Journal of Hydrology 522 (2015) 697–713

Contents lists available at ScienceDirect

Journal of Hydrology

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

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

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

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

ARTICLE INFO

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

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

SUMMARY

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

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

1. Introduction

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

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

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

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

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

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

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

1.1. Which benchmark?

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

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

1.2. Benchmarks for hydrological ensemble predictions

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

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

1.3. Aim and scope of the paper

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

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