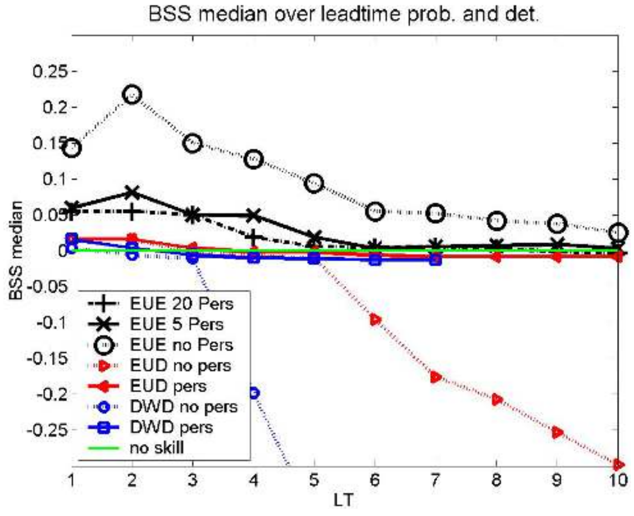

Fig. 6. BSS median over leadtime. For deterministic forecasts with and without persistence and for the EUE with 0, 5 and 20 persistent EPS.

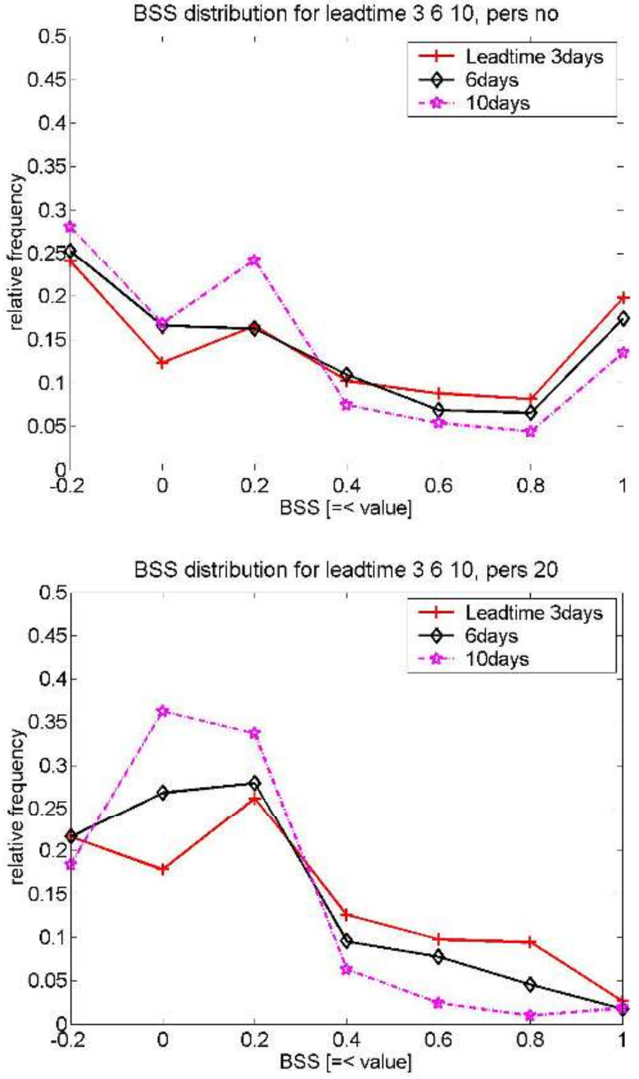

Fig. 7. Relative frequency distribution of Brier skill score (BSS) values (0=no skill, 1=perfect forecast). 1 curve per leadtime. Top no persistence, bottom 20 EPS persistent.

0.4 for upper and lower Danube (and tributaries) and almost zero or even negative for middle Danube and Tisza (Hungarian tributary).

4.1.4 POD and HK

Figure 9 shows the POD (Eq. 6), the POFD (second term of HK/TSS in Eq. 2) and the Hanssen-Kuipers skill score (HK or TSS). It can be seen that the POFD is very close to zero, while the POD and the HK basically take the same values. This is a consequence of the fact that for relatively rare events, like the exceedances of the EFAS HAL threshold, the HK, which is the difference between POD and POFD, is completely dominated by the POD (as the POFD is almost zero due to the high number of "positive rejects"). These skill scores all steadily decrease with leadtime. There is almost no difference in performance between the deterministic forecasts (based on DWD and EUD), while the scores calculated for the EPS-based forecasts (EUE) indicate a much higher skill. Besides, the skill of EUE for these skill scores is highest when persistence is not considered, and goes down with increasing persistence thresholds (number of EPS-based simulations exceeding EFAS HAL). The values of these scores come close to the deterministic ones at ca. 30 persistent EPS simulations above EFAS HAL. This is due to the increasing number of *misses* and decreasing number of *hits* for higher persistence thresholds.

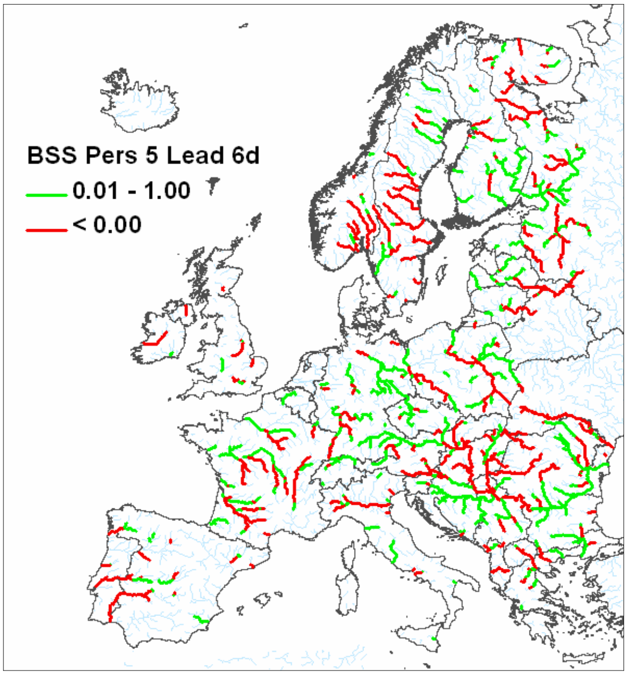

Fig. 8. Areal distribution of BSS median values for 6 days of leadtime and persistence (>5 EPS)

A different picture is obtained for the skill expressed as odds (Fig. 10): the decrease with leadtime for EUE is steeper than that observed for the POD skill score and, with increasing persistence thresholds of EPS-based forecasts, the skill score yields higher (i.e. better) values.

4.2 Reliability diagram

The reliability diagrams obtained from EFAS forecasts for leadtimes 3, 6 and 10 days are shown in Fig. 11. The three histograms presented below the reliability diagrams show the numbers of hits that correspond to the respective relative frequency in the reliability diagram for the leadtimes 3, 6 and 10 days. These histograms show that every data point in the reliability diagram is calculated with at least 250 pixels that had a hit.

It can be seen that for most EPS thresholds the results for EFAS hydrological forecasts are below the diagonal (perfect reliability), meaning that during the study period the EFAS forecasts were over-predicting – i.e. predicting a higher probability than the actually observed frequency of occurrence. The probability of a hit with 51 EPS members predicting a discharge greater than EFAS HAL is only 80%, and, with 48 EPS members, it is less than 60%.

In Fig. 11, the notion of persistence in the EPS threshold refers only to the number of EPS simulations above EFAS HAL in the previous forecast, when the actual forecast has at least 1 EPS member above HAL. This choice reveals a result that would have been obscured if we had omitted the results for EPS numbers smaller than the persistence threshold. Namely, with increasing the threshold of EPS exceeding HAL in the previous forecast, the probability to have a hit becomes higher for the lower EPS numbers, leading to the tendency to under-predict (as observed in the reliability diagram of Fig. 11 bottom). Actually, if the previous forecast predicted 20 EPS>HAL, the probability of a hit with 3 days leadtime is around 30%, no matter if the current forecast is 1 or 25 EPS>HAL (see lower diagram in Fig. 11).

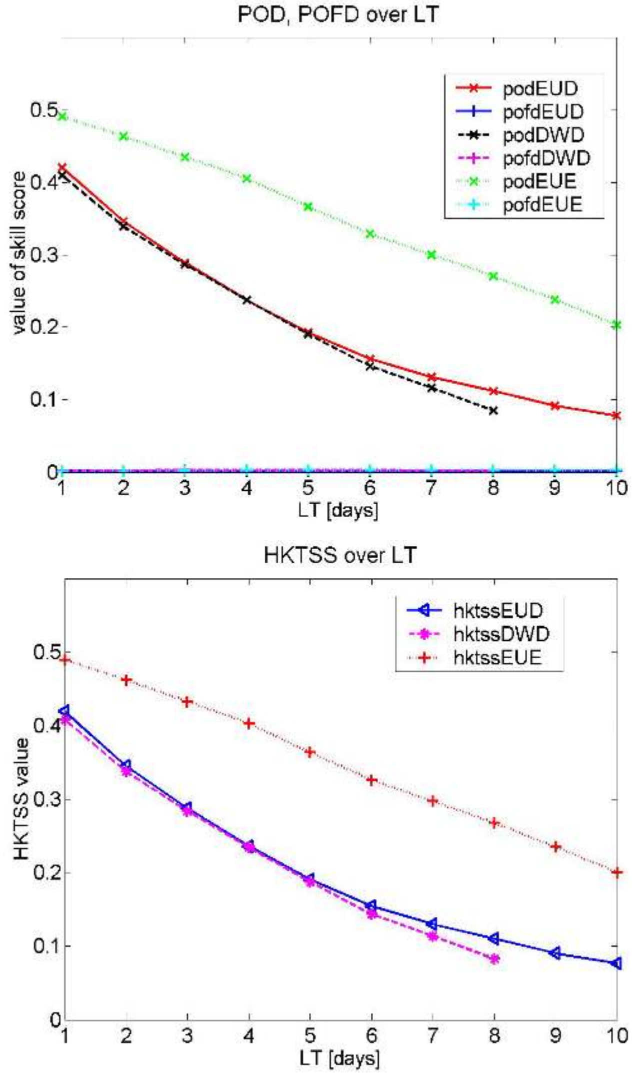

Fig. 9. POD and POFD (top) and Hanssen-Kuipers (HKTSS) (bottom), for both deterministic forecasts based on EUD and DWD and for 15 persistent EPS (EUE).

5 Discussion

Assessing the skill of EFAS forecasts with different skill measures shows different tendencies in the results when taking or not taking into account a criterion of persistence of

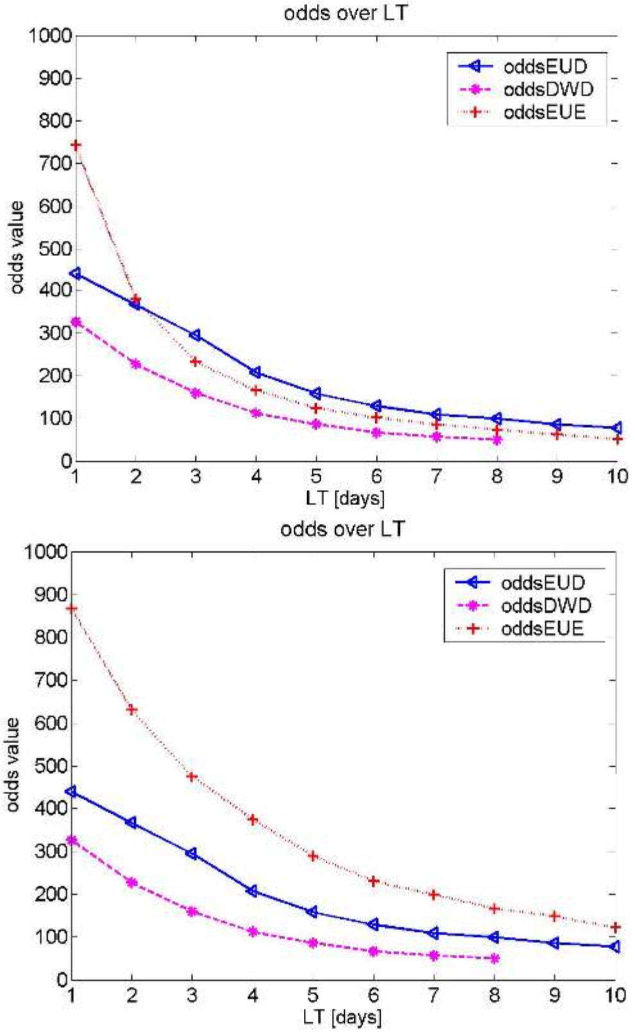

Fig. 10. Skill measures of odds over leadtime for EUD, DWD, 5 persistent EPS (top) and 25 persistent EPS (bottom).

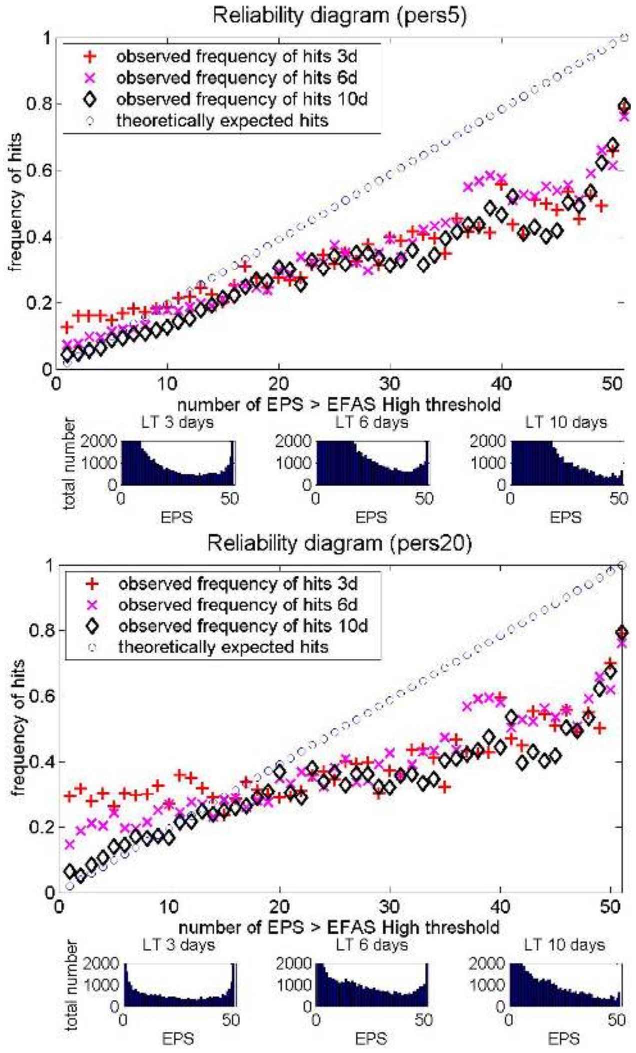

Fig. 11. Reliability diagrams for the cases that the previous forecast had<5EPS (upper diagram), or ≥20EPS (lower diagram). In each diagram 3 leadtimes (3, 6, 10 days) are reported. Points on the diagonal line have perfect reliability (i.e. forecasted probability=observed frequency). In the small histograms the absolute number of hits for each EPS number threshold that was used to create the reliability diagrams is reported.

the forecasted signal in two consecutive forecasts. The frequency bias is influenced positively by the persistence criterion and waiting for persistence of at least 5 EPS members, even results in decreasing bias over leadtime (Fig. 5). In general, the use of the persistence criterion leads to a strong decrease of *false alerts (f)*. However, this comes at the cost of a moderate increase of *misses (m)*, as well as a moderate decrease of *hits (h)*. The probability to have a *hit* with a low number of EPS is strongly increased through the use of the persistence criterion (see Fig. 11). Increasing the threshold for persistent EPS members, the skill expressed as odds increases, while the scores BSS, POD and HK (TSS) decrease. The odds skill score decreases with leadtime and increases for increasing EPS thresholds. The interpretation of the results obtained with complex skill scores is not straightforward. One should be aware of the specific behaviour of each skill score and their tendency to be strongly influenced by one of the fields of the contingency table. Therefore, it is very important to look also at the absolute numbers from the contingency tables, as well as to consider simple skill measures like FOH, FOM and POD that give a direct idea of the ratios between the fields of a contingency table.

The fact that the events studied here – forecasted discharges greater than EFAS high alert level (HAL) – are events that, by definition, should not occur very often in a two-year study period, influenced the whole analysis. Not always were there enough events in a pixel to fill all the fields of the contingency tables, which made it impossible to calculate skill scores like odds or HK that need values greater than zero in all fields. Additionally, even when all fields were filled in, we often faced the fact that the number of *hits*, *false alerts*, and *misses* was very low and, consequently, the number of positive rejects was very high – i.e., in the order of two magnitudes higher than *hits*, *false alerts* and *misses.* For skill scores like the HK, this meant that it was principally reduced to the same values as the POD.

In the analysis of the deterministic forecasts, the skill of EFAS hydrological forecasts using DWD meteorological data as input was in general lower than the one obtained with ECMWF products. Other studies in the literature indicate that the DWD precipitation forecasts tend to have also lower scores than the ones from ECMWF (see for instance Pedemonte et al., 2005). In the present study, this result might be due to a scaling issue: the spatial resolution of the DWD data used in EFAS is much higher than the spatial resolution of the JRC-MARS observed meteorological data that was used to calibrate the EFAS LISFLOOD model and to define a proxy for observed discharges in the skill assessment. The spatial resolution of ECMWF data was much more similar to the resolution of the JRC-MARS data. It is expected that forecasted precipitation intensities are higher at higher spatial resolutions and that convection in the DWD model is better resolved. Given this, it seems plausible that the EFAS discharges simulated with DWD data as input were sometimes too high (compared with the proxy-observed ones), resulting in a relatively higher number of false alerts, while the number of hits and misses was very similar to the results based on ECMWF data. This aspect needs to be further investigated, and is beyond the scope of this paper.

In the probabilistic forecasting framework, it is of high importance for the forecaster to know the probability that a forecasted event will happen. This study shows that the assumed equi-probability (i.e. reliability) of predictions from EPS in meteorology – which assumes that if 5 out of 50 EPS members forecast an event, there is a 10% chance for the event to happen – is not linearly translated into the hydrological probabilistic forecasts issued by EFAS. There are in fact strong biases (see Sect. 4.2) between the observed and expected frequencies.

It was shown that in the case of EFAS forecasts the probability of a current forecast is heavily conditioned by the probability that was forecasted in the previous forecast. The use of the persistence criterion successfully incorporates this information from past forecasts into the present EFAS forecast.

This study aims at analyzing the past performance of EFAS forecasts regarding a proxy for observed discharges. It has no ambitions on identifying at which point (in river location, forecasting leadtime, upstream area or number of EPS-based simulations forecasting the event) there is a skill in the forecasts. Skill scores are applied to a two-year period of existing operational forecasts in order to obtain a first picture of the performance of the forecasting system on a European scale. However, it must be stressed that the notion of skill and quality of a forecasting system depends heavily on the needs and acceptable trade-offs of the end-user. For a system like EFAS, aimed at complementing local forecasts issued by different hydrological services in different countries and river basins, the role of the end-user in defining references and thresholds to be used in skill assessment studies is crucial. In this context, the general goal is to build skill scores that are tailored to the specific needs of the user and that will contribute to increase the utility (and economic value) of a forecast.

Additionally, it must be noted that although the "number-crunching" exercise as performed in this study tried to mimic the behaviour of the human EFAS forecaster (notably by introducing a persistence criterion when flagging an event as a "forecasted" one), all efforts in this regard can only be done to a certain degree. It is certainly difficult (maybe even impossible) to translate the expertise of the forecaster into totally objective rules. Over the past two years, EFAS performance, in terms of success rate of external alerts sent out to national authorities whenever EFAS forecasts a potential flood situation, showed a hit rate of roughly 80%, which is much higher than what is indicated in the results of this study. Such a high performance was mainly achieved by the gaining of experience of the EFAS forecast team over time, as well as by the adoption of a more conservative attitude when sending out EFAS alerts, through which the number of false alerts was lowered drastically. Therefore, the results of this study should be taken as indicative of the system's performance, and the reader should bear in mind that they have a tendency to be more pessimistic than what has been observed in the past two years.

Finally, it should be noted that the verification of forecasts against observed flood events that are spread all over Europe is quite a complicated task. At such large scale of action, any observed data (real or proxy) will hardly include all observed flood events, and the introduction of a bias from taking into account too few misses is practically inevitable.

6 Conclusions

This study presents the statistical analysis of two full years of operational EFAS forecast data with different skill scores. It showed that the use of a persistence criterion, which considers the persistence of the forecasted signal in consecutive forecasts, can have a positive influence (i.e., false alert reduction) on the deterministic as well as the probabilistic EFAS EPS-based forecasts. Absolute numbers of *hits, misses* and *false alerts* were reported and different skill scores were analysed. The study highlighted some problems with using more complex skill scores for hydrological ensemble applications and advocates the use of simple, intuitively understandable skill scores. The use of EPS in hydrological forecasting proved to be of great added value to a flood early warning system, as the EPS-based forecasts showed, in general, higher skill than the deterministic-based ones.

As expected from such a global approach at a European scale, there are significant differences in skill for different rivers. This is linked to the different hydrometeorological conditions encountered in European transnational river basins, but also to calibration issues (data scarcity and not equally representative in space) and varying skill in the meteorological forecast input.

Finally, the findings of this study will be incorporated into the pre-operational EFAS at a pixel basis. The aim is to give the forecaster the possibility to assess past performance of the system at any time and to give guidance to the forecaster in estimating the forecast probability of an event to happen. The first step to assign a probability to EFAS flood forecasts might come from the results indicated in the reliability diagrams that were presented in this study. A second step will go further, with the incorporation of weather forecasts from different types (deterministic and probabilistic weather predictions). The general idea is to treat the deterministic forecasts as part of the ensemble probabilistic forecasting system by assigning them a weight and then assessing the resulting total probability of the ensemble flood forecasts.

Acknowledgement. The authors gratefully acknowledge the support of all staff of the weather driven natural hazards action and LMU IT support. The authors would also like to thank the Deutsche Wetterdienst and the European Centre for Medium-Range Weather Forecasts, as well as the JRC IPSC. Thanks go also to F. Pappenberger for fruitful brainstorming. For their financial support the authors thank the European Parliament, DG Environment and DG Enterprise.

Edited by: L. Pfister

References

Atger, F.: Verification of intense precipitation forecasts from single models and ensemble prediction systems, Nonlin. Processes Geophys., 8, 401–417, 2001, http://www.nonlin-processes-geophys.net/8/401/2001/.

Baldwin, M. E. and Kain, J. S.: Examining the sensitivity of various performance measures, 17th Conf. on Probability and Statistic in the Atmospheric Sciences, 84th AMS Annual Meeting, Seattle, WA, 2.9, 1–8 January, 2004.

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

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

Bradley, A. A., Schwartz, S. S., and Hashino, T.: Distributions-Oriented Verification of Ensemble Streamflow Predictions, J. Hydrometeorol., 5(3), 532–545, 2004.

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

Clark, M. P. and Hay, L. E.: Use of Medium-Range Numerical Weather Prediction Model Output to Produce Forecasts of Streamflow, J. Hydrometeorol., 5(1), 15–32, 2004.

De Roo, A., Wesseling, C. G., and Van Deurssen, W. P. A.: Physically based river basin modelling within a GIS: the LISFLOOD model, Hydrol. Process., 4(11–12), 1981–1992, 2000.

EEA, European Environment Agency: Mapping the impacts of recent natural disasters and technological accidents in Europe, *Environmental issue report* No. 35, published by European Environment Agency, Copenhagen, 2003, p. 47, 2003.

Flueck, J. A.: A study of some measures of forecast verification, Preprints, 10th Conf. on Probability and Statistics in Atmospheric Sciences, Edmonton, AB, Canada, Amer. Meteor. Soc., 69–73, 1987.

Franz, K. J., Hartmann, H. C., Sorooshian, S., and Bales, R.: Verification of National Weather Service Ensemble Streamflow Predictions for Water Supply Forecasting in the Colorado River Basin, J. Hydrometeorol., 4(6), 1105–1118, 2003.

Gandin, L. S. and Murphy, A.: Equitable skill scores for categorical forecasts, Mon. Weather Rev., 120, 361–370, 1992.

Gilbert, G. F.: Finley's tornado predictions, Am. Meteorol. J., 1, 166–172, 1884.

Göber, M., Wilson, C. A., Milton, S. F., and Stephenson, D. B.: Fairplay in the verification of operational quantitative precipitation forecasts, J. Hydrol., 288, 225–236, 2004.

Gordon, N. D. and Shaykewich, J. E.: Guidelines on Performance Assessment of Public Weather Services – Geneva: WMO, 2000, (WMO/TD 1023), 2000.

Hamill, T. M. and Jura, J.: Measuring forecast skill: is it real skill or is it the varying climatology?, Q. J. Roy. Meteorol. Soc., 132, 2905–2923, 2006.

Hanssen, A. W. and Kuipers, W. J. A.: On the relationship between the frequency of rain and various meteorological parameters, Meded. Verh., 81, 2–15, 1965.

Heidke, P.: Berechnung des Erfolges und der Gute der Windstarkevorhersagen im Sturmwarnungsdienst, Geogr. Ann., 8, 301–349, 1926.

Legg, T. P. and Mylne, K. R.: Early warnings of severe weather from ensemble forecast information, Weather and Forecasting, 19, 891–906, 2004.

Kalas, M., Ramos M. H., Thielen, J., and Babiakova, G.: Evaluation of the medium-range European flood forecasts for the March–April 2006 flood in the Morava River, J.Hydrol. Hydromech., 56(2), 116–132, 2008.

Marzban, C.: Scalar measures of performance in rare-event situations, Weather and Forecasting, 13, 753–763, 1998.

Mason, I.: Dependence of the critical success index on sample climate and threshold probability, Aust. Met. Mag., 37, 75–81, 1989.

Ebert, E. and McBride, J. L.: Methods for verifying quantitative precipitation forecasts: application to the BRMC LAPS model 24-hour precipitation forecasts, BRMC Techniques Development Report No. 2, 87 pp., 1997.

McBride, J. L. and Ebert, E.: Verification of quantitative precipitation forecasts from operational Numerical weather prediction models over Australia, Weather and Forecasting, 15, 103–121, 2000.

Mullusky, M., Demargne, J., Welles, E., Wu, L., and Schaake, J.: Hydrologic applications of short and medium range ensemble forecasts in the NWS Advanced Hydrologic Prediction Services (AHPS), AMS 16th Conference on Numerical Weather Prediction Symposium on Forecasting the Weather and Climate of the Atmosphere and Ocean, Seattle, Paper J11.5, 2004.

Murphy, A. H. and Epstein, E. S.: A note on probability forecasts and "hedging", J. Appl. Meteor., 6, 1002–1004, 1967.

Murphy, A. H. and Winkler, R. L.: A general framework for forecast verification, Mon. Weather. Rev., 115, 1330–1338, 1987.

Murphy, A. H.: The Finlay Affair. A signal event in the history of forecast verification, Weather and Forecasting, 11, 3–20, 1996.

Murphy, A. H.: Forecast verification, in: Economic Value of Weather and Forecasts, edited by: Katz, R. W. and Murphy, A. H., 19–74, Cambridge: Cambridge University Press, 1997.

Nash, J. E. and Sutcliffe, J. V.: River flow forecasting through conceptual models Part 1 – A discussion of principles, J. Hydrol., 10(3), 282–290, 1970.

Pappenberger F., Scipal, K., Buizza, R.: Hydrological aspects of meteorological verification, Atmos. Sci. Lett., 9(2), 43–52, 2008.

Pedemonte L., Corazza, M., Sacchetti, D., Trovatore, E., and Buzzi, A.: Verification of limited-area models precipitation forecasts during the map-sop, Proceedings ICAM/MAP 2005 Zadar, Croatia, 23–27 May, 2005.

Peirce, C. S.: The numerical measure of success of predictions, Science, 4, 453–454, 1884.

Roulin, E.: Skill and relative economic value of medium-range hydrological ensemble predictions, Hydrol. Earth Syst. Sci., 11, 725–737, 2007, http://www.hydrol-earth-syst-sci.net/11/725/2007/.

Roulin, E. and Vannitsem, S.: Skill of Medium-Range Hydrological Ensemble Predictions, J. Hydrometeorol., 6(5), 729–744, 2005.

Schaefer, J. T.: The Critical Success Index as an Indicator of Warning Skill, Weather and Forecasting, 5, 570–575, 1990.

Stanski, H. R., Wilson, L. J., and Burrows, W. R.: Survey of common verification methods in meteorology, Technical Report No. 8 Geneva: WMO, 1989 (WMO/TD 358), 1989.

Stephenson, D. B.: Use of the "odds ratio" for diagnosing forecast skill, Weather and Forecasting, 15, 221–232, 2000.

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

Van Der Knijff, J. and De Roo, A.: LISFLOOD – distributed water balance and flood simulation model, User manual; EUR Report EUR 22166 EN, 2006.

Wilson, C.: Review of current methods and tools for verification of numerical forecasts of precipitation, COST-717 Working document WDF_01_200109_1, www.smhi.se/cost717/, 2001.

Wilks, D. S.: Statistical Methods in the Atmospheric Sciences, Academic Press, San Diego,CA, 265–267, 1995.

Younis, J., Ramos, M. H., and Thielen, J.: EFAS forecasts for the March–April 2006 flood in the Czech part of the Elbe river basin – a case study, Atmos. Sci. Lett., 9, 88–94, doi:10.1002/asl.179., 2008.