Journal of Hydrology 424–425 (2012) 143–153

Contents lists available at SciVerse ScienceDirect

Journal of Hydrology

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

Ensemble hydro-meteorological simulation for flash flood early detection in southern Switzerland

Lorenzo Alfieri [a,*], Jutta Thielen [a], Florian Pappenberger [b]

[a] European Commission – Joint Research Centre, Institute for Environment and Sustainability, via E. Fermi, 2749, 21027 Ispra, VA, Italy
[b] European Centre for Medium Range Weather Forecasts, Shinfield Park, Reading RG2 9AX, UK

ARTICLE INFO

Article history:
Received 22 July 2011
Received in revised form 12 December 2011
Accepted 27 December 2011
Available online 3 January 2012
This manuscript was handled by Konstantine P. Georgakakos, Editor-in-Chief, with the assistance of Marco Borga, Associate Editor

Keywords:
Flash floods
Hydrological ensemble predictions
Early warning
Discharge threshold exceedance

SUMMARY

Ongoing changing climate has raised the attention towards weather driven natural hazards. Local floodings and debris flows following exceptional downpours often come without any adequate warning and cause heavy tolls to the human society. This work proposes a novel flood alert system for small catchments prone to flash flooding, capable of monitoring a large portion of the European domain. Operational streamflow simulations are produced through distributed hydrological modeling of ensemble weather forecasts. A long-term reforecast dataset is run through the same hydrological model to derive coherent warning thresholds. These are compared with operational discharge ensembles in a threshold exceedance analysis to produce early warnings.

A case study in the southern Switzerland is tested over a 17-month period and system skills are evaluated by means of different quantitative and qualitative analyses. Results from three different predictors derived from the streamflow ensemble are shown and compared, also by accounting for the persistence of lagged forecasts. Significant improvements in predicting discharge thresholds exceedance are achieved by fitting gamma probability distributions to the raw ensemble. Further discussion underlines the limits of predictability of extreme events in small catchments due to the comparatively coarse space–time resolution of current weather forecasts.

1. Introduction

Advances in the skill of Numerical Weather Predictions (NWPs) have fostered the creation of a number of flood forecasting and early warning systems based on Ensemble Prediction Systems (EPS) as inputs (Cloke et al., 2009; Cloke and Pappenberger, 2009; Thielen et al., 2009). Recent work shows that operational flood alert systems driven by hydrological EPS (HEPS) are currently capable of detecting upcoming flood events in large river basins up to 10 days in advance (Hopson and Webster, 2010; Pappenberger et al., 2008). Such warning lead times are of crucial importance to both increase the preparedness of the population and coordinate the crisis management with timely intervention plans.

Flood warnings can only be effective if the forecasting system provides sufficient skill and consistency (Pappenberger et al., 2011; Persson and Grazzini, 2007). Recent works show that taking into account the persistence of lagged forecasts can reduce the number of false alarms and thus improve the performance of warning systems based on threshold exceedance (Bartholmes et al., 2009; Dietrich et al., 2009). In addition, lagged forecasts increase forecast consistency (Persson and Grazzini, 2007). The European Flood Alert System (EFAS) is one example of a medium range flood forecasting system in which lagged forecasts are used successfully (Thielen et al., 2009).

EPS and Limited-area EPS (LEPS) are applied with increasing frequency for riverine flood forecasting (Rotach et al., 2009). Fewer attempts have been specifically focused on flash flood early detection (Addor et al., 2011; Marty et al., 2008; Philipp et al., 2008; Reed et al., 2007; Younis et al., 2008), as the space–time resolution of weather predictions is often too coarse to reproduce unbiased streamflow estimates at the scale of small catchments. In fact, the skill of those systems much relies on the estimation of coherent warning thresholds, which are particularly difficult to reproduce at the scales of interest of flash floods. Firstly, many operating meteorological gauging networks are not dense enough to capture the small-scale horizontal and vertical variability of storms producing flash floods – this is if any gauge exists at all for these events which take place in small watersheds in mountainous areas. In this regard, weather radars have reached widespread coverage over large areas in Europe and will become an interesting option for use, once datasets of some years are collected. Secondly, seamless time series of meteorological variables are usually not available at fine (i.e., sub-daily) temporal resolution for time spans long enough to

* Corresponding author. Tel.: +39 0332 78 6999; fax: +39 0332 78 6653.
E-mail address: lorenzo.alfieri@jrc.ec.europa.eu (L. Alfieri).

0022-1694/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.jhydrol.2011.12.038

reproduce a robust climatology. Thirdly, the analysis of extreme events should be based on climatological datasets (e.g., 20 years or longer) in order to estimate reliable flood flows linked to low probability of occurrence.

Previous work has shown the dramatic improvement of meteorological forecasts after calibration with long-term reforecast datasets (Fundel et al., 2010; Hamill et al., 2006). Yet, little effort has been put to transfer the same idea to correct streamflow predictions driven by weather forecasts, particularly at the scale of interest of flash floods. In this study we describe a new flash flood alert system, based on hydrological simulation of probabilistic ensemble forecasts. Coherent warning thresholds are derived from a long-term reforecasts dataset generated with the same model used for operational forecasts. The system is hereafter referred to as EFAS-FF (European Flood Alert System for Flash Floods).

The objective of this paper is to propose a novel predictor for flash floods and compare it to more traditional methods. Also, we analyze the impact of lagged forecasts on the forecast performance and suggest that any modern flash flood forecasting system will have to rely on lagged forecasts as additional source of information.

A schematic description of the data and the different steps involved in the early warning system is presented in Section 2, while Section 3 shows the evaluation methods used to test the system performance for a selected case study. Results are shown in Section 4 and discussed in Section 5, while concluding remarks are drawn in the last section.

2. Data and methods

2.1. A European flash flood alert system

EFAS-FF is aimed to provide a tool for improving preparedness to small size catchments prone to flash floods, where flood events are generated by severe downpours producing high rainfall intensities over short durations. EFAS-FF is activated by positive signals from the European Precipitation Index (EPIC, Alfieri et al., 2011b), which has already been successfully used in monitoring potential flash flood warnings at the continental scale. The EPIC index runs on daily basis on a 5 million km^2 area, covering most of the European domain, and gives an indication on the severity of upcoming rainfall events. It compares forecasted accumulated rainfall, over typical durations for flash floods, with corresponding reference thresholds derived from the climatology. Where a positive signal is detected from EPIC, a catchment-scale hydrological simulation is activated at 1-km resolution, by taking as initial conditions the operational results provided by EFAS simulations, run at 5-km resolution. Ensemble weather forecasts are run through a distributed hydrological model to predict discharges, namely, one hydrological simulation for each member of the meteorological ensemble. These are compared to the reference climatology through a threshold exceedance analysis to derive probability-based warnings. The advantage of such a cascade of approaches is that the location of the warnings and the accuracy of the discharge predictions are improved given limited CPU resources. In addition, model version developed specifically for flash flood applications can be nested into a more general framework.

Three discharge warning thresholds are calculated following the same approach used in the European Flood Alert System and are named Medium, High and Severe alert. In details, a meteorological climatology based on weather reforecasts covering a large portion of the European domain (see Section 2.2) is routed through a hydrological model at 1 km resolution to derive a continuous discharge climatology. For each grid point on the river network, annual maxima of discharge are extracted from the climatology and used to estimate warning thresholds. The mean of the annual maxima is considered as Medium threshold for flood warning, as it usually corresponds to peak flows around the bank-full conditions (Carpenter et al., 1999). Although it does not correspond to significant flooding conditions, it is the first level of the warning system and it is useful for monitoring upcoming high flow events potentially increasing in severity. It is a robust indicator as it is based only on the sample of annual maxima, with no assumptions on their statistical distribution. In addition, a Gumbel extreme value distribution is hypothesized for the annual maxima of discharge at each point, and peak flows corresponding to return periods of 5 and 20 years are chosen as High and Severe warning thresholds. A similar approach is used to derive thresholds of heavy rainfall to be used to calculate the EPIC index.

EFAS-FF makes use of a distributed hydrological model named LISFLOOD, described in detail by Van der Knijff et al. (2010). LISFLOOD is a hybrid between a conceptual and physically based rainfall–runoff model combined with a routing module for river channels. It simulates canopy and surface processes, snow accumulation and melting, soil and groundwater processes and flow in the river network. LISFLOOD has been specifically designed for large river basins (De Roo et al., 2001) but has shown positive results in applications to smaller watersheds (e.g., Alfieri et al., 2011a; Younis et al., 2008).

The proposed EFAS-FF system is designed to run hydrological simulations at 1-km spatial resolution and 3-h time resolution, which is currently the optimal tradeoff between model resolution and computational feasibility. It is fit to catchments with drainage area up to 1000–2000 km^2 where the most hazardous events are induced by storms of duration up to 24 h (Gaume et al., 2009; Reed et al., 2007). The error induced by using initial conditions at 5-km resolution for the 1-km resolution model was shown to be often positive but negligible for high flow conditions (Alfieri et al., 2010). In particular, for flash-flood-prone catchments, the effective rainfall producing flood peaks is entirely included in one meteorological forecast. Therefore, the initial conditions at coarser resolution that are used to initialize each hydrological simulation, are mostly useful to represent those runoff components linked to slow response (i.e., soil moisture, snow cover, cumulative interception, groundwater storage) rather than the initial discharge/water stage in the drainage network.

2.2. Meteorological data

Meteorological forecasts were provided by the Consortium for Small-scale Modeling (COSMO). The Limited-Area Ensemble Prediction System (COSMO-LEPS) is the operational 16-member ensemble forecast of COSMO (Marsigli et al., 2005). It is run once a day at 12:00 UTC and spans 132 h. Operationally, the two latest 51-member ensemble forecasts of the European Centre for Medium-range Weather Forecasts (ECMWF) are combined in a 102-member super-ensemble, from which 16 representative members are selected through cluster analysis technique and used to initialize the COSMO model run. Forecast fields of precipitation, 2 m temperature and potential evapo-transpiration are provided on a rotated spherical grid covering most of southern and central Europe, with horizontal resolution of $0.09^\circ \times 0.09^\circ$ (about 10 km $\times$ 10 km) and temporal resolution of 3 h.

In addition, a continuous meteorological climatology was created from a set of 30-year reforecasts (Fundel et al., 2010), which was made available from the COSMO Consortium. It consists of a set of deterministic reforecasts, initialized every 3 days from ECMWF control run, by using ERA 40 re-analysis dataset (Uppala et al., 2005) as initial and boundary conditions. A continuous climatology is obtained by attaching together the first three days of data of each forecast, to produce a seamless dataset with the same

spatial and temporal resolution of operational COSMO-LEPS forecasts. Hereafter, it is referred to as COSMO-30y. Daily COSMO-LEPS forecasts at 10-km resolution were available from 7th July 2008 to 30th November 2009.

2.3. Case study

To evaluate the performances of the proposed warning system, we carried out a simulation experiment by running the system continuously (i.e., mimicking the daily operational runs in hindcast mode) on a test catchment using COSMO-LEPS forecasts as input data, and comparing the resulting ensemble discharges with observed values at the outlet. The considered case study is the Verzasca, an alpine catchment in the southern Switzerland. It is characterized by a V-shaped narrow valley with steep slopes, shallow soils (mostly <30 cm) and elevation ranging between 2870 and 490 m a.s.l. at the gauging station of Lavertezzo (upstream area A_U = 186 km^2). The catchment area is little affected by urban settlements and human activities. However, the artificial Vogorno Lake lies just downstream the river gauge. It is bounded by a 220 m high dam which is mainly operated for hydropower production. Forests (30%), shrub (25%), rocks (20%) and alpine pastures (20%) are the predominant land cover classes. The hydrological regime is governed by snowmelt in spring and early summer and by heavy rainfall events in fall (Wöhling et al., 2006). Baseflow in winter can be less than 1 m^3 s^{-1}, while the mean annual flood peak in Lavertezzo is about 400 m^3 s^{-1}.

Seamless hourly discharge observations at Lavertezzo were provided by the Swiss Federal Office for the Environment (FOEN), together with a set of 19 annual maxima of observed discharge from 1990 to 2008. Fig. 1 shows a map of the Verzasca catchment, together with the 1-km drainage network considered by the LISFLOOD model.

The hydrological simulation was run on the Verzasca catchment set up at 1 km resolution, with a model calculation sub-step of 30 min and output maps stored every 3 h. Meteorological input data were COSMO-LEPS forecasts, ranging 132 h each, for 512 consecutive days, starting in July 7th, 2008. For each hydrological forecast, simulated ensemble discharges at the outlet are extracted and split according to the lead time between 1 and 5 days. Forecasted and observed discharges are first normalized by the corresponding mean of the annual maxima, namely, the mean of the simulated and of the observed maxima, respectively. Previous works (Alfieri et al., 2010) have shown the advantages of the normalization, such as the reduction of bias in simulated discharge, due to modeling small catchments at a spatial resolution comparatively coarse with respect to their size. Also, similar findings were drawn by Norbiato et al. (2009) in the context of the Flash Flood Guidance (FFG) approach. The normalized discharge K_Q is an intuitive indicator of the state of the river in terms of flood warning (i.e., K_Q = 1 corresponds to the mean annual maxima of discharge) and can be easily compared at different locations and different catchments. As an example, we show in Fig. 2 a comparison between normalized observed and simulated ensemble discharges at the catchment outlet, for a fixed lead time of 4 days (i.e., 72–96 h ahead). It is worth noting that this simulation is meant to test the EFAS-FF system in an operational way, therefore no specific calibration was carried out on the hydrological model. Operationally, the same model setting would be applied to any catchment in the considered domain, which boundaries are set by the meteorological input data. In addition, this study aims at testing the limits of predictability of such system, by choosing a catchment with area of the same magnitude as the meteorological input (i.e., 10 × 10 km^2). Sangati and Borga (2009) showed that the error in normalized rainfall and normalized peak discharge becomes significant when the rainfall resolution is similar to the characteristic basin length. However, the proposed approach is totally independent of in situ measurements, so the evaluation carried out in the following allows one to extrapolate objective findings of wide validity.

2.4. Predictors for flash flood detection

Three kinds of predictor are created from the discharge ensemble forecasts and their performance towards the observations are analyzed and compared together. They are described in the following.

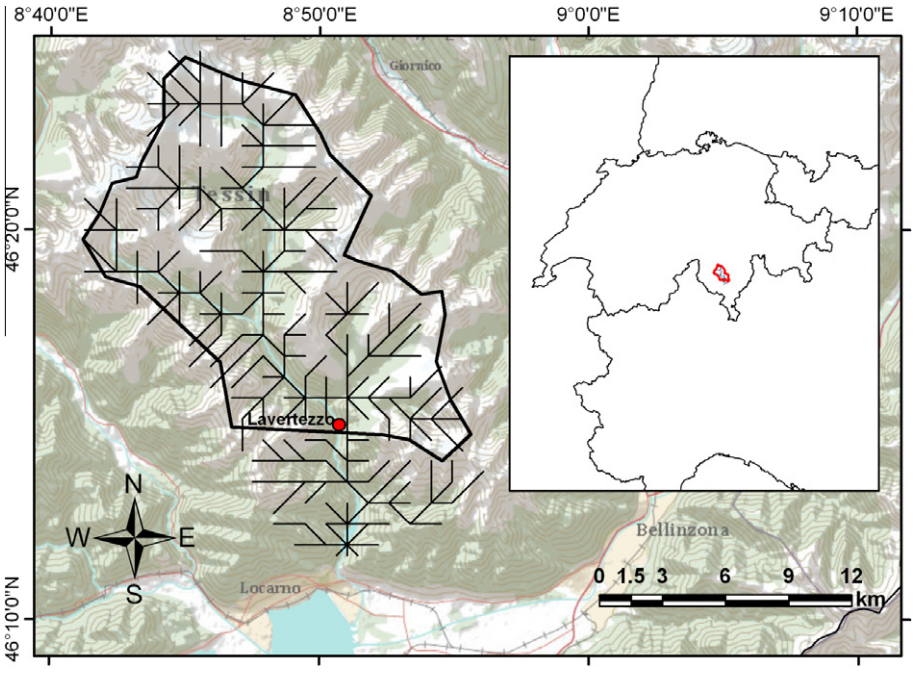

Fig. 1. Map of the Verzasca catchment (Switzerland) and 1-km drainage network for the simulated model. The catchment outlet in Lavertezzo is shown with a circle.

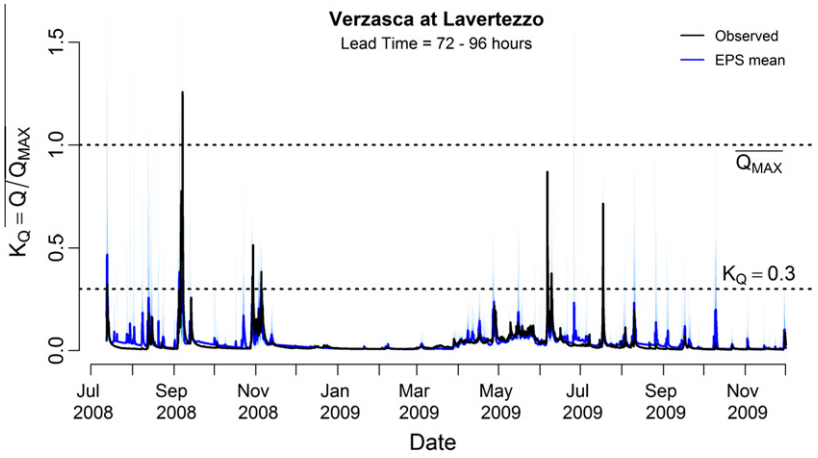

Fig. 2. Normalized observed and simulated ensemble discharges at Lavertezzo, for the simulation period (4 day lead-time).

1. The first predictor is the full ensemble of discharge forecasts as obtained from the hydrological simulation, sorted from the lowest to the highest value for each time step. In the following it is referred to as EPS. It provides probabilistic predictions, though its robustness is likely to decrease in the upper and lower quantiles, which can be heavily affected by outliers.
2. The ensemble mean is the second option chosen, being the most widely used deterministic predictor derived from an EPS. Despite its robustness, due to considering the full set of members, it gives no information on the ensemble spread. In addition the ensemble mean loses information on the temporal development, which is given by individual ensemble members. However, this is irrelevant in the adopted decision framework, which is based on probabilistic threshold exceedance for each individual forecast horizon.
3. The third considered predictor is obtained by fitting an analytical probability distribution to each ensemble of 16 members, corresponding to every selected time-step and forecast lead-time. While it can be argued that the fitting generates additional uncertainty, on the other hand it provides a probabilistic predictor with increased robustness, compared to the EPS, particularly in the tails of the distribution.

While the first two predictors have been widely used and discussed in the literature (see Cloke and Pappenberger, 2009 and references therein), the third option is proposed and tested within this work. In details, each ensemble is inferred with a 2-parameter gamma distribution by means of L-moment estimators. The gamma is a family of very flexible probability distributions which have long been used to model many natural phenomena such as rainfall and runoff data, including extreme events (Bobée and Ashkar, 1991; Loucks and van Beek, 2005). The probability density function (pdf) of a gamma-distributed random variable x is defined as:

$$f(x) = \frac{x^{\alpha-1}e^{-x/\beta}}{\beta^{\alpha}\Gamma(\alpha)} \quad \text{for } x \geqslant 0 \text{ and } \alpha, \beta > 0, \tag{1}$$

where α is the shape parameter, β the scale parameter, and $\Gamma(\cdot)$ denotes the gamma function

$$\Gamma(\alpha) = \int_0^{\infty} t^{\alpha-1}e^{-t}dt. \tag{2}$$

Parameters α and β are estimated by equaling the first two sample L-moments with those of the gamma distribution (λ_1, λ_2) given by the following equations (see e.g., Hosking, 1990):

$$\lambda_1 = \alpha\beta, \tag{3}$$

$$\lambda_2 = \pi^{-1/2}\beta\Gamma\left(\alpha+\frac{1}{2}\right)/\Gamma(\alpha). \tag{4}$$

L-moment estimators are known for being nearly unbiased for a wide range of sample sizes and distributions (Vogel and Fennessey, 1993), and become particularly useful for short samples as in the present work (i.e., 16 values for each fitting).

3. Evaluation methods

A single index is unlikely to provide deep understanding of all the different properties of a forecast system (Napolitano et al., 2011). In this paper we show a number of different analyses.

3.1. Visual assessment

Visual assessment of flow hydrographs is a standard and quick method to illustrate differences in behavior. It gives useful indications on systematic problems in a forecast chain.

3.2. Probability plot

When comparing the observations with the simulated dataset an informative perspective is given by considering their probability distributions. Ideally, the probability density function $p_i(K_{Q,F})$ of the predictions should match that of the observations $p_i(K_{Q,O})$. The hypothesis H_0: $p_i(K_{Q,F}) \equiv p_i(K_{Q,O})$ can be verified by defining the probability integral transform $z_i = P_i(K_{Q,O})$, that is, the cumulative distribution function of the predictions $K_{Q,F}$ in correspondence of each observed value. Under the hypothesis H_0, the sample of z_i follows a standard uniform distribution $U(0,1)$. A useful graphical representation to verify the uniformity hypothesis is obtained by plotting the set of sorted z_i values versus their theoretical quantiles j_i taken from a standard uniform distribution, which are calculated by simply dividing the rank of each observation R_i by the sample size: $j_i = R_i/n$. The benefit of these graphs is twofold: first, the shape of the resulting curve and its departure from the bisector reveals different qualitative information on the estimation skill (see e.g., Laio and Tamea, 2007). Second, deviations from the uniformity can be quantified objectively through statistical tests at selected significance levels by using the Kolmogorov statistic. This graphical tool accounts for the full probability distribution of theoretical and forecasted values, therefore it should not be confused with the apparently similar reliability diagram (e.g., see Wilks, 2006), used for the verification of probability forecasts of a binary predictor.

3.3. Mean squared error and variance

The quantitative performance of the three predictors against the observations can be expressed by making use of the relation between the mean squared error (*MSE*), the variance (σ^2) and the bias (Δ) of a variable x,

$$MSE(x) = \sigma^2(x) + \Delta^2(x), \tag{5}$$

where x are the estimation residuals between the forecasted ($K_{Q,F}$) and the observed ($K_{Q,O}$) normalized discharge: $x = K_{Q,F} - K_{Q,O}$. The mean squared error and the variance are calculated from all the n residuals with the same lead time and estimation quantile:

$$MSE(x) = \frac{1}{n}\sum_{i=1}^{n}(K_{Q,F} - K_{Q,O})_i^2, \tag{6}$$

$$\sigma^2(x) = \frac{1}{n}\sum_{i=1}^{n}(K_{Q,F} - \overline{K_{Q,O}})_i^2, \tag{7}$$

where $\overline{K_{Q,O}}$ is the sample mean of the observed time series. The mean squared error and variance can be plotted in the same graph. The difference between the two lines represents the squared bias of the estimation residuals ($\Delta^2(x)$). This graphical representation is very informative, as it allows one to distinguish the contribution to the total error of the bias, which measures the reliability, and of the variance, which measures the resolution of the predictor. This kind of graph is particularly useful in flood early warning systems based on threshold exceedance analysis, where "deterministic" decisions have to be made on the basis of probabilistic information at selected quantiles. The system performance depends much on its reliability (i.e., its bias) and in turn by the choice of appropriate warning thresholds. This is substantially different to flood forecasting systems that focus on the probability distribution of the estimates and the uncertainty analysis, and in turn on improving the resolution of the system (i.e., the minimization of its variance).

3.4. ROC area

The previous scores provide a quantitative evaluation of the forecasted discharge towards the observed one but they are not representative of the performance of a warning system based on threshold exceedance analysis. Indeed, the performances of such a system are determined by its reliability in the range of high flows, in the proximity of warning thresholds. The ultimate goal is to detect all the events exceeding a certain threshold without providing false alarms, and its outcome is translated in the issuing or not issuing a (flood) warning. A number of scores have been proposed and described in the literature to evaluate the skills of both deterministic and probabilistic forecasts for threshold exceedance analysis (see e.g., Jolliffe and Stephenson, 2003; Wilks, 2006, Chapter 7). In this work, we focus on testing the skill of the two probabilistic predictors, the EPS and the gamma fit, through Relative Operating Characteristic (ROC) curves. ROC curves have been widely used to measure the skill of dichotomous forecasts based on probabilistic information, as they plot the empirical relation between the Hit Rate (HR) and False Alarm Rate (FAR) for different probability thresholds. An interesting property is that the area under the ROC curve, hereafter referred to as AROC, can be used as skill score to assess the overall system performance through just one value. In particular, the range of interest measured by AROC is between $AROC = 0.5$, which corresponds to random forecasts, and $AROC = 1$ for perfect forecasts (while $AROC = 0$ means forecasts perfectly opposed to the observations).

3.5. Rank histogram

A rank histogram analyzes the location of the verifying observations in an ensemble system. In particular, it shows the probability density function of the observations versus the corresponding rank of the interval between each couple of consecutive EPS members in which each observed value falls in. It shows the sum of the ranks of individual observations in a corresponding forecast. In an ensemble with perfect spread, each member represents an equally likely scenario, so the observation is equally likely to fall between any two members. A flat rank histogram does not necessarily indicate a good forecast; it only measures whether the observed probability distribution is well represented by the ensemble. A U-shaped histogram indicates that the ensemble spread is too small and that many observations fall outside the extremes of the ensemble. A dome-shaped indicates a too large ensemble spread with most observations falling near the centre of the ensemble. Asymmetric rank histograms are caused by a biased forecast system.

4. Results

4.1. Visual comparison of flow hydrographs

The three kinds of predictor described above are derived from the simulated ensemble discharges for five lead times between 1 and 5 days. Thus, we obtained 15 predictors for a 512-day period of 3-h normalized discharge to compare with the observations. A graphical comparison of the three predictors is shown in Fig. 3, for an observed event occurred in June 2009, together with the corresponding forecasts with fixed lead time of 4 days (i.e., 72–96 h). The top panel shows the EPS with blue shadings indicating different probabilities around the median (i.e., the darkest polygon). Each line among two different shadings of blue corresponds to a value from the sorted streamflow ensemble of 16 members. Similarly, the bottom panel indicates with different purple shadings, K_Q quantiles taken from the gamma distributions obtained for each time step. For coherent graphical comparison with the top panel, K_Q values are calculated at quantiles corresponding to $F(K_Q) = \{1, 2, \ldots, m\}/(m+1)$, with $m = 16$ as the EPS sample size. However, for each time step, a full gamma distribution is defined within the domain $[0, \infty)$. This is a first important property of continuous distributions, as it overcomes the inconsistency of the EPS, which assumes a probability of occurrence equal to zero for each value outside its range. By comparing top and bottom panel in Fig. 3, one can see the more regular variations of the fitted distributions compared to the EPS, as well as a reduction of the probability associated to values much different from the other EPS members. For example, the outlier in the EPS on 01/06/09 is significantly damped in the corresponding gamma fit. Further, the inset figure on the right shows the empirical (EPS) and fitted (gamma) cumulative distribution function of the normalized discharge for one time step, indicated with a dashed line in the two panels. The observed value is also plotted with a vertical dashed line in the inset figure.

Two skill scores widely used in hydrological forecasting have also been calculated for each of the five forecast lead times (LTs) and are displayed in Table 1. In the first row, the Continuous Ranked Probability Score (CRPS, e.g., Hersbach, 2000) evaluates the integrated squared differences of the EPS against the vector of observations. In the second row, the Nash–Sutcliffe efficiency (NS, Nash and Sutcliffe, 1970) of the EPS mean is shown. Both skill scores denote the highest performance with lead time between 3 and 5 days. In details, the CRPS ranges between 1.4% and 1.8% of the mean of the annual maxima of discharges, while the

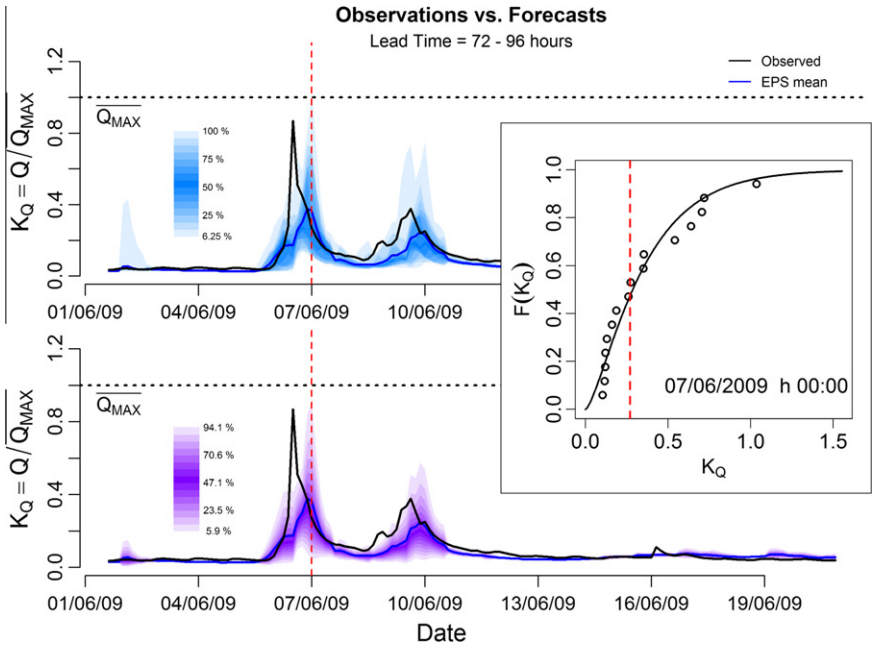

Fig. 3. Four-day forecast of normalized discharge at Lavertezzo for one event. EPS (top panel) and gamma fitted quantiles (bottom panel) are shown with blue and purple shadings together with the EPS mean (blue solid line) and the normalized observations (black solid line). Empirical and fitted cdf for one time-step are shown in the inset figure, with observed value plotted with a red dashed line. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 1
CRPS of the EPS and Nash–Sutcliffe (NS) efficiency of the EPS mean, towards the forecast lead time between 1 and 5 days (17-month simulation).

	LT1	LT2	LT3	LT4	LT5
CRPS (EPS)	0.018	0.015	0.014	0.014	0.014
NS (EPS mean)	0.32	0.36	0.46	0.48	0.44

Nash–Sutcliffe efficiency is always positive, with a maximum value on the 4-day lead time.

4.2. Probability plot

Fig. 4 shows the probability plot of the ensemble predictions for each considered forecast lead-time. Kolmogorov statistic (see e.g., Stephens, 1974) at 5% significance level is calculated and shown in the graph through confidence bands parallel to the bisector. The five curves show a similar behavior, with significant over-predictions around the central quantiles (i.e., $j_i = 0.3$–0.8). Note that this bias is quantified in the graph in terms of cumulative distribution, but does not necessarily translate to large quantitative overestimations in the range of high flows. For example, the median quantile of the forecasted distributions ($z_i = 0.5$; $K_{Q,F} \approx 0.025$), corresponds to a much larger theoretical quantile (i.e., of the observations) $j_i = 0.65$–0.70, though the quantitative forecast is about 1% larger (i.e., $K_{Q,O} \approx 0.035$), compared to $\overline{K_{Q,O}}$, for all the lead times. For larger quantiles (i.e., $j_i \geq 0.8$) the observed and forecasted distributions become more uniform, with only the 1-day lead time being outside the 5% confidence bands. A possible explanation to this behavior is the use of initial conditions taken from a different model at coarser resolution (Alfieri et al., 2010). In fact, as mentioned in Section 2.1, this approximation was shown to produce positive prediction errors becoming negligible for high flow conditions, which are the cases of highest interest in flood warning systems.

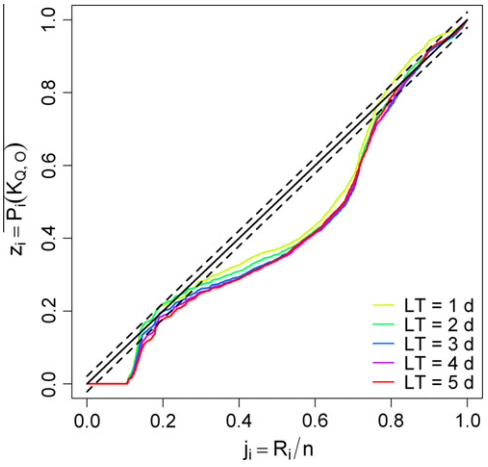

Fig. 4. Probability plot of the ensemble predictions for each considered forecast lead-time, together with Kolmogorov confidence bands at 5% significance level.

4.3. Mean squared error and variance

The left column of Fig. 5 shows the quantitative performance of the three predictors against the estimation quantile, for the five aforementioned lead times between 1 and 5 days (one for each panel). For each predictor two lines are plotted: the top one is the $MSE(x)$, while the bottom one is the variance $\sigma^2(x)$. In the five left panels of Fig. 5 (i.e., full data set) almost all the error is attributed to the estimation variance, while the bias is negligible for all the quantiles of estimation. Results for the EPS mean are shown as horizontal lines, as it does not depend on the estimation quantile. The ensemble mean proves to be a reliable (i.e., with little bias) and precise (i.e., with low variance) estimator. The EPS and the gamma fit provide similar results, with optimal values of both *MSE* and bias always around the median (i.e., 50% quantile). Further, the error variability tends to increase with the forecast lead time, though

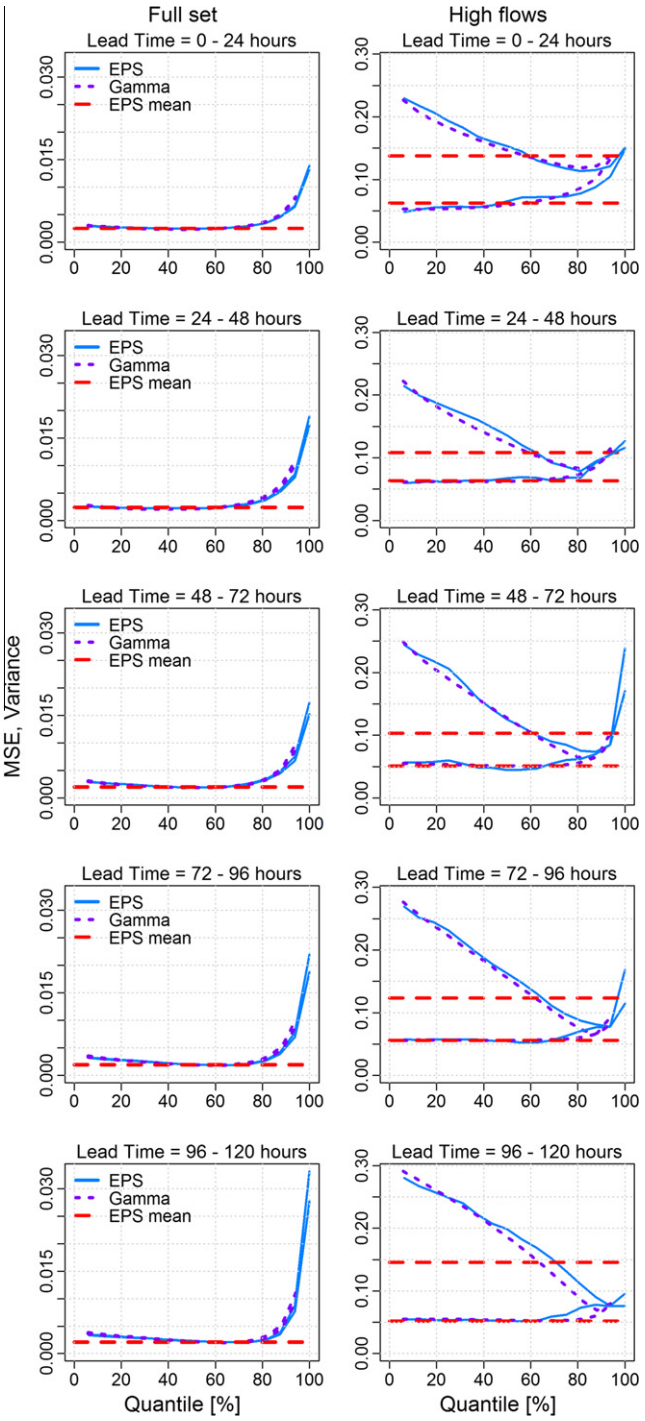

Fig. 5. Mean squared error (top lines) and variance (bottom lines) of the estimation residuals towards the quantile, for the three tested predictors. Results are displayed for the full data set (left column) and for a subset of high flows with $K_{Q,O} > 0.3$ (right column). Each row shows a different lead time range between 1 and 5 days.

values with the lowest *MSE* occur in the 3 and 4 day lead-time windows.

The quantitative analysis of the prediction error was carried out also for a subset of observed high flow, that is, for those time steps when the observed discharge was above a certain threshold. The threshold for this analysis $K^T_{Q,O}$ was chosen as a tradeoff value between (I) being representative of an actual high flow condition that follows a significant rainfall event and (II) having a minimum number of events to analyze. Results are shown in the right panels of

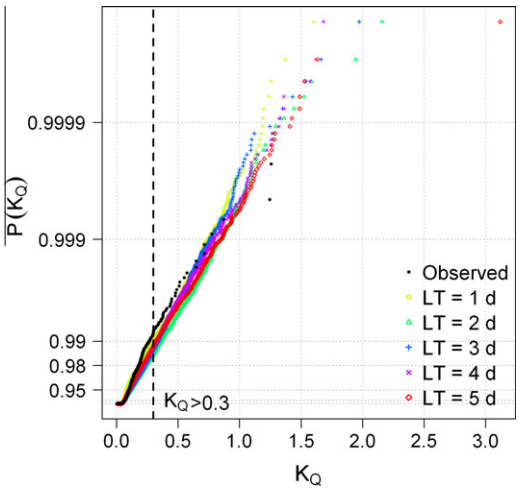

Fig. 6. Empirical cumulative distribution of observed and forecasted normalized discharge for five lead times on gamma probabilistic paper. Threshold value for high flow is shown with a dashed vertical line.

Fig. 5. The adopted value $K^T_{Q,O} = 0.3$ resulted in eight hydrograph portions above threshold (see Fig. 2). Due to the limited period of record, the chosen threshold is a relatively low discharge in terms of flood flows and does not correspond to a significant flood hazard. However, it is a first useful step to understand the performance of the system and to address further analysis. As shown in Fig. 6, the threshold $K^T_{Q,O} = 0.3$ corresponds to about the 99% quantile of the observed discharge and the 98–99% quantile of the ensemble forecasts for the five lead times. Also, it is worth noting from Fig. 6 that forecasted high flows for the five lead times are well fitted by gamma distributions with the same shape parameter, as points of the upper tail of their empirical cumulative distributions follow straight lines in a gamma probabilistic space.

Results in Fig. 5 denote a roughly constant contribution of the estimation variance (bottom one of each line type, in the five right panels) for a wide portion of the quantile range, up to the 80% quantile, for all the different lead times and predictors. Again, the difference between top and bottom line for each predictor measures the squared bias of estimation. For high flows, the bias becomes a significant portion of the total error. The gamma fit provides slightly improved performances compared to the original EPS, particularly for the largest lead times where the *MSE* reaches lower absolute values. The EPS mean produces its best results (i.e., minimum *MSE*) for lead time of 3 days; with performance progressively decreasing towards the boundaries of the forecasting range (1 and 5 day lead time). However it generates *MSE* values that are below those of the two probabilistic predictors for most of the quantile range. In fact, the EPS and the gamma fit provide improved results around the 70–90% quantile range, suggesting a generalized underestimation of the observed values for lower quantiles. Interestingly, the increased spread of the ensemble with lead time is reflected in an increasing spread of the bias towards the quantiles, while the estimation variance does not vary significantly with the forecast lead time.

4.4. ROC area

Fig. 7 shows four plots of AROC towards the forecast lead time, for the EPS and the gamma fit (triangles and diamonds, respectively, joined by solid lines). Results in the left panels refer to threshold values $K^{T1}_Q = 0.3$ (top) and $K^{T1}_Q = 0.5$ (bottom). For example, in the top-left panel, results suggest very good performance (i.e., AROC values approaching 1) of both methods for lead times

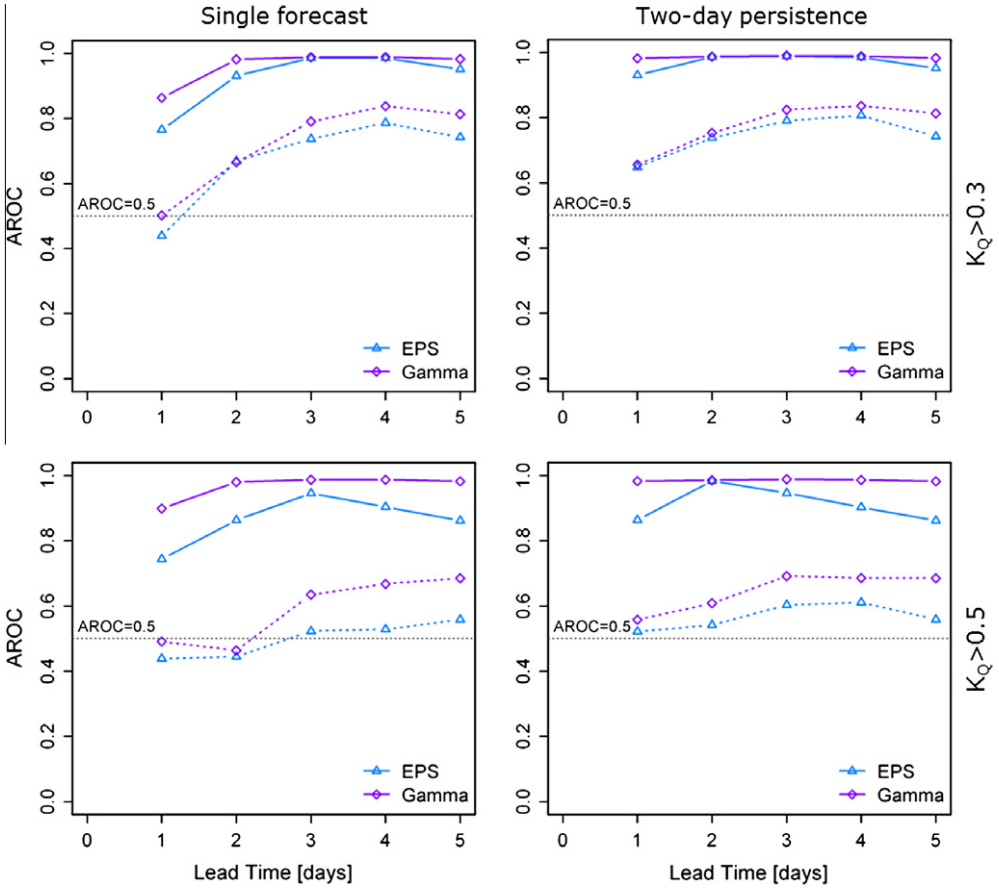

Fig. 7. Area under the ROC curve (AROC) for EPS and gamma fit, both for the full dataset (solid lines) and for high flows (dashed lines). Results are shown for two warning thresholds (top: $K_Q > 0.3$; bottom: $K_Q > 0.5$) and for single forecast (left) and 2-day persistence (right).

between 2 and 5 days, with only the 1-day lead time providing lower values, especially for the EPS. This outcome is misleading as it does not reflect the true performance in predicting threshold exceedances. In fact, it is dominated by both observed and forecasted low flow values, well below the threshold, that represent the largest proportion of the time series (98–99% as shown in Fig. 6). To correct for it, we repeated the analysis for a subset of time-steps where either the normalized observed discharge or the forecasted EPS (i.e., at least one member) was larger than the warning threshold. By its definition, this is a different (larger) subset than the one used in the quantitative analysis shown in the right panels of Fig. 5, which also enables the detection of false alarms (i.e., observed discharge below threshold along with corresponding non-zero probability in forecasted threshold exceedance). Results of the latter analysis are shown in each panel of Fig. 7 for the two probabilistic predictors, with the respective symbols connected by dashed lines. One can note in Fig. 7 (left panels) that the performance of the threshold exceedance analysis is generally higher for forecast lead times of 3–5 days, while it deteriorates for shorter lead times. The gamma fit leads to a significant improvement in comparison to the EPS, particularly for the highest of the selected thresholds (bottom-left panel).

4.5. Rank histogram

A further analysis was carried out to investigate the reason of forecast performance improving for higher lead time. Intuitively, one would think that the highest forecast skills are achieved towards the beginning of the forecasting range, though our results show the opposite. We plotted in Fig. 8 the five rank histograms, also called Talagrand diagrams (see e.g., Hamill, 2001), of the

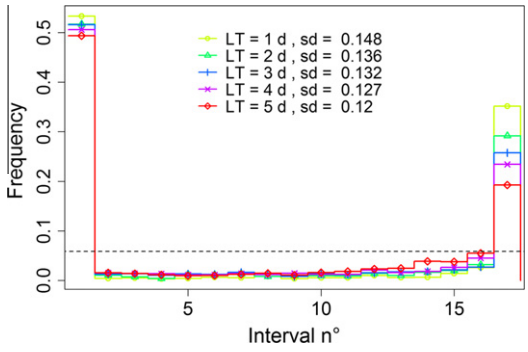

Fig. 8. Rank histograms of the normalized ensemble discharges for lead time from 1 to 5 days, together with standard deviation of the frequencies.

forecasted ensemble for the five lead times. Fig. 8 indicates a significant underdispersion of the EPS, with 68% (for 5 day lead time) to 88% (1 day lead time) of the observations falling outside the EPS range. The standard deviation (sd) of the 17 classes is shown for each line, indicating that the underdispersion of the EPS increases as the lead time decreases. In fact, optimal diagrams have values close to the uniform distribution shown by the horizontal dashed line and standard deviation approaching zero.

4.6. Lagged forecasts

Lagged forecasts have become of widespread use in many operational warning systems (Cloke and Pappenberger, 2009). Previous studies (Bartholmes et al., 2009; Thiemig et al., 2010) showed the

improved skills of persistently forecasted warnings against single forecasts by using a boolean approach. In these, a flood warning is issued when more than one consecutive meteorological forecast produces threshold exceedances of a specific quantile of the ensemble. Here, the concept of forecast persistence is applied instead in a probabilistic way and analyzed over the full range of probability thresholds through the ROC curve. The probability of exceedance is calculated from the number of forecasted discharge over threshold for each time step, taken from two consecutive forecasts. For the EPS predictor, this is equivalent to considering a new ensemble with twice the size of the original one, which means 32 members for COSMO-LEPS forecasts. Similarly, for the gamma fit, the total probability of threshold exceedance is calculated by fitting a gamma distribution for each time step and lead time, on the new ensemble of increased size (32 members).

Results of the threshold exceedance analysis with 2-day forecast persistence are shown in the right panels of Fig. 7 for two different warning thresholds, both for the full dataset (solid lines) and on the subset of discharge over threshold (dashed lines). For each lead time d in the abscissa, AROC is calculated from the forecasts at day d and $d+1$, while results for the highest lead time (i.e., 5 days) are left as in the left panels, without considering the persistence. The graphs denote a generalized benefit derived by considering the forecast persistence, which is more evident for the first two forecast lead times and in general for the EPS predictor.

5. Discussion

An overview to figures and results of different analyses allows us to draw some considerations on the performance of the estimation, both as comparison among the tested predictors for the various forecast lead times and as overall evaluation.

5.1. Model performance

Both the quantitative analysis and the threshold exceedance analysis show similar results from the 3rd to 5th day of the forecast. The performance of estimation decreases for shorter lead times, particularly in the range of high flows as shown in Figs. 5 and 7. This behavior is partly due to the underdispersion of the EPS, clearly visible in Fig. 8, which is progressively reduced as longer lead times are considered. Indeed, COSMO-LEPS was mainly designed for the medium range forecast between 3 and 5 days (Marsigli et al., 2005).

The second important reason is the uncertainty in the initial conditions as they are taken from a different hydrological model at coarser resolution. However, this is a necessary step to take for reducing the computational requirements of the system and providing timely flood alerts. Operationally, the hydrological simulation at fine resolution (1 km, 3 h) is run only for those catchments where a signal for possible upcoming floods is detected. Results in Fig. 4 confirm the findings of a recent work, showing that this approximation induces a positive bias in the simulated discharges, which becomes negligible for high flows. In addition, the initial conditions used do not represent correctly the initial water stage and discharge, as the modeled river networks at 1-km and at 5-km resolution are different. As a result, considerable errors are produced when the beginning of the forecast range occurs in between a rainfall event already started. Quantitatively, one can note in the top-right panel of Fig. 5 that the full quantile range of the EPS is underestimated for the 1-day lead time. Indeed, the bias becomes negligible only for the highest quantile, where the variance is large and affects the overall performance of estimation (*MSE*).

Weather predictions are commonly recognized to produce the largest proportion of the total estimation uncertainty. However, the uncertainty range would increase if the parameter uncertainty of the hydrological model was considered, compensating part of the streamflow variability that the meteorological input data cannot explain. Recent works showed possible improvements to systems based on LISFLOOD hydrological model by including the model parameter uncertainty (Feyen et al., 2007) and the total uncertainty through a Bayesian post processor (Bogner and Pappenberger, 2011).

5.2. Impact of spatial and temporal resolution

Some considerations on the overall performance of estimation must be addressed to the impact of spatial and temporal resolution of the hydrological model and of the input data, with regard to the scales of the process under study. Although COSMO-30y climatology enables the estimation of coherent warning thresholds, some bias of estimation can occur for high flows, depending on the space–time resolution of the precipitation input data and its relation to the catchment size and the response time. Several literature works (e.g., Berne et al., 2004; Carpenter and Georgakakos, 2004; Krajewski et al., 1991) showed that the use of precipitation data with coarse space–time resolution leads to considerable underestimation of the peak flow, especially for small catchments prone to flash flood events (Reed et al., 2007; Sangati and Borga, 2009). On the other hand, the proposed system is designed to be independent of local measurements, such as from rain-gauges, weather radars, so that it can be easily applied to a wide range of catchments in much larger domains (e.g., the whole COSMO-LEPS domain). Results of our simulations indicate that, although the overall estimation bias is a negligible portion of the total error (see Fig. 5, left column), high flows are on average underestimated (Fig. 5, right column). This is also reflected in the threshold exceedance analysis in Fig. 7 as a deterioration of results as higher warning thresholds are considered. As a result, a maximum threshold value could be derived, above which the system does not provide any added value compared to random forecasts. The limited length of the available dataset does not enable robust analyses which consider higher warning thresholds. However, this can be overcome in future analyses by expanding the test dataset through the simulation of several catchments at the same time, to include a higher number of extreme events. Such assessment is likely to show improvements of the system performance as bigger catchments are considered. Indeed, increasing the catchment area has a twofold benefit of (I) reducing the scale issues concerning the resolution of the meteorological input data versus the catchment size. (II) Similarly, location errors in the weather predictions have a reduced impact on the streamflow prediction, as the catchment area increases (see Vincendon et al., 2011).

A further source of uncertainty arises in the comparison of results, due to the temporal resolution of the observed dataset used for validation. For coherent comparison, observed discharge are resampled with the same temporal resolution of the forecasted discharge, which in EFAS-FF is set to the temporal resolution of the precipitation input data (i.e., 3 h). Catchment reaction to precipitation becomes faster as the upstream area decreases. Thus, the observed discharge can vary significantly within the time step duration, though only one value is taken as reference. In this work, 3-h values were derived by resampling hourly observations. We calculated a mean absolute variation of 3.7% among all the possible sets of three discharge values within each sampling interval. This becomes 21% if only the time steps corresponding to observed high flows are chosen (i.e., $K_{Q,O} > 0.3$), with a maximum 3-h range $\Delta K_{Q,O} = 0.67$ in the available time series. Such a difference is likely to be higher if finer resolution data were considered (e.g., 5 min observations). This limitation affects every flood warning system driven by weather predictions. Although there is no easy solution to overcome it, it is important to be considered in the evaluation of the system uncertainty, particularly for the smallest catchments.

5.3. Features of an early warning system

A final remark in the evaluation of the forecast performance is related to errors in the timing of simulated and observed high flow events. In early warning systems, when the considered forecast lead time is of the order of 3–5 days, a shift of the predicted peak flow from the observed one of ±12 h can be accepted and should not be accounted as a bad forecast (see, for example, Fig. 3). When observed and predicted discharges are compared at the same time steps, as in this work, the threshold exceedance analysis results in an increased number of both missed forecasts and false alarms, while some of them actually correspond to good forecasts. Such issue mainly affects flash floods in small catchments, where the threshold exceedance of flood hydrographs is often sudden and short-lived. Thus, the actual forecast performance, as shown in Fig. 7, would increase if a time buffer of some hours were considered in the comparison. Operationally, it translates to switching to an event-based analysis, where only simulated and observed peaks over threshold are matched in time and compared. Such analysis was not carried out here, as a larger number of events with peak discharge over threshold are needed to draw consistent evaluations. Similarly to the issue discussed in Section 5.2, this can be tackled by increasing the number of severe events, selecting them from several catchments in different regions.

6. Conclusions

This paper presents a new system for flash flood early warning, designed to monitor small to medium size catchments (up to 1000–2000 km^2) within a large portion of the European domain. The system is based on the hydrological simulation of ensemble meteorological forecasts at selected catchments, where a significant probability of upcoming severe precipitation is detected. Hydrological simulations were run in operational-mode over a 17-month test period for a case study in the southern Switzerland. Forecasted ensemble discharges are compared with observations at the catchment outlet and the system performance is assessed through quantitative evaluation and threshold exceedance analysis. Three novelties have been introduced and evaluated in the proposed approach. They are summarized below.

- The use of a simulated meteorological reforecasts dataset (COSMO-30y), consistent with operational ensemble forecasts, is used to derive a discharge climatology and, in turn, coherent warning thresholds. The broad spatial extent and the comparatively fine space–time resolution of COSMO-30y can enable unprecedented applications of such methodology for small-scale phenomena such as flash floods, at the continental level. Future work will focus on setting up such a system on the full spatial domain covered by COSMO-LEPS forecasts.
- Ensemble streamflow predictions are inferred with gamma probability distributions and tested as predictors within the flood warning system. This approach is aimed to describe continuously the spectrum of possible future evolutions and the probability linked to each value. Also, it overcomes the inconsistency of assuming a zero probability of occurrence for values outside the EPS range.
- The benefit of forecast persistence in predicting threshold exceedances is evaluated objectively. New ensembles of increased size are considered (i.e., 32 members), which include discharge prediction for the same time steps, derived from two consecutive forecasts.

Further, this study has led to important findings, which are summarized in three main points:

- The fitting of ensemble streamflows through gamma probability distributions was found to be the best choice to use within EFAS-FF. The improvements of this approach against the raw EPS are most evident in the quantitative analysis of high flows and in the threshold exceedance analysis, which is of main interest in flood warning systems. Quantitatively, the EPS and the gamma fit yield increased performances provided that appropriate probability thresholds are chosen. In this regard, the EPS mean is a valid alternative which avoid the choice of a probability threshold and provides accurate results compared to the two probabilistic predictors for a wide portion of the quantile range.
- Considering the persistence of consecutive forecasts is found to improve the performance in predicting threshold exceedances, through the use of ROC curves, particularly in the lead time range 0–48 h. This confirms the findings of previous works by means of a new objective approach.
- Results show that the space–time resolution of forecasted precipitation input is often coarse to reproduce the true variability of storms producing flash-floods. As a result, high flows were on average underestimated for the selected case study. This error would be reduced in larger catchments, where flood events are generated by storms of longer duration and broader extent, which are better captured by NWP models. At the typical scales of flash floods, uncertainties play a significant role in early warning systems. Yet, early warnings for flash floods can be extremely useful to timely intervention plans even in uncertain conditions, provided that adequate information on the related uncertainty is effectively communicated to the end users.

Despite the relatively short duration of the test period (i.e., 17 months), the continuous simulation approach allowed us to draw useful indications on the system performance, by accounting both correct predictions and false alarms. The main limiting factor was the availability of homogeneous operational weather forecasts for longer time spans, with the same model setup as the simulated climatology. In fact, these are periodically subject to updates of the space–time resolution and of the model version. The analysis of different other catchments is one possible solution to increase the number of high flow events and thus deriving robust performance analysis which consider more severe warning thresholds. Some margin of improvement is offered to the proposed approach by the recent update of COSMO-LEPS forecast model, occurred in December 2009, to a finer spatial resolution (i.e., 7 km $\times$ 7 km), that came along with an increase of the integration domain. A coherent climatology was also calculated with the same model setup, covering a 20-year time span starting in 1989.

Future analyses will focus on an event-based approach through a time window of appropriate duration, which compensates for small time lags between simulated and observed flood peak referring to the same event. This approach improves the threshold exceedance analysis both in terms of False Alarm Rate and Hit Rate, thanks to an appropriate matching of simulated and observed flood events. Thus it reflects more correctly the performance of the warning system.

References

Addor, N., Jaun, S., Zappa, M., 2011. An operational hydrological ensemble prediction system for the city of Zurich (Switzerland): skill, case studies and scenarios. Hydrol. Earth Syst. Sci. 8, 715–761.

Alfieri, L., Smith, P., Thielen, J., Beven, K., 2011a. A staggered approach to flash flood forecasting – case study in the Cevennes Region. Adv. Geosci. 29, 13–20. doi:10.5194/adgeo-29-69-2011.

Alfieri, L., Thielen, J., Smith, P., 2010. Deliverable D2.2: Flash Flood Early Warning Through Hydrological Simulation of Probabilistic Ensemble Forecasts – Downscaling the European Flood Alert System to the regional scale. IMPRINTS Project, FP7-ENV-2008-1-226555.

Alfieri, L., Velasco, D., Thielen, J., 2011b. Flash flood detection through a multi-stage probabilistic warning system for heavy precipitation events. Adv. Geosci. 29, 69–75. doi:10.5194/adgeo-29-13-2011.

Bartholmes, J., Thielen, J., Ramos, M., Gentilini, S., 2009. The European flood alert system EFAS – part 2: statistical skill assessment of probabilistic and deterministic operational forecasts. Hydrol. Earth Syst. Sci. 13, 141–153.

Berne, A., Delrieu, G., Creutin, J., Obled, C., 2004. Temporal and spatial resolution of rainfall measurements required for urban hydrology. J. Hydrol. 299, 166–179.

Bobée, B., Ashkar, F., 1991. The Gamma Family and Derived Distributions Applied in Hydrology. Springer, Littleton, CO.

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

Carpenter, T., Georgakakos, K., 2004. Impacts of parametric and radar rainfall uncertainty on the ensemble streamflow simulations of a distributed hydrologic model. J. Hydrol. 298, 202–221.

Carpenter, T., Sperfslage, J., Georgakakos, K., Sweeney, T., Fread, D., 1999. National threshold runoff estimation utilizing GIS in support of operational flash flood warning systems. J. Hydrol. 224, 21–44.

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

Cloke, H., Thielen, J., Pappenberger, F., Nobert, S., Bálint, G., Edlund, C., Koistinen, A., De Saint-Aubin, C., Sprokkereef, E., Viel, C., Salamon, P., Buizza, R., 2009. Progress in the implementation of Hydrological Ensemble Prediction Systems (HEPS) in Europe for operational flood forecasting. Hydrol. Earth Syst. Sci. 13, 125–140.

De Roo, A., Odijk, M., Schmuck, G., Koster, E., Lucieer, A., 2001. Assessing the effects of land use changes on floods in the meuse and oder catchment. Phys. Chem. Earth Pt. B. 26, 593–599.

Dietrich, J., Denhard, M., Schumann, A., 2009. Can ensemble forecasts improve the reliability of flood alerts? J. Flood Risk Manage. 2, 232–242.

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

Fundel, F., Walser, A., Liniger, M., Appenzeller, C., 2010. Calibrated precipitation forecasts for a limited-area ensemble forecast system using reforecasts. Mon. Weather Rev. 138, 176–189.

Gaume, E., Bain, V., Bernardara, P., Newinger, O., Barbuc, M., Bateman, A., Blaškovičová, L., Blöschl, G., Borga, M., Dumitrescu, A., Daliakopoulos, I., Garcia, J., Irimescu, A., Kohnova, S., Koutroulis, A., Marchi, L., Matreata, S., Medina, V., Preciso, E., Sempere-Torres, D., Stancalie, G., Szolgay, J., Tsanis, I., Velasco, D., Viglione, A., 2009. A compilation of data on European flash floods. J. Hydrol. 367, 70–78.

Hamill, T., 2001. Interpretation of rank histograms for verifying ensemble forecasts. Mon. Weather Rev. 129, 550–560.

Hamill, T., Whitaker, J., Mullen, S., 2006. Reforecasts: an important dataset for improving weather predictions. B. Am. Meteorol. Soc. 87, 33–46.

Hersbach, H., 2000. Decomposition of the continuous ranked probability score for ensemble prediction systems. Weather Forecast. 15, 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. Hydrometeorol. 11, 618–641.

Hosking, J.R.M., 1990. L-moments: analysis and estimation of distributions using linear combinations of order statistics. J. Roy. Stat. Soc. B Met. 52, 105–124.

Jolliffe, I.T., Stephenson, D.B., 2003. Forecast Verification: A Practitioner's Guide in Atmospheric Science. John Wiley and Sons.

Krajewski, W., Lakshmi, V., Georgakakos, K., Jain, S., 1991. A Monte Carlo study of rainfall sampling effect on a distributed catchment model. Water Resour. Res. 27, 119–128.

Laio, F., Tamea, S., 2007. Verification tools for probabilistic forecasts of continuous hydrological variables. Hydrol. Earth Syst. Sci. 11, 1267–1277.

Loucks, D.P., van Beek, E., 2005. Water Resources Systems Planning and Management: An Introduction to Methods, Models and Applications. UNESCO, Paris.

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

Marty, R., Zin, I., Obled, C., 2008. On adapting PQPFs to fit hydrological needs: the case of flash flood forecasting. Atmos. Sci. Lett. 9, 73–79.

Napolitano, G., Serinaldi, F., See, L., 2011. Impact of EMD decomposition and random initialisation of weights in ANN hindcasting of daily stream flow series: an empirical examination. J. Hydrol. 406, 199–214.

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

Norbiato, D., Borga, M., Dinale, R., 2009. Flash flood warning in ungauged basins by use of the flash flood guidance and model-based runoff thresholds. Meteorol. Appl. 16, 65–75.

Pappenberger, F., Cloke, H.L., Persson, A., Demeritt, D., 2011. HESS Opinions "on forecast (in)consistency in a hydro-meteorological chain: Curse or blessing?". Hydrol. Earth Sys. Sci. 15, 2391–2400.

Pappenberger, F., Thielen, J.B.A.J., Cloke, H.L., Buizza, R., Roo, A.D., 2008. New dimensions in early flood warning across the globe using grand-ensemble weather predictions. Geophys. Res. Lett. 35, L10404.

Persson, A., Grazzini, F., 2007. User Guide to ECMWF forecast products. Meteorol. Bull. 3, 153 pp.

Philipp, A., Schmitz, G., Krauße, T., Schütze, N., Cullmann, J., 2008. Flash flood forecasting combining meteorological ensemble forecasts and uncertainty of initial hydrological conditions. Aust. J. Water Resour. 12, 257–267.

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

Rotach, M., Paolo, A., Ament, F., Appenzeller, C., Arpagaus, M., Bauer, H., Behrendt, A., Bouttier, F., Buzzi, A., Corazza, M., Davolio, S., Denhard, M., Dorninger, M., Fontannaz, L., Frick, J., Fundel, F., Germann, U., Gorgas, T., Hegg, C., Hering, A., Keil, C., Liniger, M., Marsigli, C., Mctaggart-Cowan, R., Montaini, A., Mylne, K., Ranzi, R., Richard, E., Rossa, A., Santos-Muñoz, D., Schär, C., Seity, Y., Staudinger, M., Stoll, M., Volkert, H., Walser, A., Wang, Y., Werhahn, J., Wulfmeyer, V., Zappa, M., 2009. Map D-phase real-time demonstration of weather forecast quality in the alpine region. B. Am. Meteorol. Soc. 90, 1321–1336.

Sangati, M., Borga, M., 2009. Influence of rainfall spatial resolution on flash flood modelling. Nat. Hazard. Earth Sys. 9, 575–584.

Stephens, M.A., 1974. EDF statistics for goodness of fit and some comparisons. J. Am. Stat. Assoc. 69, 730–737.

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

Thiemig, V., Pappenberger, F., Thielen, J., Gadain, H., de Roo, A., Bodis, K., Del Medico, M., Muthusi, F., 2010. Ensemble flood forecasting in Africa: a feasibility study in the Juba-Shabelle river basin. Atmos. Sci. Lett. 11, 123–131.

Uppala, S., Kållberg, P., Simmons, A., Andrae, U., da Costa Bechtold, V., Fiorino, M., Gibson, J., Haseler, J., Hernandez, A., Kelly, G., Li, X., Onogi, K., Saarinen, S., Sokka, N., Allan, R., Andersson, E., Arpe, K., Balmaseda, M., Beljaars, A., van de Berg, L., Bidlot, J., Bormann, N., Caires, S., Chevallier, F., Dethof, A., Dragosavac, M., Fisher, M., Fuentes, M., Hagemann, S., Hólm, E., Hoskins, B., Isaksen, L., Janssen, P., Jenne, R., McNally, A., Mahfouf, J., Morcrette, J., Rayner, N., Saunders, R., Simon, P., Sterl, A., Trenberth, K., Untch, A., Vasiljevic, D., Viterbo, P., Woollen, J., 2005. The ERA-40 re-analysis. Q. J. Roy. Meteor. Soc. 131, 2961–3012.

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

Vincendon, B., Ducrocq, V., Nuissier, O., Vié, B., 2011. of convection-permitting NWP forecasts for flash-flood ensemble forecasting. Nat. Hazard. Earth Syst. 11, 1529–1544.

Vogel, R., Fennessey, N., 1993. L moment diagrams should replace product moment diagrams. Water Resour. Res. 29, 1745–1752.

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

Wöhling, T., Lennartz, F., Zappa, M., 2006. Technical Note: real-time updating procedure for flood forecasting with conceptual HBV-type models. Hydrol. Earth Syst. Sci. Disc. 33, 925–940.

Younis, J., Anquetin, S., Thielen, J., 2008. The benefit of high-resolution operational weather forecasts for flash flood warning. Hydrol. Earth Sys. Sci. 12, 1039–1051.