Hydrological Sciences Journal

ISSN: 0262-6667 (Print) 2150-3435 (Online) Journal homepage: https://www.tandfonline.com/loi/thsj20

Evaluation of a global ensemble flood prediction system in Peru

Konstantinos Bischiniotis, Bart van den Hurk, Ervin Zsoter, Erin Coughlan de Perez, Manolis Grillakis & Jeroen C. J. H. Aerts

To cite this article: Konstantinos Bischiniotis, Bart van den Hurk, Ervin Zsoter, Erin Coughlan de Perez, Manolis Grillakis & Jeroen C. J. H. Aerts (2019) Evaluation of a global ensemble flood prediction system in Peru, Hydrological Sciences Journal, 64:10, 1171-1189, DOI: 10.1080/02626667.2019.1617868

To link to this article: https://doi.org/10.1080/02626667.2019.1617868

Published online: 27 Jun 2019.

HYDROLOGICAL SCIENCES JOURNAL
2019, VOL. 64, NO. 10, 1171–1189
https://doi.org/10.1080/02626667.2019.1617868

Evaluation of a global ensemble flood prediction system in Peru

Konstantinos Bischiniotis[a], Bart van den Hurk[a,b], Ervin Zsoter[c], Erin Coughlan de Perez[a,d,e], Manolis Grillakis[f] and Jeroen C. J. H. Aerts[a]

[a]Institute for Environmental Studies (IVM), Vrije Universiteit, Amsterdam, the Netherlands; [b]Deltares, Delft, the Netherlands; [c]European Centre for Medium-Range Weather Forecasts (ECMWF), Reading, UK; [d]International Research Institute for Climate and Society, Earth Institute, Columbia University, Palisades, New York, USA; [e]Red Cross Red Crescent Climate Centre, the Hague, the Netherlands; [f]Department of Environmental Engineering, Technical University of Crete, Chania, Greece

ABSTRACT

Flood early warning systems play a more substantial role in risk mitigation than ever before. Hydrological forecasts, which are an essential part of these systems, are used to trigger action against floods around the world. This research presents an evaluation framework, where the skills of the Global Flood Awareness System (GloFAS) are assessed in Peru for the years 2009–2015. Simulated GloFAS discharges are compared against observed ones for 10 river gauges. Forecasts skills are assessed from two perspectives: (i) by calculating verification scores at every river section against simulated discharges and (ii) by comparing the flood signals against reported events. On average, river sections with higher discharges and larger upstream areas perform better. Raw forecasts provide correct flood signals for 82% of the reported floods, but exhibit low verification scores. Post-processing of raw forecasts improves most verification scores, but reduces the percentage of the correctly forecasted reported events to 65%.

ARTICLE HISTORY
Received 1 August 2018
Accepted 28 March 2019

EDITOR
A. Castellarin

ASSOCIATE EDITOR
K. Kochanek

KEYWORDS
flood; early warning; forecast; bias-correction; risk; ensemble streamflow predictions

1 Introduction

Riverine floods rank among the most frequent, deadly and damaging natural hazards worldwide (Guha-Sapir *et al.* 2012; Wallemacq *et al.* 2015). Climate change, population increase and economic growth have led to an upward trend in the damages they have caused (Tanoue *et al.* 2016, Arnell and Lloyd-Hughes 2014, Hirabayashi *et al.* 2013; IPCC 2012), exceeding USD 1 trillion globally over the period 1980–2013 (Munich Re 2015). The losses caused by floods are lower in high-income countries thanks to the long term, high protection levels these countries usually adopt (Doocy *et al.* 2013). Lower-income countries cannot afford similar risk mitigation policies and are more often dependent on ex-ante preparedness measures to reduce risks. Therefore, the development and effective implementation of early warning systems is one of the first priorities of both governmental and humanitarian organisations in such flood-prone countries.

While many early warning systems are solely based on weather predictions (Alfieri *et al.*, 2012), streamflow forecasting systems have started to play a key role in flood preparedness. Recently developed models run operationally in several spatial scales and show increasing potential for being incorporated into (flood-) risk management (Roulin 2006; Dale *et al.* 2014). However, due to the inherent uncertainties in hydrological predictions, probabilistic, rather than deterministic, streamflow forecasts are usually preferred (Palmer 2001), as their refined estimates are valuable in risk-based decision-making (Raiffa and Schlaifer 1961, Laio and Tamea 2006, Todini 2007, Verbunt *et al.* 2007). These prediction uncertainties are often quantified and expressed by means of ensemble streamflow prediction (ESP) systems, which have gained popularity in recent years (Cloke and Pappenberger 2009, Wetterhall *et al.* 2013). Ensemble streamflow predictions of large-scale hydrological models have demonstrated their capabilities in historical flood events (e.g. De Roo *et al.* 2003, Gouweleeuw *et al.* 2005, Pappenberger *et al.* 2005, Webster *et al.* 2011) and have also been used operationally to trigger humanitarian action (Coughlan De Perez *et al.* 2015). Their continuous performance evaluation over different domains and temporal scales is not only necessary to gain the trust of the end users (Wetterhall *et al.* 2013), but is also essential to guide their further improvement.

The performance assessment in probabilistic streamflow forecasts is more complicated than in deterministic ones, since observed events are compared against forecast probabilities (Bartholmes *et al.* 2008). For this reason, a wide variety of quantitative verification scores exists (Brown *et al.* 2010). However, since no single score contains all the necessary information for a complete skill evaluation (Bartholmes *et al.* 2008), there is a need for a careful selection of skill scores that examine different aspects of forecast attributes and are subject to the end user's needs (Franz *et al.* 2003, Clark and Hay 2004, Roulin and Vannitsem 2005, Randrianasolo *et al.* 2010, Alfieri *et al.* 2013). Most skill scores are calculated by comparing the threshold exceedance probabilities for each point or section of a river system to the observed discharges (Bartholmes *et al.* 2008, Gourley *et al.* 2012). However, such a statistical approach only works when the record of in situ discharges covers a sufficiently long time (Hannah *et al.*

CONTACT Konstantinos Bischiniotis kbischiniotis@gmail.com

2011). When this is not the case (e.g. in data-sparse areas), forecasted streamflows are compared to modelled ones, which are used as proxies for observations to compute the so-called "*model's theoretical skill*" (Thiemig *et al.* 2010, Alfieri *et al.* 2014, Candogan Yossef *et al.* 2017). This approach provides the flexibility to assess the skill of the probabilistic streamflow forecast for each grid point on the river network, rather than for single points of observed data (Thielen *et al.* 2009). However, since modelled streamflows can significantly differ from the real ones, the theoretical skill mainly demonstrates the skill of the precipitation forecast, with a reference precipitation the forecast used to initiate the model. Therefore, to optimise its usefulness, the theoretical skill should be carefully interpreted by the end-users in real-world applications.

Except for possible deficiencies in the model structure and model parameterisation, forecast predictive skill is primarily affected by errors in initial hydrological states and meteorological forcing (Hamill *et al.* 2008), which vary according to location, season and lead time (Bierkens and van Den Hurk 2007, Shukla and Lettenmaier 2011). Both have improved in recent years thanks to the increasing number of reporting stations, the continuous assimilation of hydrological observations (Candogan Yossef *et al.* 2017), the evident progress in meteorological forcing quality (McBride and Ebert 2000, Hamill *et al.* 2008) and the development of effective bias-correction methods performed on meteorological inputs (e.g. Kang *et al.* 2010).

Ensemble streamflow prediction output results are often subject to biases and post-processing techniques are often applied to reduce these biases (Kang *et al.* 2010). A popular method is to recalibrate the model results to reproduce the climatological distribution (Madadgar *et al.* 2014). Other post-processing techniques include event bias correction (Smith *et al.* 1992), LOWESS regression (Cleveland 1979), variance inflation (INFL) method (Fundel and Zappa 2011, Roulin and Vannitsem 2015), ensemble copula coupling (ECC) (Schefzik *et al.* 2013, Bellier *et al.* 2018) and quantile mapping (Wood and Lettenmaier 2006, Baigorria *et al.* 2007). Although all of them improve forecast quality significantly (Kang *et al.* 2010), Hashino *et al.* (2007) have shown that the choice of the optimal technique is subject to application requirements.

Despite the known uncertainties and limitations of ESP forecast models, global scale systems such as the global flood awareness system (GloFAS) (Alfieri *et al.* 2013) are being used by humanitarian responders, even without any post-processing, especially in countries with limited flood forecasting systems (Coughlan de Perez *et al.* 2016). An example is Peru, which has experienced several devastating flood events over the past years and is subject to large climatological gradients in flood risk characteristics. Therefore, Peru has received increasing attention from humanitarian organisations aiming to reduce flood impacts. For example, the forecast-based financing project is being applied in the north-western regions, using GloFAS forecasts to trigger humanitarian action (Coughlan De Perez *et al.* 2015).

The aim of the current study is to assess the skills of GloFAS using a holistic evaluation framework in Peru for the years 2009–2015. The simulated discharges produced by the model are compared against 10 river gauges that are located in different regions of the country. The predictive capability of GloFAS is assessed in a so-called hindcast mode, using operational daily forecasts over a lead time (LT) of 1 to 15 days, from two complementary perspectives: (a) by calculating several verification scores at every grid point of the river network, comparing the forecasted discharge to the simulated one, and (b) through an event-based analysis comparing the GloFAS flood signals against collected information from multiple observational disaster databases. The first forecast skill assessment provides a spatial and temporal indication mainly of the meteorological forecast skills that are used as input by GloFAS. The second one compares the forecast information to reported damaging events. After these two assessments had been carried out, the raw streamflow forecasts were post-processed using the quantile mapping technique (Madadgar *et al.* 2014) to evaluate whether a simple removal of biases could increase the forecast skill and whether this would lead to better preventive flood risk management planning by humanitarian organisations and decision makers.

The paper is organised as follows: In Section 2, we describe the methodological framework and the data used; in Section 3, we present the results; and, finally, Section 4 presents the concluding remarks and discusses the findings and the limitations of this study.

2 Data and methods

2.1 Study area

Peru is a country highly vulnerable to flood events. More than 2000 people have lost their lives due to floods over the period 1980–2013, while the reported damages exceeded US$2 billion (Munich Re 2015). More recently, the catastrophic 2017 floods caused more than US$3 billion worth of damage to infrastructure and houses and over 100 casualties (El Comercio 2017). The country has pronounced spatial gradients of climatological precipitation and aridity (Fig. 1(a)) (Peel *et al.* 2007). The coastal areas are characterised by desert and semi-arid climates, the southern areas by humid sub-tropical climates and the central and northern areas by equatorial and tropical climates. The northern coastal region of Peru experiences more influence from El Niño (ENSO) (Bayer *et al.* 2014). Its recurrent nature and its relationship with flooding in that area has led to the development of an ENSO-based index insurance (Khalil *et al.* 2007).

In Figure 1(b), we demonstrate the mean annual discharges for the main river network as calculated by GloFAS. The highest discharges are seen in the northern, eastern and central part of the country (particularly at the Amazon, Rio Napo, Huallaga and Ucayali rivers), while the lowest discharges are found in the relatively short rivers of the coastal areas. Figure 1(c) shows the administrative separation into regions and the 10 gauging stations that are used for the comparison of the simulated and the observed discharge (see Section 2.4.1). Figure 1(d) shows the mean annual precipitation (Balsamo *et al.* 2015), with the highest precipitation observed in the north-eastern part and some places areas in Puno region, close to lake Titicaca, and the lowest along the coast.

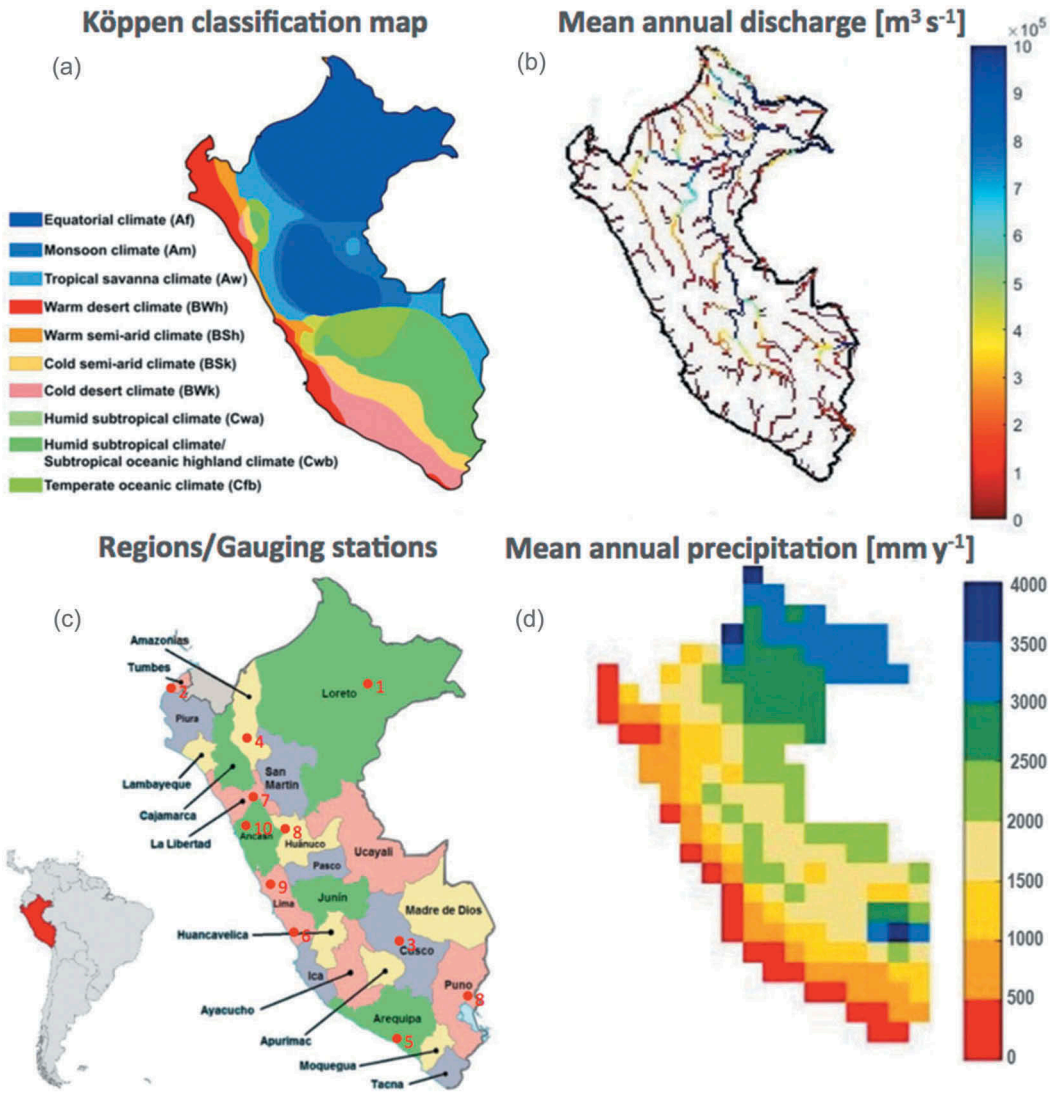


Figure 1. Overview of the study area: (a) Köppen classification map of Peru (Peel *et al.* 2007), (b) mean annual discharge (m^3/s) in GloFAS (Alfieri *et al.* 2013), (c) regions of Peru and gauging stations, and (d) mean annual precipitation (mm) (adopted from Boelee *et al.* 2017).

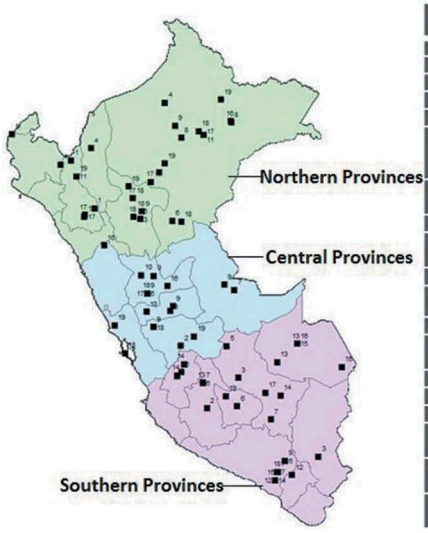
2.2 Flood archives

Natural disaster databases lack standardised procedures in flood monitoring and therefore, there are wide disparities in the number of disasters observed (Gall *et al.* 2009, Wirtz and Below 2009). In order to compile an event-based analysis, we combined information derived from various disaster databases; the Dartmouth Flood Observatory (DFO) (Brakenridge 2015), the NatCatService (Munich Re 2015), the Emergency Events Database (EM-DAT) (Guha-Sapir *et al.* 2014) and the Reliefweb.[1] From this combined database, 19 medium- to large-scale riverine floods were identified in Peru in the period 2009–2015. The most catastrophic event was the January 2010 flood, during which damages exceeded US$2 billion and 26 people lost their lives (Munich Re 2015). Information regarding the flood dates, the dataset used and the affected locations of each event are shown in Fig. 2. According to the datasets, most events affected more than one region. To increase our sample size, we considered each flood in an administrative region as an individual event, leading to a total of 61 region/flood combinations.

2.3 Model framework

The Global Flood Awareness System (GloFAS) (Alfieri *et al.* 2013) is designed for flood early warning purposes and compares ensemble streamflow forecasts to climatological distributions at a global scale. The warnings are produced on a daily basis and are freely available online.[2] Although other global streamflow forecast models exist (Sperna Weiland *et al.* 2010, Wang *et al.* 2011, Candogan Yossef *et al.* 2012), to our knowledge, only GloFAS is used to trigger humanitarian action (Coughlan de Perez *et al.* 2016), having shown its potential before the August 2013 floods in Pakistan and the September 2013 floods in Sudan.

[1] https://reliefweb.int/.
[2] http://www.globalfloods.eu/.



ID	Onset	End	Year	Affected Locations	(N)	(D)	(E)	(R)	Affected Provinces
1	1 Jan	30 Jan	2009	NW	x				Cajamarca, Loreto
2	12 Dec	19 Dec	2009	C, SW	x	x		x	Junín, Ayacucho
3	21 Jan	12 Feb	2010	C, S	x	x		x	San Martín, Huánuco, Cusco, Puno
4	10 Feb	17 Feb	2010	N, NW		x		x	Piura, Amazonas, Loreto
5	2 Mar	4 Mar	2010	C		x		x	Cusco
6	3 Jan	17 Jan	2011	C, S, SW	x		x		San Martín, Loreto, Huánuco, Pasco, Huancavelica, Ayacucho, Apurimac, Arequipa
7	10 Feb	20 Mar	2011	C, S, SW	x	x			Ucayali, Huancavelica, Ayacucho, Cusco, Arequipa
8	24 Apr	25 Apr	2011	NE	x				Loreto
9	14 Feb	11 May	2012	Entire country	x		x	x	Loreto, San Martín, Ucayali, Huánuco, Pasco, Junín, Arequipa
10	28 Dec	30 Dec	2012	C	x	x	x		Huánuco, Pasco
11	7 Feb	12 Feb	2013	Entire country	x	x	x	x	Cajamarca, Loreto, Arequipa
12	10 Mar	17 Mar	2013	S		x			Arequipa, Moquegua
13	1 Jan	8 Jan	2014	C, S	x	x		x	Huánuco, , Madre de Dios, Ayacucho, Arequipa
14	20 Jan	22 Jan	2014	S	x			x	Huancavelica, Cusco, Arequipa
15	26 Jan	11 Feb	2014	SE	x		x		Madre de Dios, Arequipa
16	22 Feb	5 Mar	2014	NE, C, W	x				Ancash, Loreto, Madre de Dios, Huánuco
17	10 Mar	23 Jun	2014	C, N, W	x			x	Loreto, San Martín, Cajamarca, Huánuco, Cusco
18	31 Jan	31 May	2015	Entire country	x		x	x	Loreto, San Martín, Huánuco, Callao, Junín, Arequipa
19	1 Dec	31 Dec	2015	C, N, W	x			x	Loreto, Cajamarca, San Martín, Lima, Junín, Apurimac

Figure 2. Flood events in Peru from 2009 to 2015, as reported in NatCatService (N), DFO (D), EM-DAT (E) and Reliefweb (R). In the Affected Locations column the meaning key to symbols is: N: north, C: centre, S: south, W: west, NW: north-west, NE: north-east, SW: south-west.

Numerical ensemble weather predictions are used as an input for the GloFAS hydrological simulations. Every day, an ensemble of 51 streamflow forecasts is produced over a LT of 30 days. Meteorological forcing is limited to the first 15 days. Observation estimates are not used to initialise the model. Instead, this is done by using the meteorological data of day 0. Discharge simulated for the last 15 days is derived from the water routing of the overland flow produced in the first 15 days. Figure 3 provides an overview of the GloFAS structure. The reference climatology of the meteorological data is produced by ECMWF's global atmospheric reanalysis ERA-Interim (Dee *et al.* 2011). Daily ensemble forecasts of meteorological parameters are computed by its integrated forecast system (IFS) (Miller *et al.* 2010), whose main components are a data assimilation system (DAS) and a general circulation model (GCM). The vertical water fluxes are transformed to run-off using the HTESSEL[3] model (Balsamo *et al.* 2011) and subsequently the Lisflood model (Van Der Knijff *et al.* 2010) simulates the horizontal water fluxes along the river network on a daily basis and on a resolution of 0.1°. In this way, the ensemble streamflow predictions for each grid point of the river network are made. By comparing the ensembles with reference thresholds that are derived from the simulated discharge climatology, flood alerts are issued (for further information about GloFAS, see Alfieri *et al.* 2013). The GloFAS system has been rigorously validated against observed daily flow data from 620 stations globally, for watersheds ranging between 450 and 4 680 000 km^2. In 58% and 60% of these stations, the Nash-Sutcliffe efficiency (Nash and Sutcliffe 1970) and the coefficient of variation were skilful. In Peru, the validation was done for only a few stations in the eastern part of the country, but further details are not publicly available.

2.4 Evaluating glofas skills

We used daily hydrological hindcasts from 1 January 2009 to 31 December 2015. These hindcasts are produced for each point of the Peruvian river network with an upstream area greater than 2000 km^2 (2780 points in total). The evaluation strategy of GloFAS skills was split into three parts. Initially, we compared the simulated discharge of 0 days LT (for simplicity, referred to as "simulated discharge") with observed discharge, wherever this was available. Subsequently, we compared the ensemble streamflow forecasts to the simulated discharge and then evaluated GloFAS forecasts against reported, damaging flood events.

2.4.1 Comparison of simulated and observed discharge

The simulated daily discharge that was produced by GloFAS was compared with observed daily discharge for 10 stations that are located in different regions of Peru and include discharges during our study time period. This data was obtained from the website of the Peruvian National Water Authority.[4] The agreement between observations and simulated discharge was estimated using the Nash-Sutcliffe efficiency (NS) (Nash and Sutcliffe 1970) (Appendix A, equation (A1)). This score is an objective function for reflecting the overall fit of a hydrograph (Servat and Dezetter 1991), indicating how well the observed temporal variability is reproduced by the simulations (Moriasi *et al.* 2007). Furthermore, a threshold of the 90th percentile of the daily times series was considered for each station data for the observed and the simulated discharges explicitly. We selected this percentile to indicate a flood event, similar to (Alfieri *et al.* 2013) and, at the same time, bring our evaluation sample to a size that allows representative and statistically meaningful conclusions about the performance of the system. Then, the GloFAS simulated discharges were compared against the observed discharges day-by-day to calculate the number of times both thresholds were exceeded (Hit: H) and the number of times that the simulated discharge was not in agreement with the observed one (Miss: M), thus calculating the probability of detection (POD), which shows the proportion of successfully detected events:

[3]Tiled ECMWF Scheme for Surface Exchanges over Land, revised for Hydrology.
[4]http://snirh.ana.gob.pe/visors2/

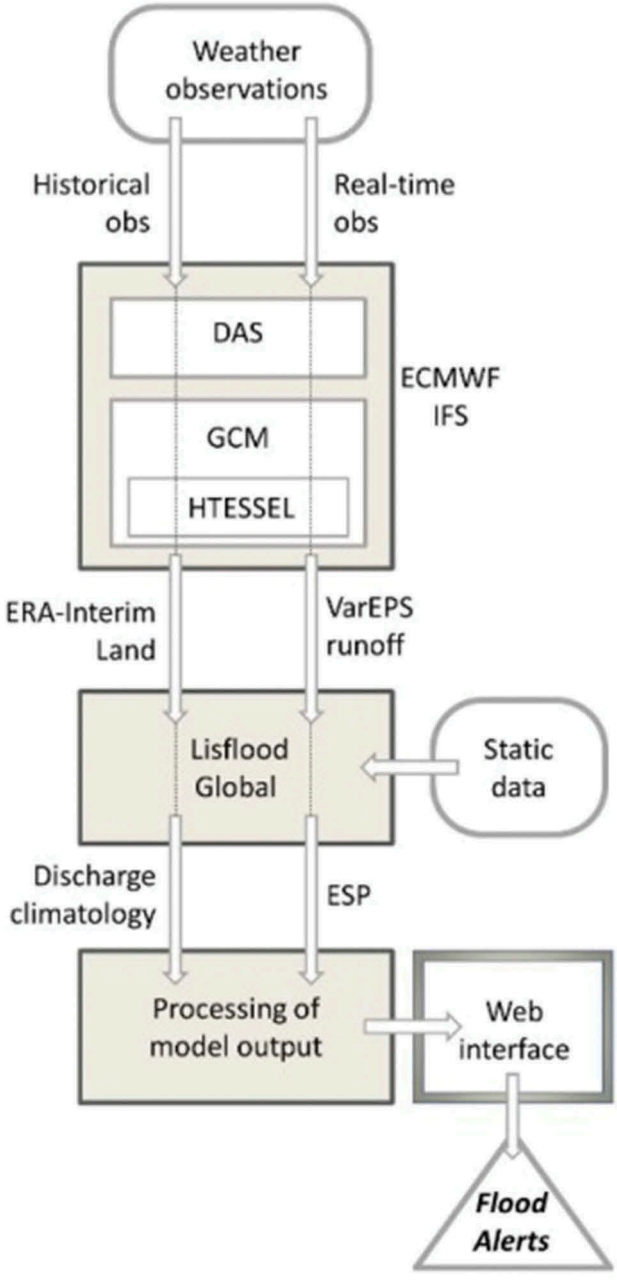


Figure 3. Overview of GloFAS structure. DAS: data assimilation system, GCM: general circulation model, IFS: integrated forecast system, ESP: ensemble streamflow prediction. Adopted from (Alfieri *et al.* 2013).

$$\mathbf{POD} = \frac{H}{H + M} \tag{1}$$

2.4.2 Forecasting simulated discharge

The ability of GloFAS to forecast the simulated discharge was assessed. This procedure mainly evaluates how errors in precipitation prediction affect the simulation. For that, we calculated several verification scores over the period 2009–2015 for which hindcasts were available. We calculated the NS efficiency, first, based on average calculated climatological discharges similarly to Alfieri *et al.* (2014) (Appendix A, equation (A2)), and second, based on a persistent forecast (NS_{pf}) (Appendix A, equation (A3)) (Plate and Lindenmaier 2008). In hydrological forecasting, the latter acknowledges the role of initial conditions, making it particularly useful for slowly varying rivers, it is independent of seasonal variations and is usually "harder to beat" than the NS that uses climatological discharge as a benchmark.

The other scores used are the coefficient of variation (CV) and the percentage bias (Pbias), which are calculated based on the ensemble mean, and the continuous ranked probability skill score (CRPSS), which takes into account the entire ensemble. These scores are briefly described below and discussed analytically in Appendix A.

- Percentage bias (Pbias): This score measures the average tendency of the forecast values to be smaller or greater than the observed ones and has the ability to indicate systematic model deficiencies (Gupta *et al.* 1999). Positive and negative values show that forecasts under- and over-estimate discharge, respectively (see Appendix A, equation (A4)).
- Coefficient of variation of the root mean squared error (CV): This score is used to measure the standard deviation between the forecast and the observed values, while allowing comparison between river cells with very different discharges (Reed *et al.* 2007) (Appendix A, equation (A5)).
- Continuous ranked probability skill score (CRPSS): This score, proposed by Hersbach (2000), evaluates the probabilistic skill of the forecast, measuring the weighted average skill over a range of discrete threshold levels for which exceedance probabilities are computed (Bradley and Schwartz 2011) (Appendix A, equations (A6)–(A8)).

All river cells of the Peruvian territory were aggregated to create separate boxplots for each LT to gain insights into how the skill of GloFAS varies temporally. Subsequently, we plotted the skill scores on the map using LT7 (7-day lead time), to obtain a spatial overview of the model performance.

Although other LTs could have been used, LT7 was chosen as it is in the middle of the meteorological forcing period and provides a sufficient time window for preparation, should a large event be forecast. Finally, we classified the river cells into 11 groups, based on the size of their upstream area. For each of these groups, we created boxplots to demonstrate the effect of the upstream area size on forecasting skills.

2.4.3 Forecasting observed flood events

An event-based analysis was carried out to evaluate the ability of GloFAS to provide accurate flood warnings based on reported, damaging events. Large-scale model skills are usually evaluated against individual big events, e.g. the Pakistan 2010 floods (Alfieri *et al.* 2013), or evaluated for a short time period, e.g. African floods in 2003 (Thiemig *et al.* 2015). In this paper, we evaluated the performance of GloFAS against the recorded riverine floods of the examined period (2009–2015).

The GloFAS flood signals were compared to the reported floods obtained from the disaster databases (Section 2.2) to calculate the POD.

As in Section 2.4.1, the 90th percentile was used as a threshold to transform the ensemble forecast into

a forecast of a dichotomous event (i.e. Flood/No Flood) for each river cell. We used forecasts up to LT15, the forecast range for which meteorological forcing is applied. For this period and for any given cell, a flood signal is identified if (a) there is at least 50% probability that the discharge on a -specific day will (b) exceed the 90th percentile discharge for (c) at least three consecutive forecasts, similarly to Thiemig *et al.* (2015). If 10% of the river cells in an administrative region meet these criteria, a flood signal is counted. If a flood was reported in this region, there was a correct hit (H); otherwise, there was a miss (M).

2.4.4 Post-processing: quantile mapping

As a post-processing technique, we applied quantile mapping to correct the systematic distributional biases, which come from the precipitation forecast that is used as an input in the model. Déqué (2007) used the same methodology, referring to it as "simple unbiasing" and applying it to the entire cumulative distribution function (cdf) to correct a regional climate scenario.

In our case, for each LT, the cumulative probability distribution of all forecast ensemble members was transformed in order to match the daily cumulative distribution of the simulated discharge. This adjusts the forecast distributions, but the temporal variability of the mean and variability of the ensemble forecasts are largely retained. Each river cell and each forecast LT is treated individually, addressing LT and space-dependent systematic biases. First, we derived the reference cdf based on the simulated discharges. Subsequently, we created cdfs of the forecast discharges for each ensemble member and LT. These are rescaled to the reference cdf, correcting the discharge probabilities at intervals of 1%. The advantage of dealing with each LT independently is that the biases at different LTs can be corrected (e.g. when the forecast tends to predict higher discharges in short LTs and lower discharges in long LTs). The procedure is described by:

$$\mathrm{Mod}_i = \mathrm{Raw}_i + \left(\overline{\mathrm{Ref}_i} - \overline{\mathrm{Raw}_i}\right) \quad (4)$$

where $\mathrm{Raw}_i$and Mod_i are, respectively, the raw and post-processed data of the ith percentile (from all ensemble members), while $\overline{\mathrm{Ref}_i}$ and $\overline{\mathrm{Raw}_i}$ are the average values for the ith percentile of the reference and raw data, respectively (from all ensemble members).

We repeated the procedure of sections 2.4.2 and 2.4.3 to calculate the verification scores and the event-based metrics using the post-processed discharges and allowing inspection of the effect of the bias adjustment on the forecasting skill.

The post-processing methodology was also applied to each gauging station discharge and the NS obtained in Section 2.4.1 was recalculated. In this case, the reference discharge was the observed one, and the transformed discharge was the simulated one.

3 Results

3.1 Discharge validation using gauging stations

Figure 4 displays the hydrographs of the observed discharge, the raw simulated discharges and the post-processed discharge for the 10 gauging stations. As can be seen from Fig. 4, only two stations have observational discharges that allow a comparison for the entire study period (2009–2015). Through comparison of the raw and the post-processed discharge data, we observe an improvement when post-processing is applied. This improvement is also depicted by the NS values shown in Table 1. In the case of raw forecasts, the results show a fair performance based on NS, considering that the GloFAS model is not calibrated. It is also shown by the NS that the post-processing was highly beneficial for all stations except for the one in Amazonas, which has the lowest mean annual discharge. Given the explicit estimation of the 90th percentile on the simulated and observed discharge, the POD is the same for both the raw and post-processed discharge data, ranging between 0.27 and 0.73, with a weighted average of 0.46.

3.2 Forecast versus simulated discharge

3.2.1 Performance versus lead time

Figure 5 displays the NS values when the mean value of the simulated discharges (NS) and the persistent forecasts ($\mathrm{NS_{pf}}$) are used as benchmarks for both the raw (left) and the post-processed forecasts (right) for each LT of the forecast range. The median of all river points of the Peruvian river network is shown by a horizontal line, the boxes represent 25–75% and the whiskers 1–99%. Outliers are not plotted for better readability of the graphs. In both the raw and post-processed forecasts, the median of the NS decreases with LT, demonstrating a decrease in forecasting skill. Regarding the $\mathrm{NS_{pf}}$, its median is always below 0 for all LTs when using the raw forecasts, demonstrating that, in most cases, persistent forecasts would be more useful than the forecasts of the model. However, the median of the post-processed forecasts is above 0 after a lead time of 5 days (LT5), showing that, at longer ranges and after post-processing, the model is more skilful than a persistent forecast.

In Figure 6 (left), it may be seen that the median of CV increases and CRPSS decreases with LT, demonstrating a decrease in forecasting skill at longer LTs for raw forecasts. The negative values in Pbias clearly show that the forecasts in most river cells produce much higher discharge values than simulated climatology at all LTs. A large fraction of all river cells (75%) displays CRPSS > 0 up to LT12, showing that these cells perform better than simulated climatology. The variability of all verification scores increases over LT.

The results for the post-processed forecasts (Fig. 6, right) display a decrease in the variability. Similarly to the raw forecasts, the skill scores show that performance goes down with increasing LT. However, skill scores are improved compared to the raw outputs. For example, NS increases around 15% for LT12, CV decreases around 10% for the same LT and the median of Pbias is much closer to 0 for all LT.

3.2.2 Performance on spatial scale

Figure 7 displays NS score of the raw (left) and post-processed forecast (right) at LT7. The maps of the other skill scores are given in Appendix B (Fig. B1). The results of the raw forecasts demonstrate performance is better in cells of rivers that exhibit higher discharges and larger catchments

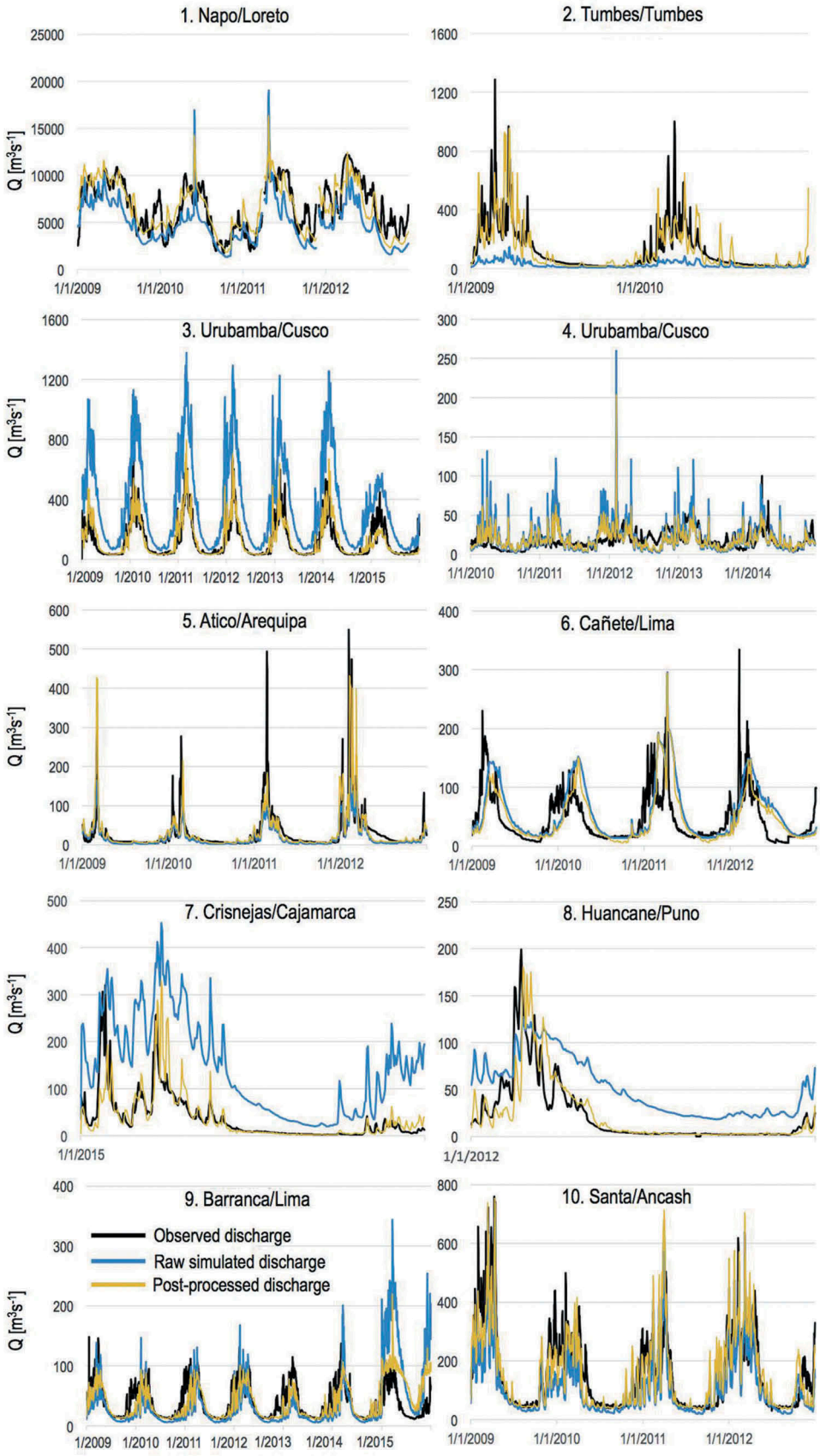


Figure 4. Observed discharge (black line), raw simulated discharge (blue line) and post-processed discharge (yellow line) for the 10 gauging stations of the study area.

Table 1. GloFAS validation results using the raw and the post-processed simulated discharges against observed data for 10 gauging stations.

	Province/Region	Station coordinates		Data availability (years)	Average daily observed discharge (m^3/s)	NS	Post-processed data	POD
		Lat (°)	Lon(°)			Raw data		
1	Napo/Loreto	−3.48	−73.08	4	6602.0	0.12	0.61	0.39
2	Tumbes/Tumbes	−3.71	−80.46	2	117.0	−0.17	0.42	0.47
3	Urubamba/Cusco	−13.18	−72.53	7	126.1	−4.49	0.72	0.63
4	Utcubamba/Amazonas	−5.89	−78.18	5	17.9	−1.89	−0.29	0.27
5	Atico/Arequipa	−17.02	−71.69	4	30.8	0.46	0.41	0.73
6	Cañete/ Lima	−13.02	−76.19	4	46.7	0.00	0.13	0.32
7	Crisnejas/Cajamarca	−7.46	−78.11	1	36.7	−5.38	0.46	0.64
8	Huancane/Puno	−15.12	−69.79	1	22.2	−0.15	0.57	0.68
9	Barranca/Lima	−10.54	−77.22	7	39.4	−0.82	0.35	0.31
10	Santa/Ancash	−8.65	−78.25	4	152.8	0.56	0.65	0.52
	Data availability weighted average					−1.23	0.39	0.46

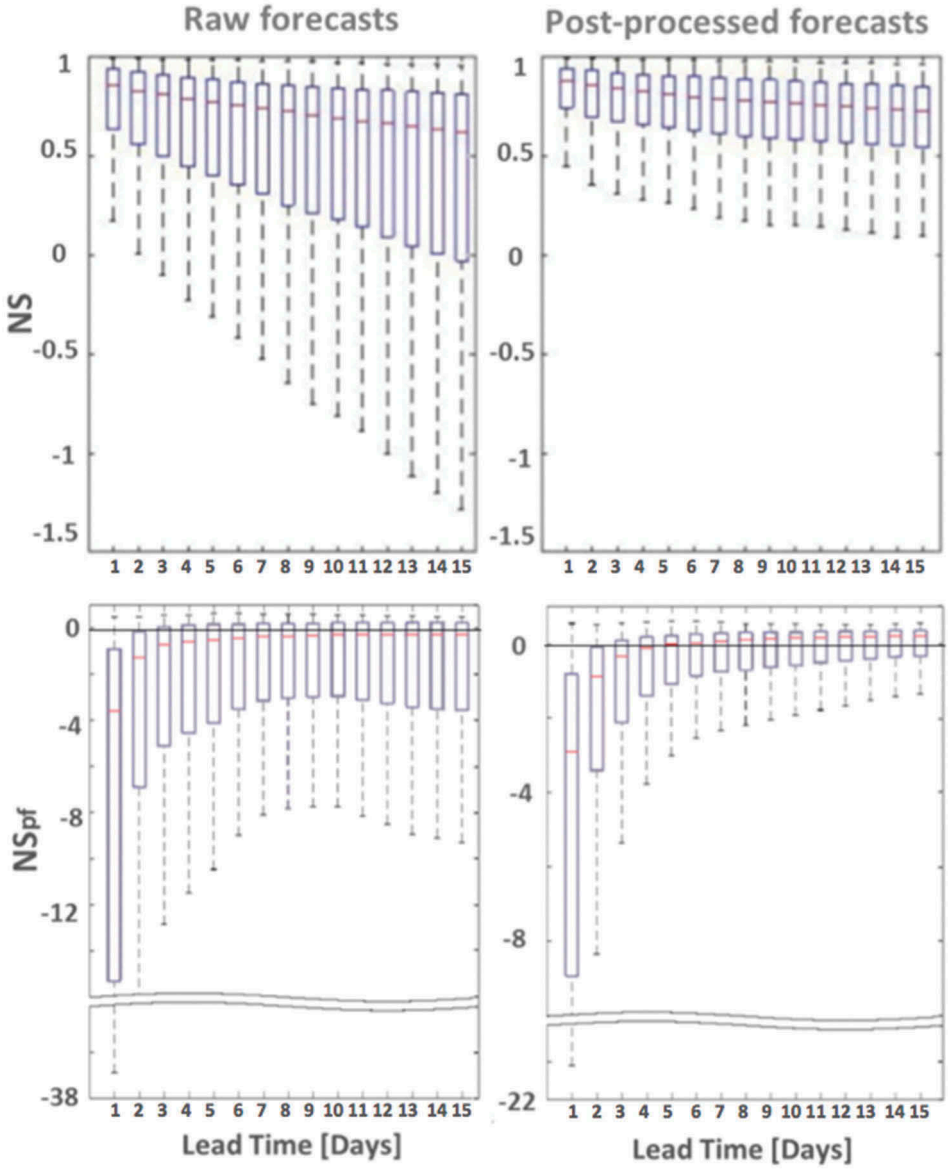


Figure 5. Boxplots of NS and NS_{pf} *versus* the forecast LT over the period 1 January 2009–31 December 2015, calculated for 2780 river points for raw forecasts (left) and post-processed forecasts (right). The horizontal (red) line shows the median, the edges of the boxes (blue) indicate the 25th–75th percentiles and the whiskers the first–99th percentiles.

(i.e. the Amazon, Ucayali and Urubamba), as their NS scores are closer to 1. High skill is shown for river cells of the largest regions (i.e. Loreto and Ucayali). On the other hand, many grid cells along coastal areas, where catchments are small and rivers are shorter (e.g. Tumbes and Piura in the north-west and Ica in the south), have NS values below 0. Similar results are found for CV and CRPSS. The Pbias values show that the meteorological forecasts that are used as an input in the

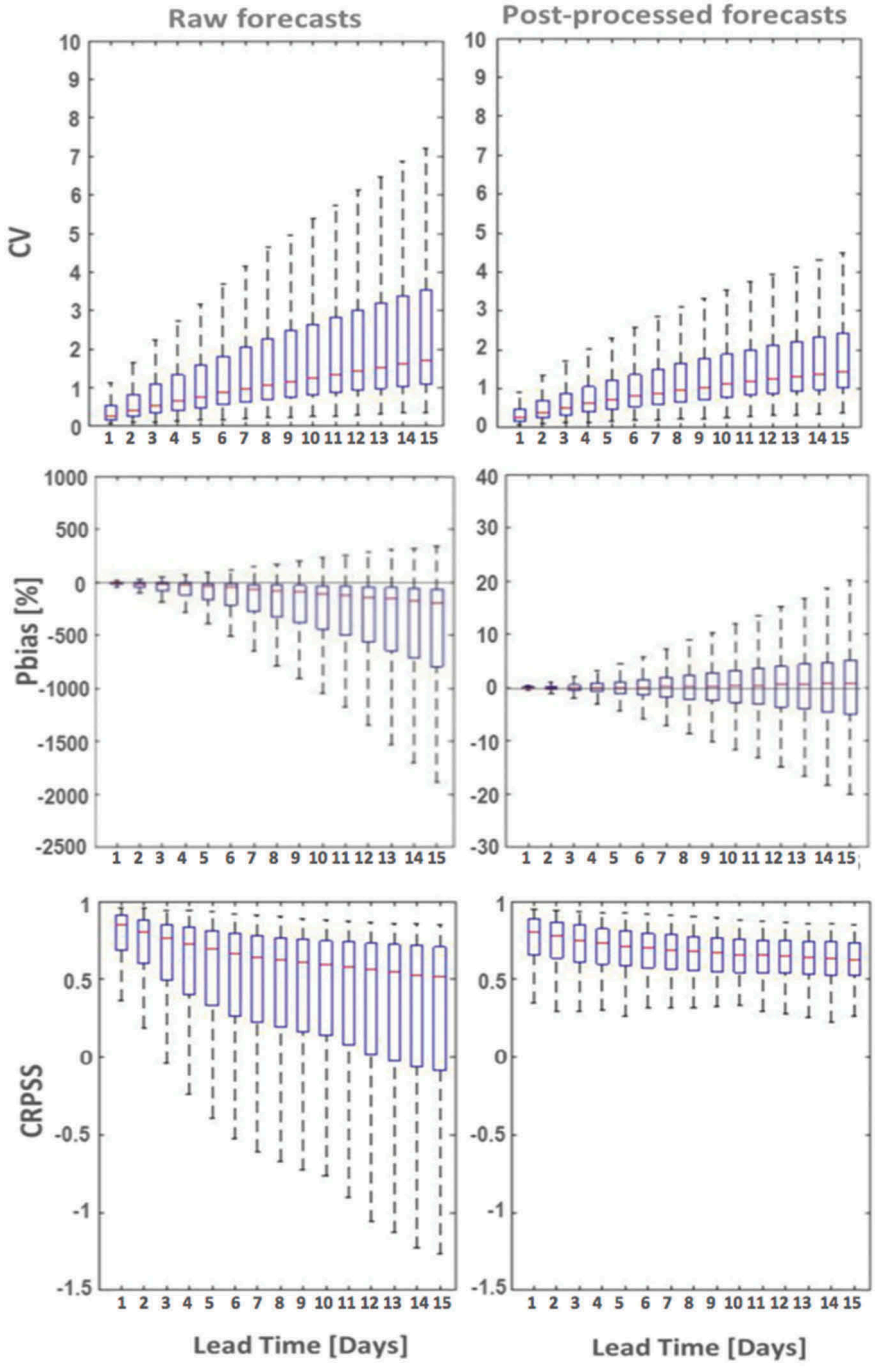


Figure 6. Boxplots of CV, Pbias and CRPSS *versus* the forecast LT over the period 1 January 2009–31 December 2015, calculated for 2780 river points for raw forecasts (left) and post-processed forecasts (right). The red line shows the median, the edges of the blue boxes indicate the 25th–75th percentiles and the whiskers the 1st–99th percentile.

model have a clear tendency to forecast higher precipitation than the one produced at LT0. As a result, the model forecasts higher discharge at longer LTs than the simulated discharge at LT0. This is the case for the entire country, with some exceptions in river cells in the Huancavelica, Cuzco, Loreto and Puno regions.

Post-processing improves the scores for most river cells (Fig. 5, right). More specifically, NS improves largely for the river cells whose raw forecasts exhibit very low scores. The CRPSS becomes better in the river cells that are located along the coast and in the central northern areas shown, while CV scores improve slightly but systematically across the domain. Finally, the significant over-forecasting of the raw forecasts, demonstrated by Pbias, has been corrected considerably, which is to be expected from the application of the quantile mapping method. Values are close to 0 in most areas, except for some coastal river points and the ones in the

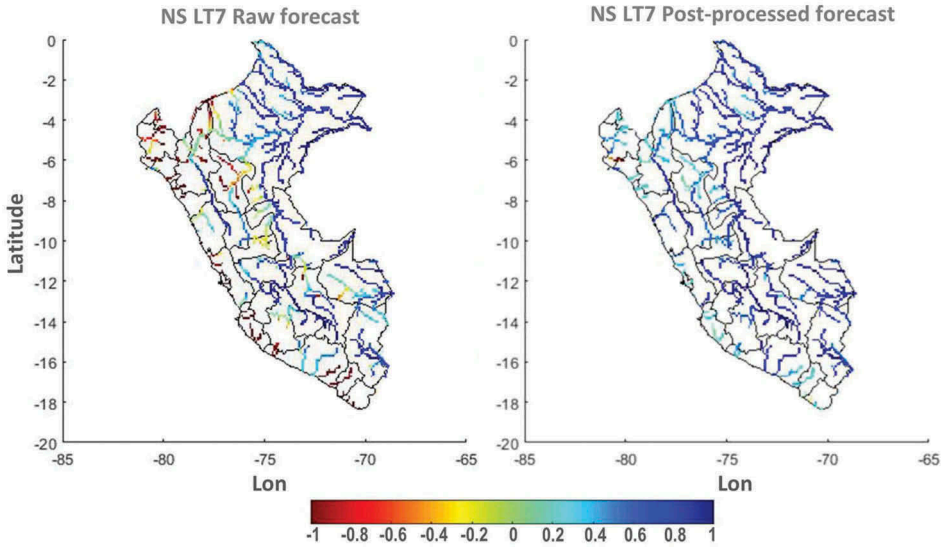


Figure 7. Nash-Sutcliffe coefficient (NS) over Peru for daily forecasts over the period 1 January 2009–31 December 2015 for 7-day LT.

Piura, Tumbes and Lambayeque regions, where the model continues to forecast higher discharges than those produced for LT0.

3.2.3 Performance versus upstream area

Figures 8 and 9 show the scores at LT7 as function of the upstream area size for the raw forecast (left) and for the post-processed forecast (right). Following the approach in Section 2.4.3, river cells were classified into 11 groups, each of which contains an equal fraction of all cells. The results indicate that performance increases and variability decreases with increasing upstream area. For example, at least 95% of river cells that belong to the two highest upstream area groups (i.e. >120 000 km^2) have NS (Fig. 8) and CRPSS (Fig. 9) values above 0 in both the raw and post-processed forecasts. Moreover, for these groups, 95% of the raw forecast river points exhibit a CV of less than 2, while the same percentage in the post-processed forecasts exhibits values of less than 1. However, more than 40% of river cells that belong to the lowest upstream area categories (i.e. up to 4600 km^2) demonstrate CRPSS values below 0 (Fig. 9). A big change between the raw and post-processed forecasts is observed in Pbias: most river cells in all groups exhibit negative values for the raw outputs, while the corrected forecasts display Pbias median values close to 0 in all groups (Fig. 9). In addition, the variability between the river cells has decreased considerably. Finally, the NS_{pf} median value of the raw forecasts is negative for all categories, but it becomes positive after the post-processing for most river cell groups, revealing that, in most cases, the post-processed output should be preferred rather than a persistent forecast at this LT.

3.3 Performance in forecasting observed events

In this section, we evaluate whether the GloFAS hindcasts were able to forecast the reported, damaging flood events. The evaluation uses the reference discharge on onset dates of floods (LT0) and the raw and post-processed forecasts at LT1–15. A forecast flood is defined according to the criteria described in Section 2.4.3. The average POD using the simulated discharge over all geographical areas is 0.62, while the best score is found in northern Peru (0.67). In the central regions, POD is 0.58 and in the southern ones it is 0.61. The average POD of the raw forecasts is 0.82. The fact that this is much higher than that of the simulated discharges illustrates the over-forecasting of discharge by the model. On the one hand, this shows end users would have received a flood signal before most of the reported events; on the other hand, they would have had a very high number of false alarms. The POD of the post-processed forecasts becomes closer to the POD of the simulated discharges (0.65). These results are presented in Table 2.

Finally, Figure 10 illustrates the effect of post-processing on the over-forecasting of discharge. In Figure 10, the forecasts of LT7 are presented for Flood #8 on the map of Peru (Fig. 2). The river cells that are forecast to exceed the 90th, 95th and 99th percentiles of the climatological discharge are highlighted (in yellow, magenta and purple, respectively). It may be observed that, according to the raw forecasts (Fig. 10, left), most regions would be flooded. However, in actuality, the only flood reported was in the northern region of Loreto (black dots). This shows that, although the model has captured the flood in this location, it has also produced several false alarms in other regions, if we assume that there was no under-reporting of events. In contrast, the post-processed forecast captured this flood event fairly well (Fig. 10, right). Whereas it did provide flood signals in the south, where no flood events were reported, these signals were considerably fewer than those of the raw forecast.

The operational flood warning map that was taken from the GloFAS website[5] is presented in Appendix C (Fig. C1). This shows that GloFAS successfully forecast the catastrophic March 2017 flood in Peru that was mentioned in Section 2.1.

4 Discussion and conclusions

This study explored the potential of GloFAS as an operational flood warning system in Peru. The predictive capability was investigated for the entire Peruvian river system, which consists of 2780 river grid cells, using daily operational forecasts

[5]http://www.globalfloods.eu/.

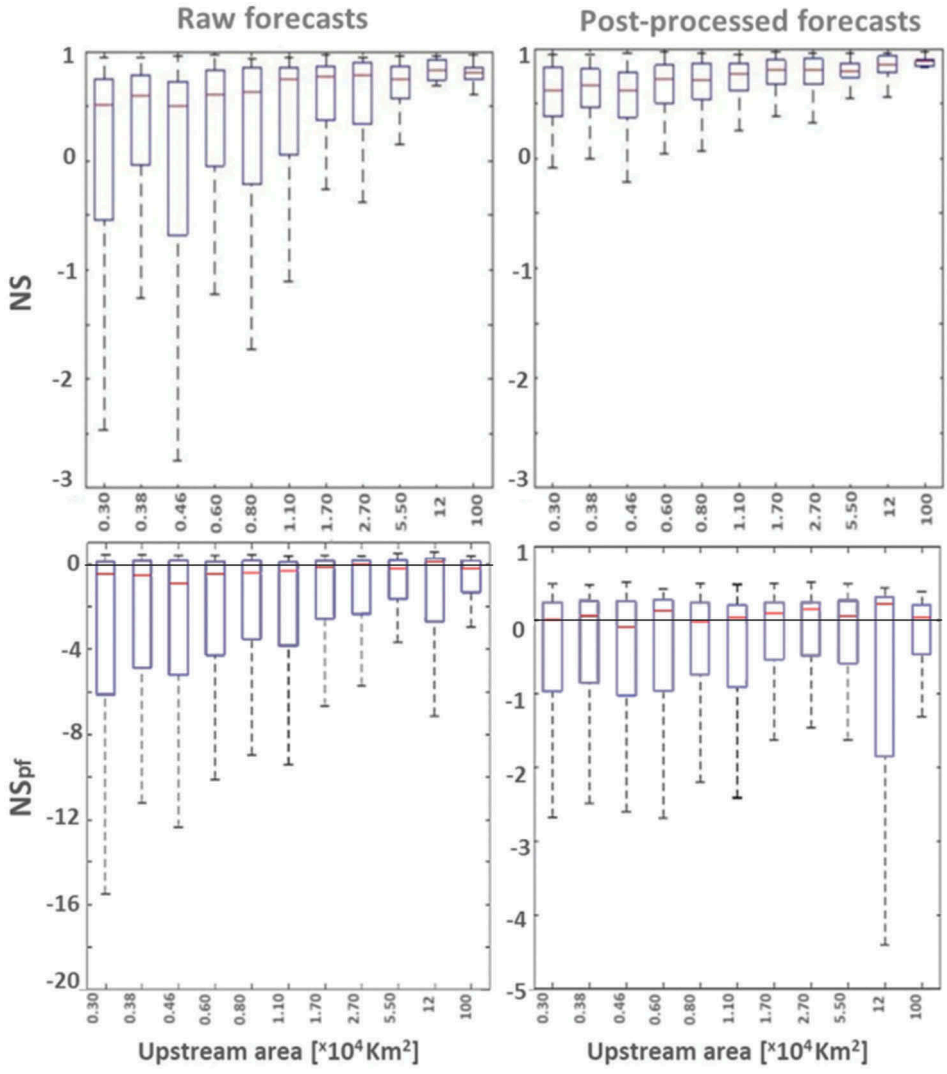


Figure 8. Boxplots of NS and NS_{pf} *versus* the upstream area over the period 1 January 2009–31 December 2015 for raw forecasts (left) and post-processed forecasts (right) at LT 7. River cells are split into 11 groups of 252 river cells each. The horizontal (red) line shows the median, the edges of the boxes (blue) indicate the 25th–75th percentiles and the whiskers the 1st–95th percentiles.

from 2009 to 2015. The study compared the simulated discharges that are produced by the model at zero lead time (LT0) with observed discharges at 10 gauging stations, located in different regions. Furthermore, we compared (a) the forecast discharge of different LTs with the simulated discharge in order to show the ability of the model in predicting, and (b) the flood signals issued by the model with the damaging flood events that were reported in several disaster databases. Next, forecasts were post-processed using the quantile mapping technique to remove bias. The evaluation was repeated and the outcomes were compared to the raw forecasts.

The comparison of the observed with the simulated discharge data indicates that, although GloFAS in general captures the seasonality of the discharge, large quantitative differences are observed between the two. This can be explained by the fact that it is not calibrated in this area. If a larger number of gauging stations had been available, the comparison of observed and simulated discharges could have provided further insights about the model's performance and the nature of errors and biases. However, although the Peruvian National Water Authority counts a network of over 100 gauging stations, the large majority is located close to the coast and, mostly, observed and forecasted discharge time series years do not coincide.

Furthermore, the results show that the performance of the raw forecasts in predicting the simulated discharges decreases with LT and highlights the tendency of the meteorological forecasts that are used as input in the model to forecast higher precipitation compared to the one used to initiate the model. This leads to higher forecast discharges than simulated ones across the entire country for all LTs. In addition, when using the persistence criterion, the results show that, in most cases, it is better to use a persistent forecast than the model itself, especially at short LTs. On average, the verification scores improve and variability decreases for river cells that belong to groups of larger upstream areas. This is an expected tendency, as the water attenuation in the case of the large, east-oriented Peruvian rivers takes place over several days, filtering out the short-term variability of the meteorological forcing. This mainly applies to the river cells of the Amazon and Napo rivers (north-eastern region), the Ucayali River (central region) and the Urubamba River (southern region). In shorter river systems that have a rapid response time, such as the ones along the coastline of Peru, the principal process that

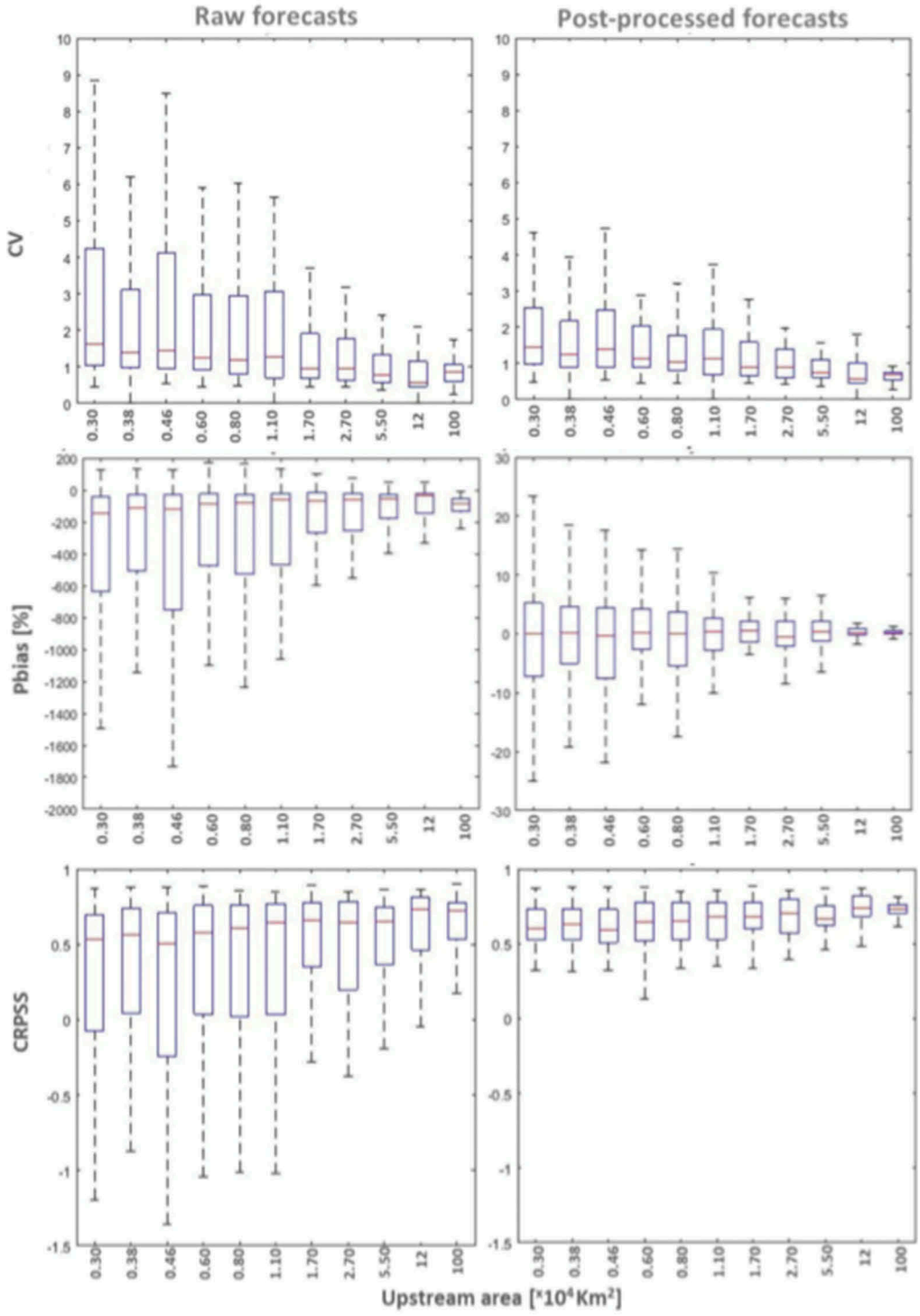


Figure 9. Boxplots of CV, Pbias and CRPSS *versus* the upstream area over the period 1 January 2009–31 December 2015 for raw forecasts (left) and post-processed forecasts (right) at LT 7. See Figure 8 caption for explanation.

controls the performance is the meteorological forcing itself and, therefore, verification scores have considerably lower values. Particularly, in the north-western coastal regions, where GloFAS is being used operationally by the Red Cross, the model has very little skill. This approach is usually followed in data-poor areas, where high-quality, daily, observational discharges are not available. The major advantage of this approach is that it can be applied to each river cell individually, allowing a skill assessment of the model itself on large spatial scales, but it has to be carefully used by the end users, since it mainly demonstrates the skill of the meteorological input and less that of the hydrological model itself. The application of verification scores, which was based on the evaluation framework developed by Alfieri *et al.* (2014), aimed to cover different aspects of forecast attributes and to allow river cells with different discharge magnitudes, upstream areas and climatic regimes to be comparable. Averaging scores over the entire 7-year period for which operational forecasts were available was preferred over using dry/wet seasons as (a) our forecast sample is

Table 2. Semi-qualitative evaluation by-product of detection (POD) of the ability of the GloFAS model to detect reported floods in Peru for zero lead time (LT0) and to forecast them at 1–15 days' LT by raw and post-processed forecasts. *Q*: discharge.

Region	LT0	Forecast (LT 1–15)	
	Simulated *Q*	Raw	Post-processed
North	0.67	0.85	0.67
Centre	0.58	0.82	0.65
South	0.61	0.78	0.65
Total	0.62	0.82	0.65

already relatively small and (b) Peru is characterised by a variety of (micro-)climate areas with different inter-annual variability.

The event-based analysis demonstrates that the majority of reported flood events were correctly forecast (POD = 0.82; the POD shows the proportion of successfully detected events). However, the POD of the simulated discharges is lower (0.62), which may be due to errors in flood reporting or the fact that raw forecasts tend to produce higher discharge than simulated ones. On the one hand, this leads to a good POD, while on the other hand this will lead to several false alarms.

When quantile-mapping is applied to the raw forecasts as a post-processing technique, the skill of the models increases for most river cells, both when comparing the simulated against the observed and the forecast against the simulated discharge. Furthermore, in this case, the median of the NS_{pf} that uses the persistency criterion is positive for LTs longer than 5 days, which shows that the model is a better option than using the persistent forecast. However, the POD for the reported flood events decreases (0.65), becoming closer to that of the simulated discharge, which demonstrates the effect of post-processing in discharge over-forecasting. Hence, a decrease in the number of false alarms is also expected.

Quantifying the number of false alarms is a rather complicated task. The collection and monitoring procedures of disaster loss data are not standardised amongst the different disaster databases. There are numerous discrepancies in the number, type and impact of the disasters (Gall *et al.* 2009, Wirtz and Below 2009). We reduced these uncertainties by combining various datasets used in similar scientific studies (Jongman *et al.* 2014, Thiemig *et al.* 2015, Hoeppe 2016; Bischiniotis *et al.* 2018) with the goal to obtain an as complete as possible and reliable list of flood events with property damage or affected population. However, we cannot claim with certainty that all floods were included, because it is likely that some of them did not cause any damage because of correctly triggered action or because of misses in reporting. Therefore, we only focused on the ones reported, without calculating the overall false alarms.

The calculation of false alarms should be done locally, in specific river points or sections, tailoring the analysis to the local boundary conditions and needs (e.g. streamflow and probability thresholds). For example, in our paper, we have used the 90th percentile of the simulated discharge time series, produced by GloFAS as a threshold to define a flood event. As mentioned before, this low percentile was used to increase our sample size, given the limited available forecast time series in combination with the rare nature of flood events. However, in reality this threshold is highly dependent on the local boundary conditions. For example, at some locations, exceeding a higher percentile may not cause any damage, while in others, very high damage can be caused by a lower percentile exceedence. A site-by-site analysis that uses detailed vulnerability and exposure data will lead to more accurate discharge thresholds, which can lead to the estimation of the false alarms in each location.

The bias correction methods are usually carried out using separate datasets for calibration and evaluation. Application of this separation did not lead to significantly different results and, therefore, we applied the quantile mapping to the entire available dataset to increase the statistical sample of our results. The improvement of skill scores after adjusting the climatological probability distribution to a reference is not trivial, as temporal variability and threshold-dependent indicators lead to a strongly non-linear propagation of forecast characteristics, which does affect skill results for individual events. Further improvement could possibly be achieved using more sophisticated post-processing methods, e.g. Bayesian model

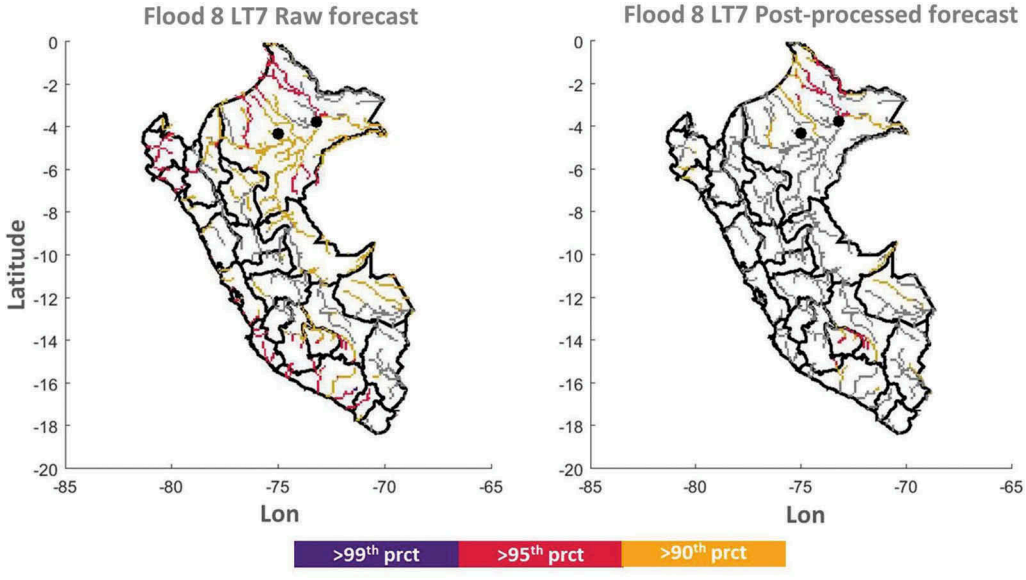


Figure 10. Example of the raw (left) and post-processed (right) forecasts for Flood #8 and for LT7. River cells are forecast to exceed the 90th percentile (yellow), the 95th percentile (magenta) and the 99th percentile (purple) of the climatological discharge. The flooded places as mentioned in the disaster databases are depicted by the black circles.

averaging (Raftery *et al.* 2005), ensemble model output statistics (Gneiting *et al.* 2005), and by conducting the analysis in different hydrological periods focusing on fewer river cells (wet/dry). A combination of pre- and post-processing methods could also be tried as it is likely to lead to greater improvements (Kang *et al.* 2010). It is also important to identify the cause of the LT-dependent systematic bias in the GloFAS calculations in Peru, which was beyond the scope of this study.

Drawing conclusions about whether and how much the model benefits the flood risk management in Peru is quite a complicated task. The actual forecast value is a product of the acceptable trade-offs that have been set in each flood risk strategy, such as those applied in forecast-based financing, developed by the Red Cross/Red Crescent (Coughlan De Perez *et al.* 2015). It is likely that the same forecast is beneficial for one strategy and less useful for another. Skill scores can be influenced to a great extent by the chosen thresholds. The acceptable level of false alarms in relation to correct hits and misses is subject to detailed cost-benefit analyses, which are largely dependent on the local boundary conditions. Such cost-benefit analyses include both tangible and intangible costs and benefits. For instance, when action is taken in vain, users lose confidence in the warnings issued, which can lead to reduced response to future warnings (LeClerc and Joslyn 2015). A careful estimation of the needs of end users is required to pick a robust forecast threshold that leads to the optimal trade-off between the costs of the mitigation measures and the achieved risk reduction. Therefore, the results of this study could be used as an indicator of the model performance for better and more effective flood risk management by humanitarian organisations acting in Peru. Future research is expected to use this evaluation framework for longer lead times in more countries, using the recently released GloFAS seasonal river flow outlook.

Acknowledgements

We thank Munich Re for providing reported flood data from the NatCatSERVICE database for the VIDI Compound Risk project by Philip Ward.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

The project was funded by NWO-VICI [Grant no. 453-13-006] and NWO New Delta [Grant no. 869.15.001];Nederlandse Organisatie voor Wetenschappelijk Onderzoek [NWO New Delta/869.15.001,NWO-VICI /453-13-006].

References

Alfieri, L., *et al.*, 2012. Operational early warning systems for water-related hazards in Europe. *Environmental Science & Policy*, 21, 35–49. doi:10.1016/j.envsci.2012.01.008

Alfieri, L., *et al.*, 2013. GloFAS-global ensemble streamflow forecasting and flood early warning. *Hydrology and Earth System Sciences.*, 17 (3), 1161–1175. doi:10.5194/hess-17-1161-2013

Alfieri, L., *et al.*, 2014. Evaluation of ensemble streamflow predictions in Europe. *Journal of Hydrology*, 517, 913–922. doi:10.1016/j.jhydrol.2014.06.035

Arnell, N.W. and Lloyd-Hughes, B., 2014. The global-scale impacts of climate change on water resources and flooding under new climate and socio-economic scenarios. *Climate Change*, 122 (1–2), 127–140. doi:10.1007/s10584-013-0948-4

Baigorria, G.A., *et al.*, 2007. Assessing uncertainties in crop model simulations using daily bias-corrected regional circulation model outputs. *Climate Research*, 34 (3), 211–222. doi:10.3354/cr00703

Balsamo, G., *et al.*, 2011. Evolution of land surface processes in the IFS. *ECMWF Newsletter*, 127, 17–22.

Balsamo, G., *et al.*, 2015. ERA-Interim/Land: A global land surface reanalysis data set. *Hydrology and Earth System Sciences*, 19 (1), 389–407. doi:10.5194/hess-19-389-2015

Bartholmes, J.C., *et al.*, 2008. The European flood alert system EFAS; part 2: statistical skill assessment of probabilistic and deterministic operational forecasts. *Hydrology and Earth System Sciences Discussions*, 5, 289–322. doi:10.5194/hessd-5-289-2008

Bayer, A.M., *et al.*, 2014. An unforgettable event: A qualitative study of the 1997-98 El Niño in northern Peru. *Disasters*, 38 (2), 351–374. doi:10.1111/disa.12046

Bellier, J., Zin, I., and Bontron, G., 2018. Generating coherent ensemble forecasts after hydrological postprocessing adaptations of ECC-based methods. *Water Resources Research*. doi:10.1029/2018WR022601,2018

Bierkens, M.F.P. and van Den Hurk, B.J.J.M., 2007. Groundwater convergence as a possible mechanism for multi-year persistence in rainfall. *Geophysical Research Letters*, 34 (2). doi:10.1029/2006GL028396

Bischiniotis, K., *et al.*, 2018. The influence of antecedent conditions on flood risk in sub-Saharan Africa. *Natural Hazards and Earth System Sciences*, 18, 271–285. doi:10.5194/nhess-18-271-2018

Boelee, L., *et al.*, 2017. Analysis of the uncertainty in flood predictions of GloFAS for Piura in the Pacific Region of Peru. *European Geosciences Union General Assembly 2017,* Vienna, Austia. doi:10.13140/RG.2.2.17486.46406

Bradley, A.A. and Schwartz, S.S., 2011. Summary verification measures and their interpretation for ensemble forecasts. *Monthly Weather Review*, 139 (9), 3075–3089. doi:10.1175/2010MWR3305.1

Brakenridge, G.R., 2015. Global active archive of large flood events. *Dartmouth Flood Obs. Univ. Color.* http://floodobservatorhttp://floodobservatory.colo [online] Available from: http://floodobservatory.colorado.edu/Archives/index.html

Brown, J.D., *et al.*, 2010. The ensemble verification system (EVS): A software tool for verifying ensemble forecasts of hydrometeorological and hydrologic variables at discrete locations. *Environmental Modelling & Software*, 25 (7), 854–872. doi:10.1016/j.envsoft.2010.01.009

Candogan Yossef, N., *et al.*, 2012. Assessment of the potential forecasting skill of a global hydrological model in reproducing the occurrence of monthly flow extremes. *Hydrology and Earth System Sciences*, 16 (11), 4233–4246. doi:10.5194/hess-16-4233-2012

Candogan Yossef, N., *et al.*, 2017. Skill of a global forecasting system in seasonal ensemble streamflow prediction. *Hydrology and Earth System Sciences*, 21 (8), 4103–4114. doi:10.5194/hess-21-4103-2017

Clark, M.P. and Hay, L.E., 2004. Use of medium-range numerical weather prediction model output to produce forecasts of streamflow. *Journal of Hydrometeorology*, 5 (1), 15–32. doi:10.1175/1525-7541(2004)005<0015:UOMNWP>2.0.CO;2

Cleveland, W.S., 1979. Robust locally weighted regression and smoothing scatterplots. *Journal of Americal Statistical Association*, 74 (368), 829–836. doi:10.1080/01621459.1979.10481038

Cloke, H.L. and Pappenberger, F., 2009. Ensemble flood forecasting: A review. *Journal ofHydrology*, 375 (3–4), 613–626. doi:10.1016/j.jhydrol.2009.06.005

Coughlan de Perez, E., *et al.*, 2016. Action-based flood forecasting for triggering humanitarian action. *Hydrology andEarth System Sciences*, 20 (9), 3549–3560. doi:10.5194/hess-20-3549-2016

Coughlan De Perez, E., *et al.*, 2015. Forecast-based financing: an approach for catalyzing humanitarian action based on extreme weather and climate forecasts. *Natural Hazards and Earth System Sciences*, 15 (4), 895–904. doi:10.5194/nhess-15-895-2015

Dale, M., *et al.*, 2014. Probabilistic flood forecasting and decision-making: an innovative risk-based approach. *Natural Hazards*, 70 (1), 159–172. doi:10.1007/s11069-012-0483-z

De Roo, A.P.J., *et al*., 2003. Development of a european flood forecasting system. *International Journal of River Basin Management*, 1 (1), 49–59. doi:10.1080/15715124.2003.9635192

Dee, D.P., *et al*., 2011. The ERA-Interim reanalysis: configuration and performance of the data assimilation system. *Quarterly Journalof the Royal Meteorological Society*, 137 (656), 553–597. doi:10.1002/qj.828

Déqué, M., 2007. Frequency of precipitation and temperature extremes over France in an anthropogenic scenario: model results and statistical correction according to observed values. *Global and Planetary Change*, 57 (1–2), 16–26. doi:10.1016/j.gloplacha.2006.11.030

Doocy, S., *et al*., 2013. The human impact of earthquakes: a historical review of events 1980-2009 and systematic literature review. *PLoS Currents*, Apr, 2013. doi:10.1371/currents.dis.67bd14fe457f1db0b5433a8ee20fb833.

El Comercio, 2017. COEN: a 145 creció número de muertos por lluvias e inundaciones. 19 May. El Comercio. Available from: http://elcomercio.pe/peru/coen-145-crecio-numero-muertos-lluvias-e-inundaciones-424733 [Accessed 10 January 2019].

Franz, K.J., *et al*., 2003. Verification of national weather service ensemble streamflow predictions for water supply forecasting in the Colorado River Basin. *Journal of Hydrometeorology*, 4 (6), 1105–1118. doi:10.1175/1525-7541(2003)004<1105:VONWSE>2.0.CO;2

Fundel, F. and Zappa, M., 2011. Hydrological ensemble forecasting in mesoscale catchments: sensitivity to initial conditions and value of reforecasts. *Water Resources Research*, 47 (9). doi:10.1029/2010WR009996

Gall, M., Borden, K.A., and Cutter, S.L., 2009. When do losses count?. *Bulletin of the Americal Meteorological Society*, 90 (6), 799–809. doi:10.1175/2008BAMS2721.1

Gneiting, T., *et al*., 2005. Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Monthly Weather Review*, 133, 1098–1118. doi:10.1175/MWR2904.1

Gourley, J.J., *et al*., 2012. Evaluation of tools used for monitoring and forecasting flash floods in the United States. *Weather and Forecasting*, 27 (1), 158–173. doi:10.1175/WAF-D-10-05043.1

Gouweleeuw, B.T., *et al*., 2005. Flood forecasting using medium-range probabilistic weather prediction. *Hydrology and Earth System Sciences*, 9 (4), 365–380. doi:10.5194/hess-9-365-2005

Guha-Sapir, D., Vos, F., and Below, R., 2012. Annual disaster statistical review 2011. *Disasters*, 52 [online]. Available from: http://cred.be/sites/default/files/2012.07.05.ADSR_2011.pdf

Guha-Sapir, D., Vos, F., and Below, R., 2014. EM-DAT: international disaster database. *Disasters*, 52 [online]. Available from: http://www.emdat.be/disaster_trends/index.html

Gupta, H.V., Sorooshian, S., and Yapo, P.O., 1999. Status of automatic calibration for hydrologic models: comparison with multilevel expert calibration. *Journal of Hydrologic Engineering*, 4 (2), 135–143. doi:10.1061/(ASCE)1084-0699(1999)4:2(135)

Hamill, T.M., Hagedorn, R., and Whitaker, J.S., 2008. Probabilistic forecast calibration using ECMWF and GFS ensemble reforecasts. part II: precipitation. *Monthly Weather Review*, 136 (7), 2620–2632. doi:10.1175/2007MWR2411.1

Hannah, D.M., *et al*., 2011. Large-scale river flow archives: importance, current status and future needs. *Hydrological Processes*, 25 (7), 1191–1200. doi:10.1002/hyp.7794

Hashino, T., Bradley, A.A., and Schwartz, S.S., 2007. Evaluation of bias-correction methods for ensemble streamflow volume forecasts. *Hydrology and Earth System Sciences*, 11, 939–950. doi:10.5194/hess-11-939-2007

Hersbach, H., 2000. Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting*, 15 (5), 559–570. doi:10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2

Hirabayashi, Y., *et al*., 2013. Global flood risk under climate change. *Nature Climate Change*, 3 (9), 816–821. doi:10.1038/nclimate1911

Hoeppe, P., 2016. Trends in weather related disasters - Consequences for insurers and society. *Weather and Climate Extremes*, 11, 70–79. doi:10.1016/j.wace.2015.10.002

IPCC, 2012. Managing the risks of extreme events and disasters to advance climatechange adaptation. *In*: C.B. Field *et al*., eds. *A special report of working groups i and ii of the intergovernmental panel on climate change*. Cambridge, UK, and New York, NY: Cambridge University Press, 582.

Jongman, B., *et al*., 2014. Increasing stress on disaster-risk finance due to large floods. *Nature Climate Change*, 4 (4), 264–268. doi:10.1038/nclimate2124

Kang, T.H., Kim, Y.O., and Hong, I.P., 2010. Comparison of pre- and post-processors for ensemble streamflow prediction. *Atmospheric Science Letters*, 11 (2), 153–159. doi:10.1002/asl.276

Khalil, A.F., *et al*., 2007. El Niño-Southern Oscillation-based index insurance for floods: statistical risk analyses and application to Peru. *Water Resources Research*, 43 (10). doi:10.1029/2006WR005281

Laio, F. and Tamea, S., 2006. Verification tools for probabilistic forecasts of continuous hydrological variables. *Hydrology and Earth System Science Discussions*, 3 (4), 2145–2173. doi:10.5194/hessd-3-2145-2006

LeClerc, J. and Joslyn, S., 2015. The cry wolf effect and weather-related decision making. *Risk Analysis*, 35 (3), 385–395. doi:10.1111/risa.12336,

Madadgar, S., Moradkhani, H., and Garen, D., 2014. Towards improved post-processing of hydrologic forecast ensembles. *Hydrological Processes*, 28 (1), 104–122. doi:10.1002/hyp.9562

McBride, J.L. and Ebert, E.E., 2000. Verification of quantitative precipitation forecasts from operational numerical weather prediction models over Australia. *Weather and Forecasting*, 15 (1), 103–121. doi:10.1175/1520-0434(2000)015<0103:voqpff>2.0.co;2

Miller, M., *et al*., 2010. Increased resolution in the ECMWF deterministic and ensemble prediction systems. *ECMWF Newsletter*, 124, 10–16.

Moriasi, D.N., *et al*., 2007. Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. *Transactions of the ASABE*, 50 (3), 885–900. doi:10.13031/2013.23153

Munich Re. NatCatSERVICE | munich Re, 2015 [online] Available from: http://www.munichre.com/natcatservice.

Nash, J.E. and Sutcliffe, J.V., 1970. River flow forecasting through conceptual models part I - A discussion of principles. *Journal of Hydrology*, 10 (3), 282–290. doi:10.1016/0022-1694(70)90255-6

Palmer, T.N., 2001. A nonlinear dynamical perspective on model error: A proposal for non-local stochastic-dynamic parametrization in weather and climate prediction models. *Quarterly Journal of the Royal Meteorological Society*, 127 (572), 279–304. doi:10.1002/qj.49712757202

Pappenberger, F., *et al*., 2005. Cascading model uncertainty from medium range weather forecasts (10 days) through a rainfall-runoff model to flood inundation predictions within the European Flood Forecasting System (EFFS). *Hydrology Earth System Sciences*, 9 (4), 381–393. doi:10.5194/hess-9-381-2005

Peel, M.C., Finlayson, B.L., and McMahon, T.A., 2007. Updated world map of the Köppen-Geiger climate classification. *Hydrology and Earth System Sciences and Discussion*, 4 (2), 439–473. doi:10.5194/hessd-4-439-2007

Plate, E.J. and Lindenmaier, F., 2008. Quality assessment of forecasts. Mekong river commission: sixth annual flood forum. *Phnom Penh*, May, 10.

Raftery, A.E., *et al*., 2005. Using Bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review*, 133, 1155–1174. doi:10.1175/MWR2906.1

Raiffa, H. and Schlaifer, R., 1961. *Applied statistical decision theory*. Boston, MA: Division of Research, Harvard Business School.

Randrianasolo, A., *et al*., 2010. Comparing the scores of hydrological ensemble forecasts issued by two different hydrological models. *Atmospheric Science Letters*, 11 (2), 100–107. doi:10.1002/asl.259

Reed, S., Schaake, J., and Zhang, Z., 2007. A distributed hydrologic model and threshold frequency-based method for flash flood forecasting at ungauged locations. *Journal of Hydrology*, 337 (3–4), 402–420. doi:10.1016/j.jhydrol.2007.02.015

Roulin, E., 2006. Skill and relative economic value of medium-range hydrological ensemble predictions. *Hydrology and Earth System Sciences Discussions*, 3 (4), 1369–1406. doi:10.5194/hessd-3-1369-2006

Roulin, E. and Vannitsem, S., 2005. Skill of medium-range hydrological ensemble predictions. *Journal of Hydrometeorology*, 6 (5), 729–744. doi:10.1175/JHM436.1

Roulin, E. and Vannitsem, S., 2015. Post-processing of medium-range probabilistic hydrological forecasting: impact of forcing, initial conditions and model errors. *Hydrological Processes*, 29 (6), 1434–1449. doi:10.1002/hyp.10259

Schefzik, R., Thorarinsdottir, T.L., and Gneiting, T., 2013. Uncertainty quantification in complex simulation models using ensemble copula coupling. *Statistical Science*, 28 (4), 616–640. doi:10.1214/12-STS443

Servat, E. and Dezetter, A., 1991. Selection of calibration objective functions in the context of rainfall-runoff modeling in a Sudanese Savannah area. *Hydrological Sciences Journal-Journal Des Sciences Hydrologiques*, 36 (4), 307–330. doi:10.1080/02626669109492517

Shukla, S. and Lettenmaier, D.P., 2011. Seasonal hydrologic prediction in the United States: understanding the role of initial hydrologic conditions and seasonal climate forecast skill. *Hydrology and Earth System Sciences*, 15 (11), 3529–3538. doi:10.5194/hess-15-3529-2011

Smith, J.A., Day, G.N., and Kane, M.D., 1992. Nonparametric framework for long-range streamflow forecasting. *Journal of Water Resources Planning and Management*, 118 (1), 82–92. doi:10.1061/(ASCE)0733-9496(1992)118:1(82)

Sperna Weiland, F.C., *et al.*, 2010. The ability of a GCM-forced hydrological model to reproduce global discharge variability. *Hydrology and Earth System Sciences*, 14 (8), 1595–1621. doi:10.5194/hess-14-1595-2010

Tanoue, M., Hirabayashi, Y., and Ikeuchi, H., 2016. Global-scale river flood vulnerability in the last 50 years. *Scientific Reports*, 6 (1), 36021. doi:10.1038/srep36021,

Thielen, J., *et al.*, 2009. The European flood alert system – part 1: concept and development. *Hydrology and Earth System Sciences*, 13 (2), 125–140. doi:10.5194/hess-13-125-2009

Thiemig, V., *et al.*, 2010. Ensemble flood forecasting in Africa: A feasibility study in the Juba-Shabelle river basin. *Atmospheric Science Letters*, 11 (2), 123–131. doi:10.1002/asl.266

Thiemig, V., *et al.*, 2015. A pan-African medium-range ensemble flood forecast system. *Hydrology and Earth System Sciences*, 19 (8), 3365–3385. doi:10.5194/hess-19-3365-2015

Todini, E., 2007. Hydrological catchment modelling: past, present and future. *Hydrology and Earth System Sciences*, 11 (1), 468–482. doi:10.5194/hess-11-468-2007

Van Der Knijff, J.M., Younis, J., and De Roo, A.P.J., 2010. LISFLOOD: a GIS-based distributed model for river basin scale water balance and flood simulation. *International Journal of Geographical Information Science*, 24 (2), 189–212. doi:10.1080/13658810802549154

Verbunt, M., *et al.*, 2007. Probabilistic flood forecasting with a limited-area ensemble prediction system: selected case studies. *Journal of Hydrometeorology*, 8 (4), 897–909. doi:10.1175/JHM594.1

Wallemacq, P., *et al.*, 2015. The human cost of weather related disasters: 1995-2015. doi:10.13140/rg.2.2.17677.33769

Wang, J., *et al.*, 2011. The coupled routing and excess storage (CREST) distributed hydrological model. *Hydrological Sciences Journal*, 56 (1), 84–98. doi:10.1080/02626667.2010.543087

Webster, P.J., Toma, V.E., and Kim, H.M., 2011. Were the 2010 Pakistan floods predictable? *Geophysical Research Letters*, 38 (4). doi:10.1029/2010GL046346

Wetterhall, F., *et al.*, 2013. HESS Opinions "forecaster priorities for improving probabilistic flood forecasts,". *Hydrology and Earth System Sciences*, 17 (11), 4389–4399. doi:10.5194/hess-17-4389-2013

Wirtz, A. and Below, R., 2009. Working paper disaster category classification and peril terminology for operational purposes, Technical Report, CRED, 1–20 [online]. Available from: cred.be/sites/default/files/DisCatClass_264.pdf.

Wood, A.W. and Lettenmaier, D.P., 2006. A test bed for new seasonal hydrologic forecasting approaches in the western United States. *Bulletin of the American Meteorological Society*, 87 (12), 1699–1712. doi:10.1175/BAMS-87-12-1699

Appendix A

Nash-Sutcliffe efficiency

Comparison between observed and simulated discharge

This NS score is defined as 1 minus the squared difference between the proxy and the observed discharge normalised by the variance of the observed discharge values during the period under investigation:

$$\mathrm{NS} = 1 - \frac{\sum_{t=1}^{N} [q_{\mathrm{obs}}(t) - q_{\mathrm{sim}}(t)]^2}{\sum_{t=1}^{N} [q_{\mathrm{obs}}(t) - \overline{q_{\mathrm{obs}}}(t)]^2} \quad \text{(A1)}$$

where t is the time index of the evaluation window, N is the number of the forecasts issued during that period, q_{sim} is the proxy discharge, q_{obs} is the observed discharge at the same time step and $\overline{q_{\mathrm{obs}}}$ is the average of the observed discharges for the entire time window.

Comparison between simulated and forecast discharge using the mean value of the simulated forecasts

This NS score is defined as 1 minus the squared difference between the proxy and the forecast discharge normalised by the variance of the proxy values during the period under investigation:

$$\mathrm{NS} = 1 - \frac{\sum_{t=1}^{N} [q_{\mathrm{sim}}(t) - q_{\mathrm{fc}}(t)]^2}{\sum_{t=1}^{N} [q_{\mathrm{sim}}(t) - \overline{q_{\mathrm{sim}}}(t)]^2} \quad \text{(A2)}$$

where q_{fc} is the forecast discharge at time step t (mean of the 51-member ensemble) and $\overline{q_{\mathrm{sim}}}$ is the average discharge for the entire time window.

The range of NS lies between $-\infty$ and 1 (perfect fit). Negative values indicate that the mean value of the proxy discharge time series is a better predictor than the model. The score indicates how well the plot of observed *versus* forecast values fits the 1:1 line (Moriasi *et al.* 2007) and it has been found to be the best objective function for reflecting the overall fit of a hydrograph (Servat and Dezetter 1991).

Comparison between simulated and forecast discharge using a persistent forecast

The $\mathrm{NS_{pf}}$ score is suggested by Plate and Lindenmaier (2008) and uses a persistent forecast a reference value:

$$\mathrm{NS_{pf}}(\mathrm{LT}) = 1 - \frac{\sum_{t=1}^{N} [q_{\mathrm{sim}}(t) - q_{\mathrm{fc}}(t)]^2}{\sum_{t=1}^{N} [q_{\mathrm{sim}}(t) - q_{\mathrm{sim}}(t - LT)]^2} \quad \text{(A3)}$$

where $q_{\mathrm{sim}}(t - \mathrm{LT})$ is the value that was used to initialise the model. The range of NS lies between $-\infty$ and 1 (perfect fit). Negative values indicate that the mean value of the proxy discharge time series (or the use of persistent forecast) is a better predictor than the model. The score indicates how well the plot of observed *versus* the simulated/forecasted values fits the 1:1 line (Moriasi *et al.* 2007) and it has been found to be the best objective function for reflecting the overall fit of a hydrograph (Servat and Dezetter 1991).

Percentage bias (Pbias)

This score measures the average tendency of the forecasted values to be smaller or larger than the observed ones and it has the ability to indicate poor model performance (Gupta *et al.* 1999). Percentage bias is a dimensionless measure that measures the forecast bias of N forecasts at an evaluation window of t days, which is rescaled by the corresponding average discharge for the same period and is defined as:

$$\mathrm{Pbias} = \frac{\frac{1}{N}\sum_{t=1}^{N} [q_{\mathrm{sim}}(t) - q_{\mathrm{fc}}(t)]}{\overline{q_{\mathrm{sim}}}} \quad \text{(A4)}$$

Coefficient of variation of the root mean squared error (CV)

The root mean squared error (RMSE) is used to measure the standard deviation between simulated and forecast values. Following Reed *et al.* (2007), it is normalised by the average simulated discharge to allow a comparison between river cells with very different discharges. The result is the so-called coefficient of variation of the RMSE, given by:

$$\mathrm{CV} = \frac{\sqrt{\frac{\sum_{t=1}^{N}[q_{\mathrm{sim}}(t) - q_{\mathrm{fc}}(t)]^2}{N}}}{\overline{q_{\mathrm{sim}}(t)}} \quad \text{(A5)}$$

Continuous ranked probability skill score (CRPSS)

To evaluate the probabilistic skill of GloFAS that is produced by the ensemble members, the CRPSS (Hersbach 2000) is used, as it measures the weighted average skill over threshold values (Bradley and Schwartz 2011). The CRPSS is calculated by normalising the continuous ranked probability score with the climatology, so that it does not depend on the magnitude of discharge and allows for spatial comparisons. It ranges from $-\infty$ to 1 (perfect forecasts) and is defined as:

$$\mathrm{CRPSS} = \frac{\overline{\mathrm{CRPS_{ref}}} - \overline{\mathrm{CRPS_{fcst}}}}{\overline{\mathrm{CRPS_{ref}}}} \quad \text{(A6)}$$

where

$$\mathrm{CRPS} = \int_{-\infty}^{+\infty} [\mathrm{F}(y) - \mathrm{F_0}(y)]^2 \mathrm{d}y \quad \text{(A7)}$$

and $\mathrm{F}(y)$ is the stepwise cumulative distribution function of the ESP of each considered forecast. $\mathrm{F_0}(y) = 0$ when $y <$ observed value; $\mathrm{F_0}(y) = 1$ when $y \geq$ observed value.

$$\mathrm{CRPS_{ref}} = \frac{1}{\mathrm{N}} \sum_{1}^{\mathrm{N}} \left| \mathrm{q_{sim}}(\mathrm{t}) - \overline{\mathrm{q_{sim}}} \right| \quad \text{(A8)}$$

Appendix B

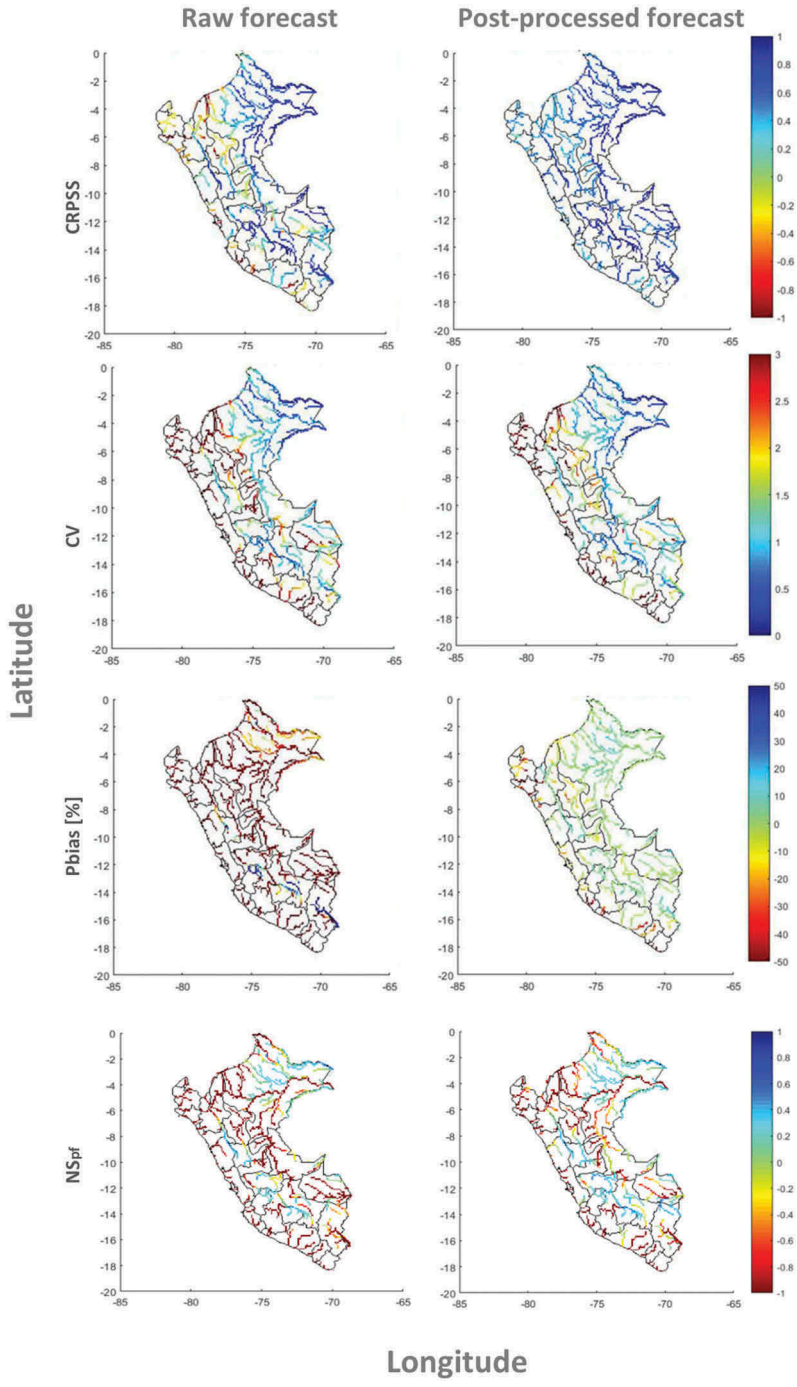


Figure B1. CRPSS, CV, Pbias and NS_{pf} for raw (left) and post-processed forecasts (right) over Peru for daily forecasts over the period 1 January 2009–31 December 2015 for 7-day LT.

Appendix C

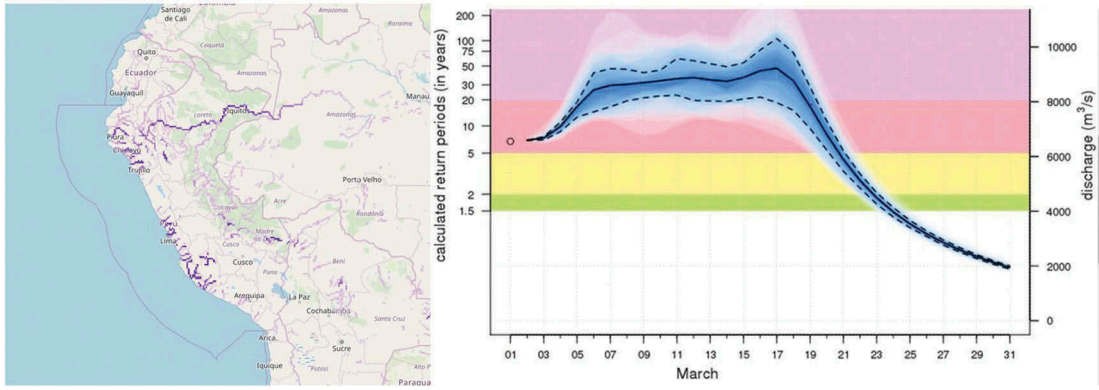


Figure C1. Operational GloFAS warning map on 1 March 2017 (left), showing that the next 15 days the discharges will exceed the 20-year return period in several parts of the Peruvian river network (purple lines) and a more detailed forecast (right) for the station Yuracyacu in Loreto region (77.55°W, 4.45°S). The snapshots were taken from the GloFAS web platform, http://www.globalfloods.eu/.