Journal of Hydrology 517 (2014) 913–922

Contents lists available at ScienceDirect

Journal of Hydrology

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

Evaluation of ensemble streamflow predictions in Europe

Lorenzo Alfieri [a,b,*], Florian Pappenberger [a], Fredrik Wetterhall [a], Thomas Haiden [a], David Richardson [a], Peter Salamon [b]

[a] European Centre for Medium-Range Weather Forecasts, Reading, UK
[b] European Commission – Joint Research Centre, Ispra, Italy

ARTICLE INFO

Article history:
Received 2 December 2013
Received in revised form 7 May 2014
Accepted 21 June 2014
Available online 30 June 2014
This manuscript was handled by Konstantine P. Georgakakos, Editor-in-Chief, with the assistance of Emmanouil N. Anagnostou, Associate Editor

Keywords:
Flood early warning
Ensemble streamflow predictions
CRPS
Skill scores
Distributed hydrological modelling

SUMMARY

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

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

1. Introduction

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

Assessing the forecast performance over large domains raises the challenge of comparing river points with different upstream area and hydrological regimes. In these cases, a widespread approach to tackle the forecast verification is to compute scores based on the probability of thresholds exceedance (e.g., warning levels), that can be defined in a consistent way for every point. While this is a standard practice for early warning systems (e.g., Bartholmes et al., 2009; Gourley et al., 2012), it is also applied to the verification of categorical events for any set of thresholds (Thirel et al., 2008). If quantitative values are considered, the choice of performance scores becomes wider (Legates and McCabe, 1999; Wilks, 2006), though only a relatively small subset is specifically dedicated to evaluate the quality of ensemble forecasts (Brown et al., 2010). The comparison of forecast skill in several river sections is often performed through benchmarking against simplified simulations (Pappenberger et al., submitted), previous model versions (Arheimer et al., 2011), different input

* Corresponding author at: European Commission – Joint Research Centre, TP 122, Via E. Fermi 2749, 21027 Ispra, VA, Italy. Tel.: +39 0332 78 3835.
E-mail address: lorenzo.alfieri@jrc.ec.europa.eu (L. Alfieri).

http://dx.doi.org/10.1016/j.jhydrol.2014.06.035
0022-1694/© 2014 The Authors. Published by Elsevier B.V.

data (e.g., Renner et al., 2009), or climatological values (Demargne et al., 2010; Verkade et al., 2013; Wood et al., 2005). An alternative method consists in normalizing forecasts and reference values before the evaluation (Pappenberger et al., 2010). Trinh et al. (2013) used a similar concept to propose a modified Continuous Ranked Probability Score (CRPS) which is suitable to compare forecast performance at different river sections. In operational systems, the forecast performance must be monitored and updated continuously in time. Hence, a skill assessment based on different scores and benchmarks (e.g., Alfieri et al., 2013a; Randrianasolo et al., 2010) is often preferred in order to analyze different aspects of the forecast performance at several locations and quickly detect trends over time or weaknesses.

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

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

2. Data and methods

2.1. Model framework

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

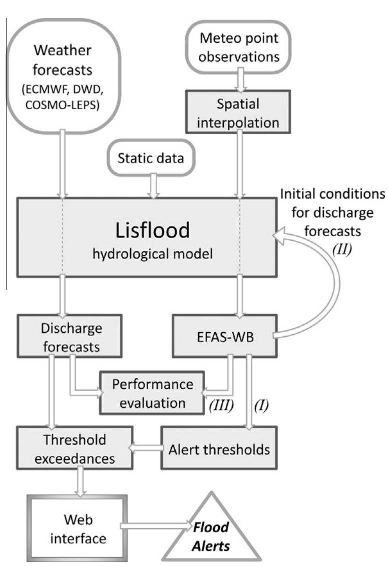

Fig. 1. Schematic view of the EFAS hydro-meteorological forecasting system.

2.2. Meteorological data

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

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

2.3. Hydrological modelling

In EFAS, hydrological simulations are performed with Lisflood, a hybrid between a conceptual and a physical rainfall–runoff distributed model, designed to reproduce the main hydrological processes of medium to large river basins (see van der Knijff et al., 2010). The considered model setup for Europe was calibrated at 481 river gauges, using the observed meteorological fields as input and up to 7 years of gauged discharge. A reference hydrological simulation starting in 1990 was run for the European window with the calibrated Lisflood model at 5×5 km resolution, using the observed meteorological fields as input. The operational model is updated daily using the initial states of the previous day and the most recent meteorological observations acquired with about 1 day lag. This simulation, hereafter referred to as EFAS Water Balance (EFAS-WB), represents our best estimate of the hydrological states in the European rivers. The EFAS-WB is used in EFAS with regard to three main aspects (see Fig. 1): (I) deriving climatological