



# Verification of two hydrological models for real-time flood forecasting in the Hindu Kush Himalaya (HKH) region

Karma Tsering<sup>1</sup> · Manish Shrestha<sup>1</sup>  · Kiran Shakya<sup>1</sup> · Birendra Bajracharya<sup>1</sup> · Mir Matin<sup>1</sup> · Jorge Luis Sanchez Lozano<sup>2</sup> · Jim Nelson<sup>2</sup> · Tandin Wangchuk<sup>3</sup> · Binod Parajuli<sup>4</sup> · Md Arifuzzaman Bhuyan<sup>5</sup>

Received: 7 February 2021 / Accepted: 26 August 2021  
© The Author(s) 2021

## Abstract

The Hindu Kush Himalayan region is extremely susceptible to periodic monsoon floods. Early warning systems with the ability to predict floods in advance can benefit tens of millions of people living in the region. Two web-based flood forecasting tools (ECMWF-SPT and HIWAT-SPT) are therefore developed and deployed jointly by SERVIR-HKH and NASA-AST to provide early warning to Bangladesh, Bhutan, and Nepal. ECMWF-SPT provides ensemble forecast up to 15-day lead time, whereas HIWAT-SPT provides deterministic forecast up to 3-day lead time covering almost 100% of the rivers. Hydrological models in conjunction with forecast validation contribute not only to advancing the processes of a forecasting system, but also objectively assess the joint distribution of forecasts and observations in quantifying forecast accuracy. The validation of forecast products has emerged as a priority need to evaluate the worth of the predictive information in terms of quality and consistency. This paper describes the effort made in developing the hydrological forecast systems, the current state of the flood forecast services, and the performance of the forecast evaluation. Both tools are validated using a selection of appropriate metrics in measurement in both probabilistic and deterministic space. The numerical metrics are further complemented by graphical representations of scores and probabilities. It was found that the models had a good performance in capturing high flood events. The evaluation across multiple locations indicates that the model performance and forecast goodness are variable on spatiotemporal scale. The resulting information is used to support good decision-making in risk and resource management.

**Keywords** Early warning system · GLoFAS-SPT · Forecast validation · Hindu Kush Himalayan · HIWAT-SPT

---

✉ Manish Shrestha  
Shrestha.manish@icimod.org

<sup>1</sup> International Centre for Integrated Mountain Development, Khumaltar, Lalitpur 44700, Nepal

<sup>2</sup> Civil and Environmental Engineering, Brigham Young University, Provo, UT, USA

<sup>3</sup> Hydrology Division, National Center for Hydrology and Meteorology, Thimphu, Bhutan

<sup>4</sup> Provincial Hydrology, Department of Hydrology and Meteorology, Kathmandu, Nepal

<sup>5</sup> Flood Forecasting and Warning Center, BWDB, Dacca, Bangladesh

## 1 Introduction

The Hindu Kush Himalayan (HKH) region is extremely vulnerable to floods caused by climate extremes, often leading to flooding disasters given the region's orogeny, rugged terrain, and monsoon-dominated atmospheric conditions (Khandu et al. 2016; Sikder and Hossian 2018). Consequent disaster impacts are often exacerbated by inefficient risk communication and poor user uptake, which are generally identified as the weakest link in the early warning system (EWS) services. With climate change-induced extreme rainfall events on the rise, flood risk communication through web-based EWS applications needs to be improved, strengthened, and made freely accessible to all (CEGIS 2006; Mirza 2011; Priya et al. 2017).

Early warning systems with the ability to predict floods days in advance can benefit tens of millions of people in the HKH region. However, countries in the HKH region face two major challenges in flood risk communication—the investment required to gauge flood-prone medium and small rivers, and the short lead time provided by instrumental flood EWSs (FEWS) (Basher 2006). Technical flood EWSs can also be rendered ineffective by stakeholders' inadequate readiness. Moreover, the warning is limited to major rivers, excluding many flood-prone rivers and tributaries. The need to enhance FEWS tailored to countries in the HKH region, therefore, is paramount to sustainable water resource management and reducing risks from flooding hazards. Rather than responding to an extreme water-related event, being able to anticipate such a situation before the actual occurrence, and taking proactive actions to alleviate consequences is the basis for every risk reduction and resilience strategy.

The commonly used prediction system for flood early warning guidance in the region is the Global Flood Awareness System (GloFAS) developed by European Commission's Joint Research Council and European Centre For Medium Range Forecasts (EC-JRC-ECMWF) Copernicus Emergency Management System (EC-CEMS 2020). However, the system is currently capable of only producing predictions at coarse spatial resolutions that do not capture hydrological responses over complex mountainous terrain. Unless it is downscaled to capture local details, GloFAS outputs will not be meaningful for local decision-making where flood risks are the greatest. Advances in computing infrastructure, hydroinformatics, and communication technologies together with hydrologic models have greatly enhanced the capability in streamflow prediction (Souffront Alcantara et al. 2019). Without these tools, technically less capable nations of the HKH region may not have access to cutting-edge technologies to address their hydroinformatic challenges. And potentially transform a country's forecasting ability to predict floods days to weeks in advance while reducing associated uncertainties.

The SERVIR ("to serve" in Spanish)-HKH program based at International Centre for Integrated Mountain Development (ICIMOD, Nepal), in partnership with National Aeronautics and Space Administration, USA-Applied Science Team (NASA-AST), has developed two web-based real-time flood forecasting tools for Bangladesh, Bhutan, and Nepal. They are branded as ECMWF-forced, GloFAS-based Streamflow Prediction Tool (GLOFAS-SPT), and the High Impact Weather Assessment Tool Streamflow Prediction Tool (HIWAT-SPT). GLoFAS-SPT tool is used primarily for larger rivers, whereas the HIWAT-SPT is used for small streams which are flashy in nature. The tools are expected to support the national public services agencies by providing reliable forecasts and early warnings for preparedness and resilience against flood. By operationalizing these decision support systems, they can concentrate on managing flood risks without being distracted by the tedious

and costly routines of collecting and processing data, setting up and running models, forecast production, and dissemination. The tools have undergone several iterations in response to user comments and feedback on the interface and enhanced functionalities. These tools can help the disaster management organization to deliver localized and timely information on water resource situations and anticipate flooding risks well in advance to take appropriate actions. Scarce resources can then be channeled into fulfilling the crucial aspects of information dissemination and service delivery.

Forecast verification is critical to justify the operational investment in sustaining flood forecast and warning services and increase socioeconomic value to the end users by informing better decision-making. Over the last several decades, substantial progress has been achieved in probabilistic hydrological forecasting theories, applications, and validation techniques (Bartholmes et al. 2009; Bogner et al. 2011; Cloke et al. 2009; Pappenberger et al. 2015; Verkade et al. 2013). However, significant challenges still exist in assessing the uncertainty of complex physical processes and improving flow and flood predictions, especially extreme hydrological events. The scientific community in weather prediction has developed ensemble methods in quantifying the uncertainty of predictions expressed through probabilistic metrics (Bauer et al. 2015). Forecast goodness defined by consistency, quality, and value (Murphy 1993) is assessed in terms of forecast attributes like accuracy, skill, reliability, resolution, and sharpness based on the probability distribution of forecasts.

The main objective of this paper is to demonstrate the predictive ability of these tools. Observed discharge is used to verify the model's forecasts and quantify uncertainties. Both GLoFAS-SPT and the HIWAT-SPT models have been extensively evaluated at several observational points across Bangladesh, Bhutan, and using historical river hydrology data, flood inundation scenarios, and impact information. Visual, commonly used error statistics, and skill scores are used in this study to evaluate the performance of the models. These error metrics are grouped into classes of (1) probabilistic verification, (2) deterministic verification, and (3) categorical verification. Multiple evaluation methods are implemented to make the study comprehensive, which helps in avoiding methodological biases (Murphy 1991, 1993; Mullen and Buizza 2001). Good decisions are made from good information, and verification results should lead to evidence-based findings based on science-informed inquiries into the model predictions as a credible information service tool.

## 2 Models

GLoFAS-SPT model generates hourly streamflow forecasts out to 15 days in advance using original GloFAS direct runoff routed with the Routing Application for Parallel Computation of Discharge (RAPID) routing model (David et al. 2011). RAPID is a vector–matrix implementation of the Muskingum routing method (David et al. 2011) to simulate discharge at a higher resolution on locally significant river reaches (Snow 2015; Snow et al. 2016). HIWAT-SPT modeled predictions are estimated by routing HIWAT precipitation forecasts through the same RAPID model but using a higher density stream network. HIWAT is a severe convection-allowing weather forecasting system to predict extreme weather phenomena spawn by localized convective disturbances and provides precipitation information for the next 3 days (48–54 h effective lead time) in advance (Gatlin 2017). It is based on Weather Research and Forecasting Model (WRF model) with a 12-km outer and 4-km nested domain positioned over South Asia.

The GLoFAS-SPT tool is a 51-member ensemble system providing information at hourly time step throughout the year, whereas the HIWAT-SPT model is a deterministic configuration providing hourly information and is active only during the monsoon season from May to September. While models are executed at the high-performance computing resources outside of the HKH region, the outputs are delivered to the ICIMOD server on daily basis for generating forecasts. The outputs from these tools are accessed and visualized using interactive web applications created within Tethys development and hosting platform (Swain et al. 2016). The user interface has been customized based on the requirements of the partner and has undergone several iterations through series of dialogs, consultative engagement, and capacity building activities. The tools are freely accessible at <http://tethys.icimod.org/apps/>. Users can view the forecast chart for any mapped river section along with the return period showing the possibility of high discharge at a particular time and day. Streamflow forecasts at reach level are updated on daily basis and color-coded based on the threshold exceedance probabilities defined by return periods of daily flow magnitude. The return periods are obtained from the frequency analysis of simulated historical discharge time series generated from ERA-Interim (Snow et al. 2016). In addition to forecast, the tools also provide an option to download historical discharge of the selected river section based on the ERA-Interim data. More detailed information on the models is available from earlier articles published elsewhere (Souffront Alcantara et al. 2019).

### 3 Data, materials, and methods

#### 3.1 Study area

The study sites span three countries of the HKH region—Bangladesh, Bhutan, and Nepal. The basins from these countries are dominated by hydrologic regimes and flow conditions that result in either riverine flooding or pluvial flash floods annually. The rivers in Bhutan and Nepal originate and flow through high mountains, whereas rivers in Bangladesh flow in the plains. Out of the several sites identified for validation initially, 34 were used in this study (Fig. 1). The stations were selected depending on the length, quality, and usability of the observed discharge. In Bangladesh, attempts were made to compile observed datasets from as many hydrometric stations as possible along the north and northeast boundaries with headwater states of India. While Bhutan and Nepal are dominated by montane-monsoonal climate, Bangladesh generally experiences a maritime-monsoonal type of climate. Tropical monsoon brings about heavy and incessant rain in summer, and the rest of the year gets peppered with often intense convective and frontal precipitation events triggering flash floods in places with predisposing catchment characteristics.

#### 3.2 Data

Reforecast datasets from the GLoFAS-SPT model were provided by science partners at Brigham Young University (BYU). The forecasts are daily forecast re-runs from January 2014 to the first quarter of 2020, with a full set of 51-member ensemble out to 15-day lead time. The model forecast verification was based on the matched set of predictions and observations from hydrological stations selected in this evaluation. While it is possible to extract real-time forecasts to a certain date in the past, such archived sets were not used in this study pending a fix in the format, but more importantly, real-time observations were

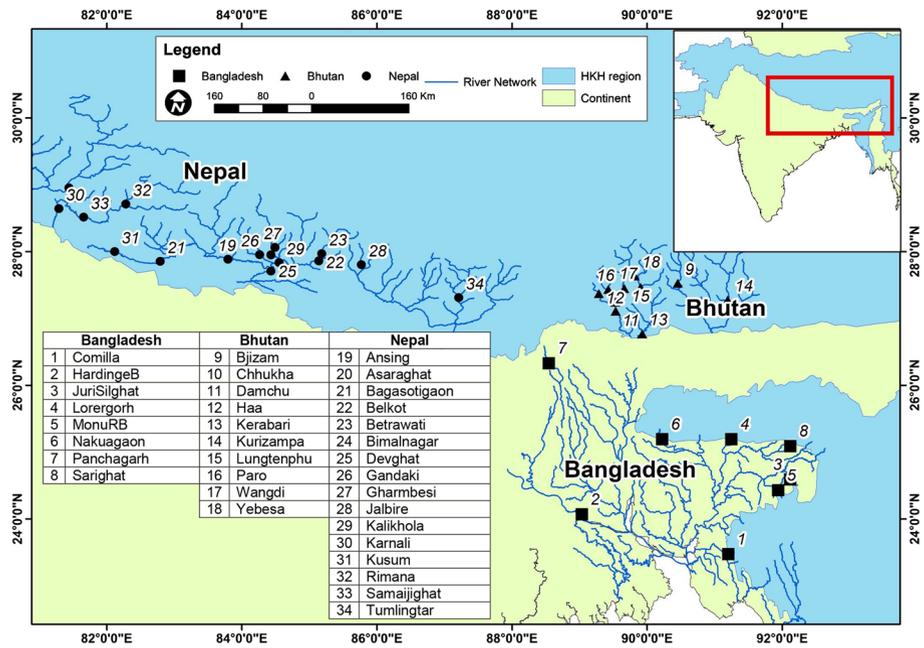


Fig. 1 Study sites across Bangladesh, Bhutan, and Nepal

not available to perform verification on the fly. Likewise, the HIWAT-SPT generated predicted outputs are also archived and maintained on a backup server in ICIMOD. Since the standard web interfaces enable only mean, standard deviation, and range of daily predictions for download, full extraction of model outputs was done programmatically with API calls and retrieval functions to the local data servers or cloud data stores in ICIMOD.

Observed discharge and/or stage data were provided by focal partner agencies from Flood Forecasting and Warning Centre (FFWC) in Bangladesh, National Center for Hydrology and Meteorology (NCHM) in Bhutan, and Department of Hydrology and Meteorology (DHM) in Nepal, either freely or against payment of a nominal fee for data processing and handling prevailing at the time of data request. For Bhutan and Nepal, discharge observations have a daily temporal resolution, while those from Bangladesh are provided with sub-daily (6-hourly) time step of river stage readings. All river stage readings are aggregated to daily averages and then converted into discharge using the rating curves provided by FFWC.

### 3.3 Methods

The model validation and forecast verification process was designed around tripartite engagement of AST, ICIMOD, and partners in Nepal and Bangladesh, and later extended to include Bhutan. Validation of the back-end modeling systems and forecast verification has emerged as a prerequisite step and priority need before technologies are accepted and transferred. Verification is being conducted focusing on predictions generated in real time and archived on daily basis to evaluate and investigate the performance of the SERVIR flood forecasting tools: GLoFAS-SPT and HIWAT-SPT. The verification period differs

across the three countries depending on the available observed discharge time. The validation period for GLoFAS-SPT starts from January 1, 2014, to the end of observed data availability from national partner agencies. In the case of HIWAT-SPT forecasts, performance is assessed from 2018 onwards from April to September each year as the model run is constrained to this period. In total, modeled forecasts are validated at over 20 observational stations in Nepal, seven in Bangladesh, and 10 in Bhutan by sampling a wide range of river basin characteristics and channel morphometry. The schematic of the verification workflow programmed into the Microsoft Excel spreadsheet application is shown in Fig. 2. An integrated verification method offers a more robust framework for greater insight into the quality, skill, value, and consistency of the forecasts to ensure the systems are fit for purpose and serve multiple interests or objectives. A select set of deterministic and probabilistic metrics and skill scores are used to evaluate the 51-member ensemble predictions over a 15-day forecast horizon. Forecast skill is assessed against climatology and persistence as reference forecasts. The following section provides information on different types of verification techniques used in the study. Since the tools are primarily developed for flood warnings, a threshold of 90% exceedance was selected for all sites across the three countries. According to Alfeiri et al. (2014), the 90th percentile of the flow time series is generally used as a discriminating criterion for flood events with a corresponding non-exceedance probability of 0.9.

### 3.3.1 Probabilistic validation

The GLoFAS-RAPID probabilistic forecasts are derived through transforming the ensemble set into a binary probability of flood dichotomy weather forecast is verified as flood or non-flood event. Forecast uncertainties are quantified using an ensemble approach and expressed in probabilistic terms. The 90th percentile of flow time series from the sampling period climatology is taken as flood threshold corresponding to the non-exceedance

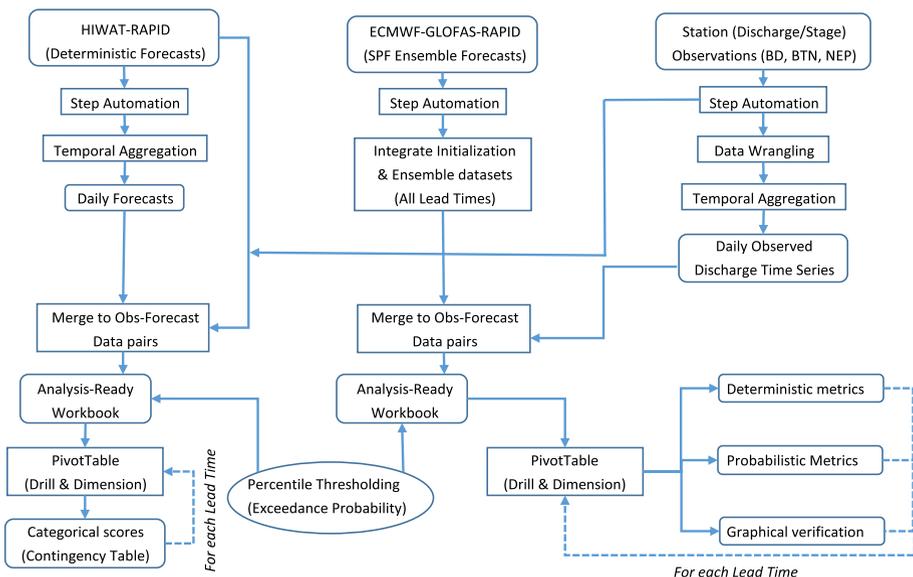


Fig. 2 Workflow for the Excel-based validation and verification procedure

probability of 0.9 for all sites across three countries (Nepal, Bangladesh, Bhutan). The GLoFAS-RAPID is further assessed through graphical measures like the reliability, talagrand, likelihood diagrams, and area under the receiver operating characteristics (AUROC). Brier score and skill score were used as numerical summary metrics to evaluate probabilistic forecasts in detecting flood days ahead of actual occurrence. These are common measures used widely in ensemble forecast assessments (Wilks 2006; Jolliffe and Stephenson 2003). This method is not applied to HIWAT-RAPID that provides only deterministic forecasts without expression of uncertainty.

### 3.3.2 Deterministic validation

Deterministic validation was carried out by comparing the time series observed data against the mean of 51 ensemble forecasts from GLoFAS-RAPID for different lead times. Using ensemble mean simplifies the evaluation of a generalized outlook of forecast performance and skill using deterministic validation metrics after filtering out smaller unpredictable scales. Popularly used error stats are like root-mean-square error (RMSE), Pearson correlation coefficient (*R*), Nash–Sutcliffe efficiency (NSE), Kling–Gupta efficiency (KGE).

### 3.3.3 Categorical verification

Categorical verification uses time-matched forecast and observation pairs based on pre-defined criteria or threshold set at the 90th percentile of observed discharge time series. This threshold to distinguish flood from non-flood events is applied to both GLoFAS-SPT and HIWAT-SPT forecasts. A simple scheme of 2×2 contingency tables is used to assess the forecast performance in terms of the joint distribution of flood dichotomy. A common set of binary scores and summary statistics are calculated from the elements of the contingency tables (Table 1).

Table 2 provides a summary of numerical metrics and graphical representations adopted in this work to ensure model performance and forecast quality are evaluated on attributes relevant to the context of forecast-based early warning systems. Limiting the scope of study around single-value error metrics may not give the full picture of forecast goodness in terms of value proposition to the decision context of hazard mitigation and risk reduction. Qualitative verification in probabilistic space is performed using graphical methods to convey often neglected but useful information about forecast quality, with the exception of BS computed as mean squared error of probabilistic forecast. Their applications are confined to verifying GLoFAS-RAPID ensemble forecasts categorized as binary floods and translated into probability measures derived from the conditional and marginal distributions. Ensemble means at each time step and over the forecast, the horizon is evaluated using deterministic statistics for accuracy, association, and correspondence. Categorical scores

**Table 1** A simple 2×2 contingency table

	Observed	
	Yes	No
<i>Forecast</i>		
Yes	Hit	False alarm
No	Miss	Correct negative

**Table 2** Summary of error stats used in the study

Error stats	Significance	Range (optimum value)
<i>Probabilistic validation (GLOFAS-SPT)</i>		
Reliability diagram	Correspondence between Observed relative frequency vs forecast probability estimated as the fraction of ensemble members exceeding certain flood threshold. Conditioned on forecast	Follow diagonal line
Sharpness	Histogram for each of the forecast probability bins	Tendency to produce 0% and 100% probability
Likelihood diagrams	frequency of each forecast probability when a flood occurred and when it did not	Mean difference of conditional distribution
Rank histogram	Frequencies of the rank of the observed data aggregated with ensemble members	Flatline
The area under the receiver operating characteristics	ROC curve is constructed by plotting False Alarm Rate (FA) against Hit Rate (HR)	0.5 to 1 (1)
Brier score	Integrated accuracy over full PDF. Mean square probability error. Verify reliability, resolution, and uncertainty attributes of probability forecasts of binary events	0–1 (0)
Brier skill score	Measures relative accuracy of the probabilistic forecast over a reference forecast	0–1 (1)
<i>Deterministic validation (GLOFAS-SPT)</i>		
Mean error (ME)	Additive bias in the forecasts, quantitative estimate of forecast error using the ensemble mean	–∞ to ∞ (0)
Root-mean-square error	Measures “average” error, weighted according to the square of the error	0 to ∞ (0)
Pearson correlation coefficient (R)	Temporal correspondence between mean forecast and observed, i.e., degree of a match at any point in time over the forecast horizon	–1 to 1 (1)
Nash–Sutcliffe efficiency (NSE),	Measures the departure from the unity of the ratio between the mean squared error of ensemble mean and the variance of the observations; quantify the accuracy of the ensemble mean	–∞ to 1 (1)
Kling–Gupta efficiency (KGE)	Composite efficiency score based on temporal error ( $r$ ), bias error, and variability error	–∞ to 1 (1)

**Table 2** (continued)

Error stats	Significance	Range (optimum value)
<i>Categorical validation (GLoFAS-SPT and HIWAT-SPT)</i>		
Heidke skill scores	It provides information on the accuracy of the forecast relative to that of random chance	$-\infty$ to 1 (1)
Peirce skill scores	It provides information on how well did the forecast separate the "yes" events from the "no" events	- 1 to 1 (1)
Gilbert skill scores	It measures the fraction of forecast events that were correctly predicted, adjusted for hits associated with random chance or how well did the forecast "yes" events correspond to the observed "yes" events accounting for hits due to chance	- 1/3 to 1 (1)
Bias	It provides information on whether the forecast system tends to under-forecast (BIAS < 1) or over-forecast (BIAS > 1) events	0 to $\infty$

are calculated for HIWAT-RAPID flood events based on threshold exceedance. The joint distribution of flood dichotomy is used to compute categorical scores describing the same forecast attributes as in deterministic and probabilistic terms.

## 4 Results

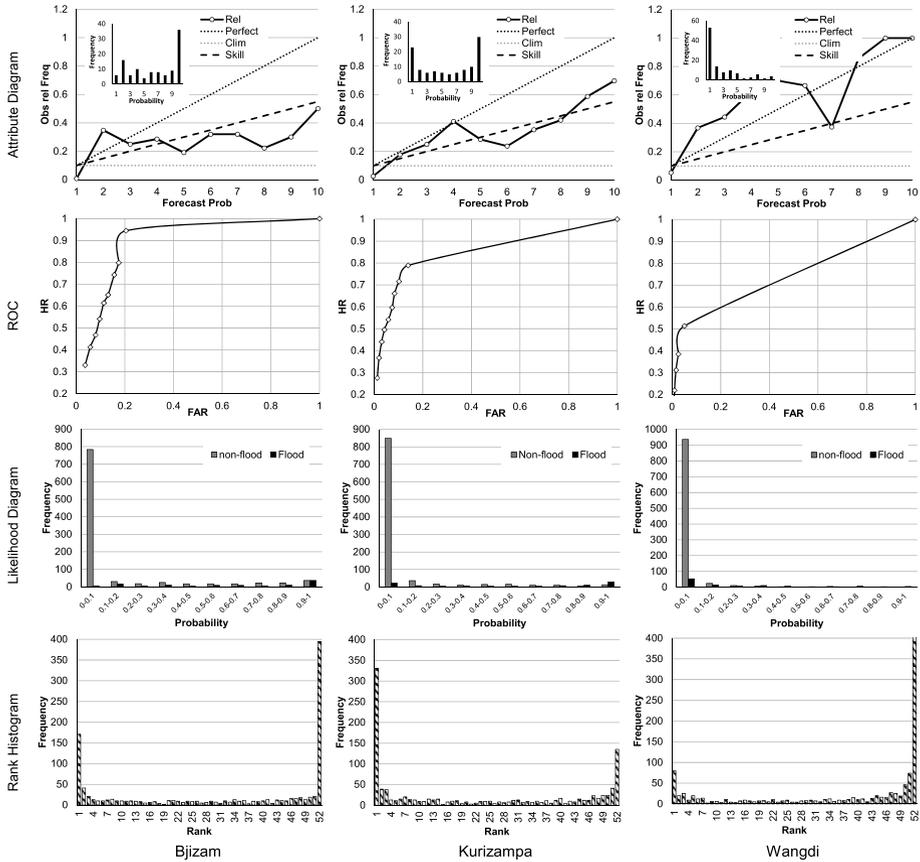
Verification is performed specifically to check on the verified claims of quality, value, and reliability of the coupled GLoFAS-RAPID flood prediction model. The HIWAT-RAPID flash flood prediction system has been evaluated only for few sites in Nepal and Bangladesh due to the want of quality observations from sites prone to flash flooding. HIWAT-based predictions present a unique challenge for verification, as precipitation forecasts are directly translated into streamflow. As such, evaluation of HIWAT-SPT predictions is limited to scores computed from the joint distribution of forecasts and observations decomposed into two sets of conditional and marginal distributions. Graphical visuals are mainly used to evaluate the correspondence in timing and magnitude of flood peaks between forecasts and verifying observations. Qualitative verification is supplemented with categorical statistics computed from the elements of the contingency table. At places throughout this article, results are sometimes presented for a select subset of verifying stations for clarity and implied representativeness over a range of variabilities in basin characteristics, climatic conditions, and the physical and human geography of the study region.

Figure 3 presents the important attributes of probability forecasts when using ensembles in flood prediction from three verifying sites in Bhutan as a case in point, and showing only results from the daily day-5 forecast lead time. Subsequent sections of the paper will dwell in greater detail on the results from Bangladesh and Nepal as these are the main focus of validation. However, references shall be made at appropriate instances to Bhutan analysis to appreciate the validity of GLoFAS-SPT forecasts beyond the SERVIR focal countries. For ensemble forecasts, verification plots are generally based on day-5 lead time of the 15-day forecast horizon unless indicated specifically. This forecast day is assessed as most representative of the general behavior of GLoFAS ensemble streamflow prediction, and a frequent lead time used by many successful flood early warning systems (Alfieri et al. 2014).

### 4.1 GLoFAS-SPT

#### 4.1.1 Diagnostic verification of probability forecasts

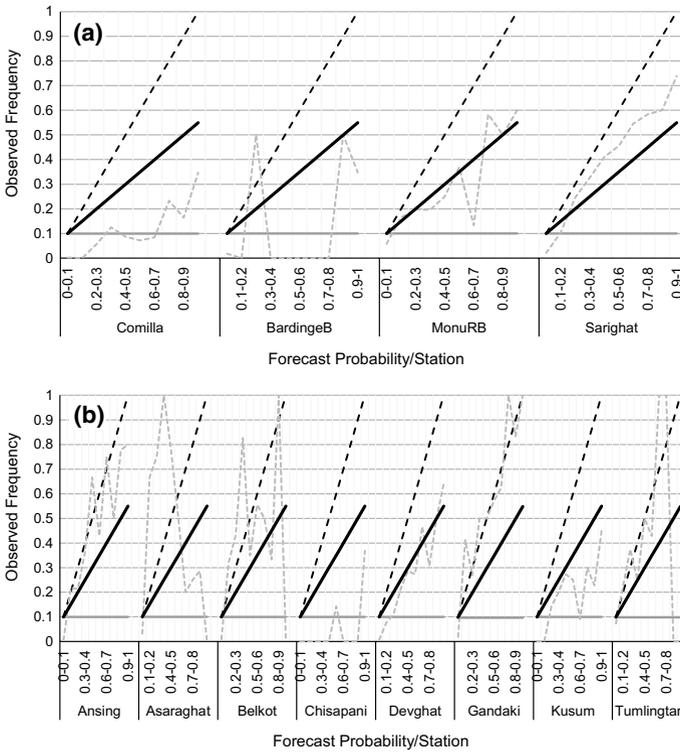
Instead of using hydrographs and scatter plots for a visual overview of the forecast systems performance, reliability diagrams, together with rank histogram and likelihood diagram, are used as diagnostic tools for assessment. Figures 3 and 4 show the reliability diagram for day-5 forecast lead time of selected stations in Bhutan, Bangladesh, and Nepal. GLoFAS-SPT coupled model has a general tendency to over-forecast discharge in most of the rivers of the verification domain. Reliability tends to be a little better on the lower end of the probability scale, but it is not a dominant characteristic of the flow conditions. The dominant trend is toward overconfident probability forecasts as evidenced by the higher forecast probabilities than observed relative frequencies. Results from Nepal sites show clear evidence of verification issues associated with under-sampling due to short periods of usable observations. The results are comparatively better from Bangladesh and Bhutan.



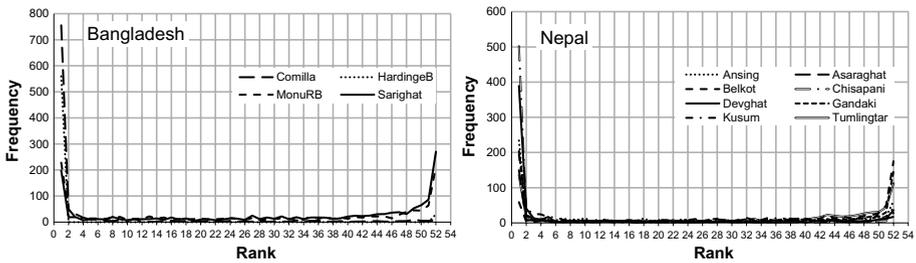
**Fig. 3** Qualitative verification of GLOFAS-SPT forecast goodness using graphical techniques for three hydrological stations in Bhutan at day-5

Overall, the GLoFAS-SPT probabilistic forecasts are not very reliable, but quite competent in resolving events with different frequencies of occurrences, i.e., high resolution. A slight trade-off is evident between reliability and forecast sharpness—a tendency to report extreme probabilities (i.e., 0 or 1). This inverse relationship suggests that sharp forecasts are not necessarily accurate. Probability forecasts are usually pessimistic at sites in the mountains with comparatively small catchment areas, and more optimistic at sites in wide floodplains commanded by large drainage basins. Results at different lead time are not presented to preserve plot readability, but suffice to note that reliability, and other attributes, degrades with increasing lead days. Best results are obtained from Wangdi (Bhutan), Sarighat (Bangladesh), and Gandaki and Ansing (Nepal).

When assessing whether ensemble distribution represents the distribution of observations, the rank histogram indicates (Figs. 3, 5) a lack of spread in early lead times and a tendency toward overconfident forecast that is consistent with forecast sharpness. The generally U-shaped histograms across sites and over lead times out to at least 10 days (not shown) suggest under-dispersive and overconfident behavior of the modeling systems. This means many observations are falling outside the extremes of the ensemble as outliers, and



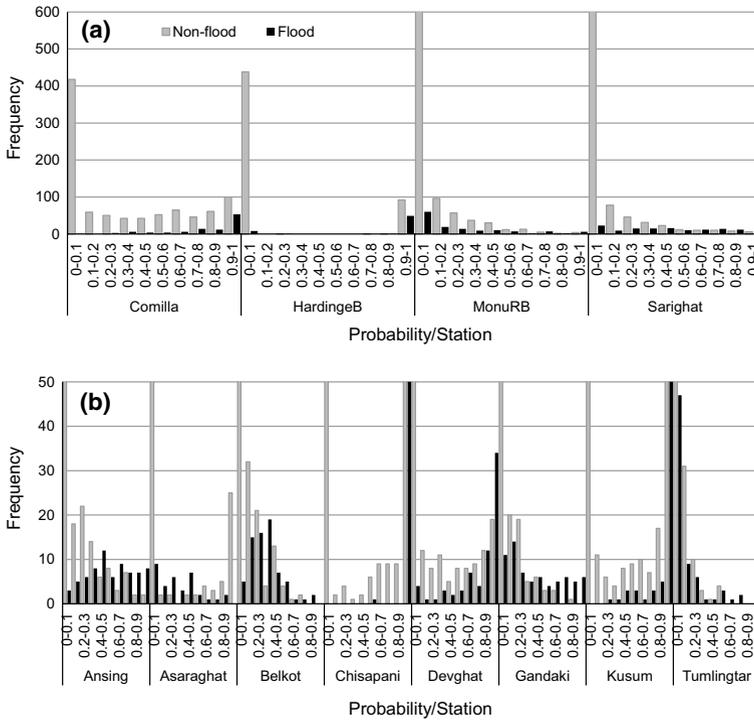
**Fig. 4** Reliability diagram for selected stations in (a) Bangladesh and (b) Nepal for day-5 lead time (Horizontal gray line shows sample climatology or no resolution, black line depicts no skill, the dash diagonal line denotes the best fit and, the gray dotted curve represents the reliability measure)



**Fig. 5** Rank histogram of observation in relation to the ensemble range for selected stations in Bangladesh and Nepal (at day-5 lead time)

the ensemble spread is too small to represent the observed distribution. The underlying “U” shape validating spread deficit also reveals smaller positive/negative early lead days biases in the streamflow forecast. However, a flat rank histogram does not necessarily indicate a skillful forecast or give a full picture of the forecast performance.

A likelihood diagram looks at the forecast distributions separately for observed flood event and non-flood event categories. Figures 3 and 6 show the likelihood diagram for the



**Fig. 6** Likelihood diagram of selected stations in (a) Bangladesh and (b) Nepal (day-4 lead time) showing the conditional distribution of forecasts given observation

marker lead day (day-5) in evaluating the discrimination attribute of the GLoFAS-SPT system conditional on discharge observations partitioned into binary yes/no of flooding. The system was able to detect non-flood with low probability (near to 0–0.1) and flood events with high probability (near to 0.9–1). Results are also suggestive of an inverse nonlinear relationship between reliability and resolution, which may partly explain why BS values are generally satisfactory across countries at least out to 7 days in advance. This will be examined further in subsequent paragraphs. Event discrimination is particularly good on rivers of large river basins and in the flood plains of Bangladesh. At sites in the interior mountains of Bhutan and Nepal, e.g., Bjizam, Wangdi, Tumlingtar, Asaraghat, Belkot, the probabilities issued for flood and non-flood show similar frequencies (Figs. 3 and 6).

We got reasonably high resolution when observations are sorted by forecasts. There is then the need to cross-verify weather forecasts are similarly sorted conditioned on observation. The receiver operating characteristic (ROC) curves in Figs. 3 and 7 show that the results across the validation sites confirm sufficiently good discrimination of flood situations by GLoFAS-SPT forecasts. ROC coordinates are computed from the hit rate (HR) and false alarm rate (FA) at each decile-binned probability threshold, which tend to cluster in the lower-left corner of the curve as expected when evaluating rare events like flooding. The area under the curve measures the discrimination skill score and is found to be close to one across the verification domain. The metric is not sensitive to the forecast bias but sensitive to the sample climatological frequency as

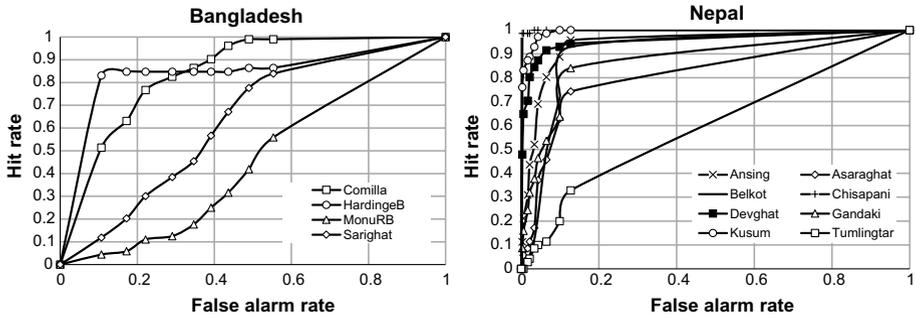
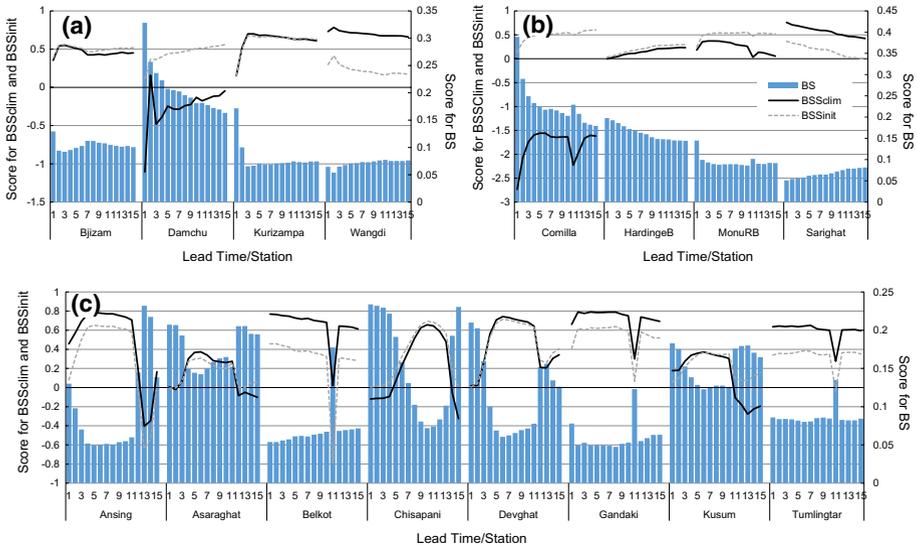


Fig. 7 ROC curves at validating sites in Bangladesh and Nepal (day-5 lead time)

explained from FA being constrained below 0.55, 0.45, and 0.1 for Bangladesh, Bhutan, and Nepal, respectively. Overall, the forecasting systems hold value in offering information to support better decision-making by the operational forecasters.

The GLoFAS-SPT prediction system that drives the service tool needs to be evaluated to ensure the probability forecasts generated are accurate, reliable, and informative under a wide range of flood-related decision contexts. While there have been several studies in verifying the system at a global scale for scientific purposes, no work was reported doing it in probability space, and few were based on real-world ground observations for administrative or economic goals. The Brier score (Brier and Allen 1951) is used to evaluate the model performance. The worst BS values were obtained at Damchu (Bhutan) and Comilla (Bangladesh), but both improve over the forecast horizon at a higher rate than in others. This contradicts our understanding that accuracy deteriorates with increasing lead time. At other locations, particularly for sites in Nepal, forecasts start at poorer accuracy on the first two lead days, improves on the 3rd–4th lead day, level off on the 10th day, abruptly show a sharp fall in the accuracy on the 11th day, and recovers thereafter (Belkot, Gandaki, Tumlingar). Others show reduced performance beyond the 10th day (Asaraghat, Devghat, Chisapani, Kusum). The rest show gradually improved accuracy over evolving lead days ahead. The first group is composed of smaller montane basins while the second group consists of rivers in larger basins in the Terai region. Widespread poor performance on the first few lead days may be due to basin lag time before runoff input to the model shows up as streamflow, or the flow may still be dominated by initial flow condition. Overall worst-performing locations are predominantly driven by very large biases.

Brier skill score (BSS) is assessed against the sample climatology and model initialization dataset taken as proxy persistence. Figure 8 shows the error and skill scores of selected stations across three countries. The model outperforms the reference forecasts throughout the entire forecast horizon, with few exceptions at Damchu in Bhutan, Comilla in Bangladesh, Ansing, Belkot, Chisapani, and Kusum in Nepal. In Nepal, both the skill scores referenced against climatology and persistence decline beyond day-10 of the lead time, with scores being better when evaluated against climatological reference than against persistence. The performance pattern is quite different in Bhutan and Bangladesh sites with some performing better against climatology and degrade with lead time (Wangdi, Sarighat), and others against persistence and improve with evolving lead time (Damchu, Comilla, MonuRB). The deficient ensemble skill in early lead days may



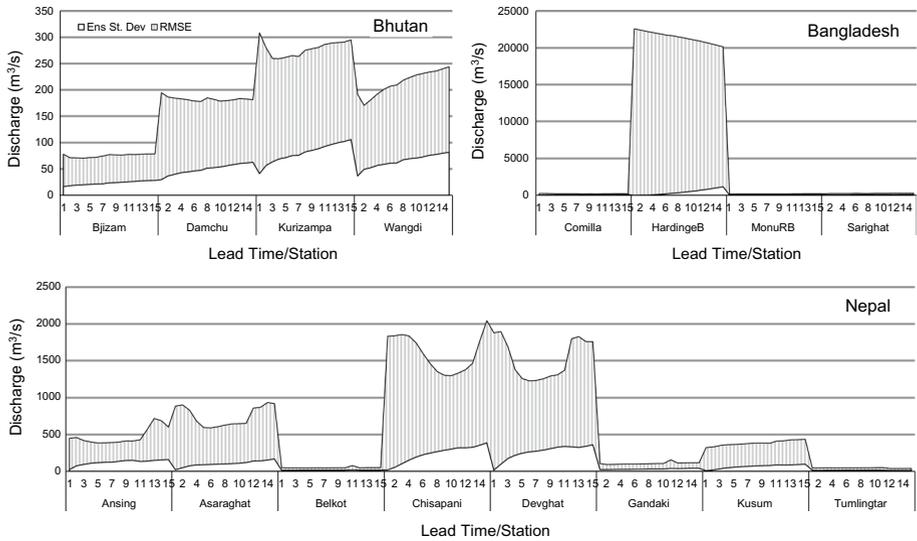
**Fig. 8** BS and BSS for selected validating stations in (a) Bhutan (b) Bangladesh, and (c) Nepal (Skill scores referenced to climatology and persistence)

be caused by the favorable initialization skill in large basins, where the inertial memory of hydrological conditions has more influence over the prediction than the runoff forcing from the model.

### 4.1.2 Deterministic verification of ensemble mean forecast time series

The mean error (ME) is used to measure the additive bias, which is rescaled to percent bias (PBias) using the observed mean discharge. PBIAS allows for comparison across widely contrasting discharge values under very different river regimes and flows conditions. Lin et al. (2019) considered a PBIAS within  $\pm 20\%$  (0.2) as an acceptable level of forecast quality. The bias in the forecasts for all of the evaluation points across selected countries shows values outside these acceptable limits indicating generally positive bias in the forecasts. It ranges from  $-0.2$  to  $2.4$  (Bhutan),  $-0.4$  to  $2.1$  (Bangladesh),  $-0.3$  to  $1.3$  (Nepal). The forecast errors in absolute magnitude are found to decrease with increasing lead time in most cases out to 10 days ahead. Small and anomalous spikes in the error were observed on day-11 of lead time in Bangladesh (Comilla and MonuRB) and Nepal, perhaps due to some unexplained modeling artifacts or abrupt change in the initial state. Generally, the model is biased to overestimate in large rivers and underestimate in smaller rivers, with few exceptions in Bhutan. The results are consistent with the previous reports that atmospheric models have a tendency to overestimating rainfall amounts (Beck et al. 2017).

The root-mean-squared error (RMSE) is fairly constant on small rivers over the forecast horizon, but gradually decreases out to day-5, stabilizes to day-10, and gradually increases again over remaining lead days on larger rivers. However, the coefficient of RMSE (not shown) is smaller for large rivers due to the attenuation of the errors and catchment size. When assessed in conjunction with ensemble standard deviation (Spread), it provides a measure of the spread-skill relationship whether ensemble predictions represent the full uncertainty range in the prediction. Figure 9 shows the RMSE is consistently greater than



**Fig. 9** RMSE score and spread-skill relation at selected stations in Bhutan, Bangladesh, and Nepal

the ensemble standard deviation, indicating the ensemble distribution is largely under-dispersive, and the distance between the two values is often in the order of magnitude particularly in larger rivers. The spread-skill assessment thus indicates that the modeled ensemble does not represent the full uncertainty range of observed streamflow, the result being consistent with measures presented earlier on a rank histogram. Since the ensemble is characterized by under-dispersive distribution, the model shows overconfidence in predictions. The poor spread-skill convergence also suggests that the ensemble did not capture the full distribution of observed outcomes.

The summary scores for KGE, NSE, and R are provided in Fig. 10 representing average performance indicating relevant aspects of changes out to 15 days in advance. In 100% of the sites, the  $R$  is above 0 with a median coefficient of 0.69 out to 15 days ahead suggesting that timing error is minimal across all sites in the three countries. The exception to this trend is with  $R$  for Kusum in Nepal falling outside the envelope delimiting values obtained at all other stations of all countries. This inconsistency is most likely related to the issue with the observed dataset as discrepancies have been identified in other metrics describing the predicted-observed relationship. A small decreasing trend is observed for MonuRB and Comilla in Bangladesh with increasing lead time, while generally found to be constant over the forecast horizon.

The fit between forecast and observation is better with KGE (mod.) compared to NSE. More than 50% of the sites gave negative values for NSE meaning that mean model predictions are worse than mean observations. The worst performance is again observed for the sites Comilla (Bangladesh), Damchu (Bhutan), Kusum (Nepal) as shown by other measures weighted heavily on the variance analysis. KGE (mod.) scores are more optimistic giving higher values trending toward the positive side, with 75% of the sites indicating better performance than merely using climatology as a surrogate forecast. The strange results from the above three sites are not replicated in this score most likely dampened by the strong positive correlation terms. While slight increasing/decreasing trends are observed at sites in Bangladesh and Bhutan, the trends in both NSE and KGE are a class aside in Nepal

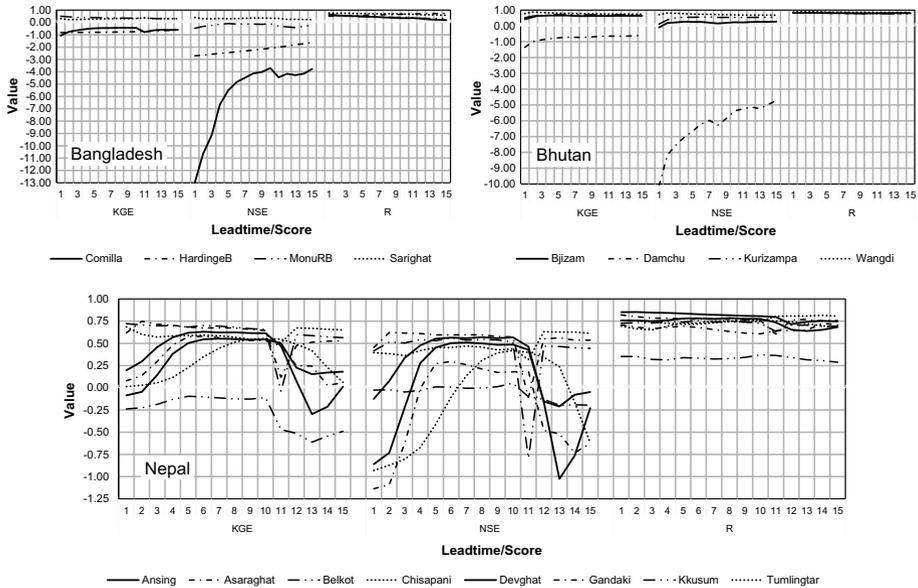


Fig. 10 KGE, NSE, and *R* value with lead time at different stations

cases. Both scores show increasing improvement up to day 5, flatten thereafter out to day 10, manifest that anomaly on day 11 with small rivers, while steeply degrading to negative scores in others. The aberration is most likely embedded in the forecast dataset for these rivers than to do with performance or observations since it persists only on the 11th day of the forecast horizon and most likely affecting all the ensemble members. In brief, the NSE and KGE (mod.) were consistently higher on validation reaches of comparatively larger upstream catchment areas. However, the efficiency scores declined with increasing lead time indicating forecast accuracy deteriorates with increasing lead time.

### 4.1.3 Categorical verification

The method is implemented to cross-verify the results obtained from metrics used in the probabilistic and deterministic sections. The analysis covers only selected stations from Bangladesh and Nepal considered adequate to serve the purpose. A select set of categorical statistics are computed from the elements in the contingency table representing the joint distribution of flood binary outcomes. Figure 11 shows those scores for flood events taken as daily discharge magnitude exceeding the 90th percentile of the observed dataset during the verification period. It was noticed that the bias remains relatively constant throughout the lead time for most of the stations. The GLoFAS-SPT is over forecasting at all the stations as indicated by the positive bias value greater than 1. The Peirce skill score for all the stations was above zero, indicating the hit rate exceeds the false alarm rate that translates into the ability of forecast to distinguish flood from non-flood events. The PSS however decreases as the lead time increases. Similar results were found for HSS and ETS. All stations had the HSS and ETS value above 0 and the skill decreases as the lead time increases, reconfirming the relative accuracy of forecasts and correspondence between the forecasts and observations. The results from these scores suggest the tool is performing

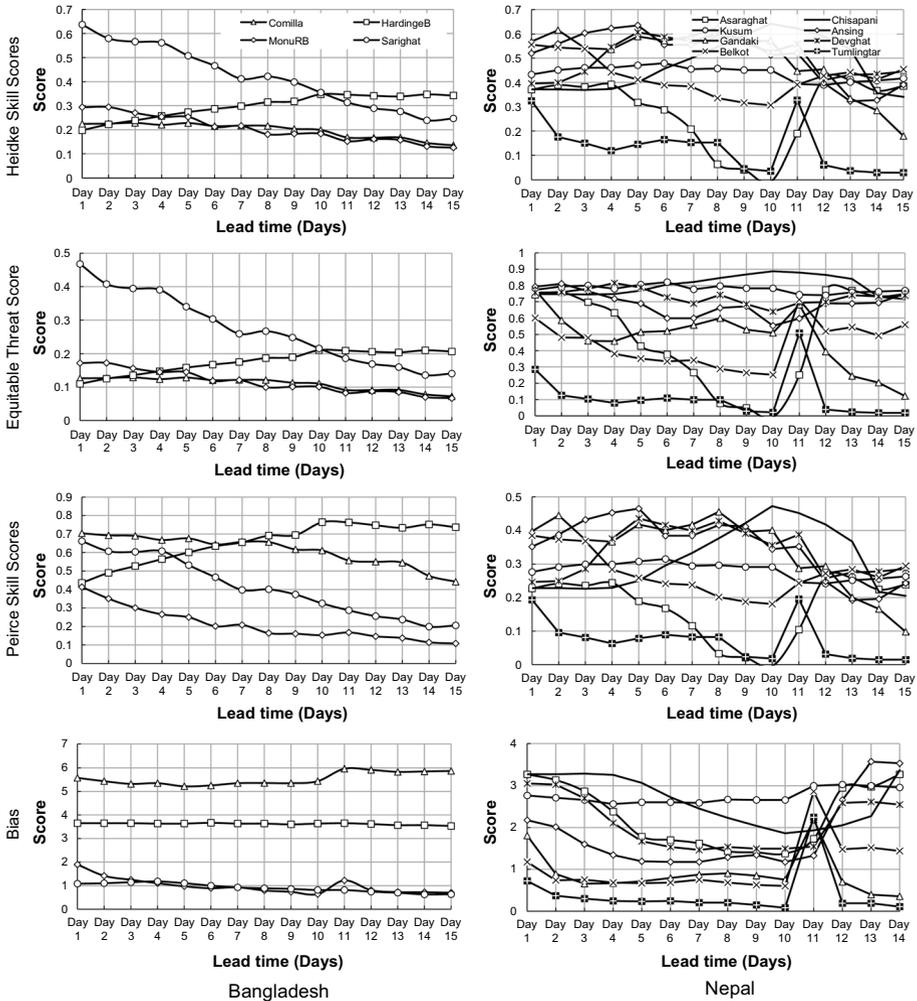


Fig. 11 Bias value and Skill score of different stations in Bangladesh and Nepal

well in detecting floods days before their actual occurrence. The results are also consistent with those obtained by other verification approaches.

### 4.2 HIWAT-SPT

The objective of these indices is to detect the overall goodness of forecasts and completeness of information considering all the cases of detection (“hit”, “false alarm”, “miss”, “correct negatives”). A perfect forecast system would produce only hits and correct negatives, and no miss or false alarm. Since the tools were developed for early warning decision support, a high value of 90th percentile signifying high flood was used as a threshold to verify forecast from HIWAT-RAPID. Due to restriction in tool output, the lead time for Bangladesh is up to three days and for Nepal is up to 2 days.

HIWAT-SPT validation was performed by categorical validation as the tool provides only a deterministic forecast from May till September. Two-by-two confusion matrix with a threshold of 90% exceedance for each station provides information about hits, false alarms, misses, and correct negatives. It was observed that the higher value lies in the lower right of the table indicating correct negatives. This is due to the threshold we chose was of higher value. It was also noticed that the value on the upper right (false alarm) was also higher indicating that the forecast of floods was made more than the observed flood, i.e., over predicting of floods. This confusion matrix was then used to calculate various scores for accessing the overall performance of the HIWAT-RAPID tool.

Figure 12 shows the bias and three skill scores for stations in Bangladesh. In all cases, the Bias was greater than 0 indicating overestimation. The bias value increases as the lead time increases. Among the five stations used for analysis, Sarighat, Comilla, and MounRB stations had comparatively fewer biases. The skill scores for three stations in Bangladesh;

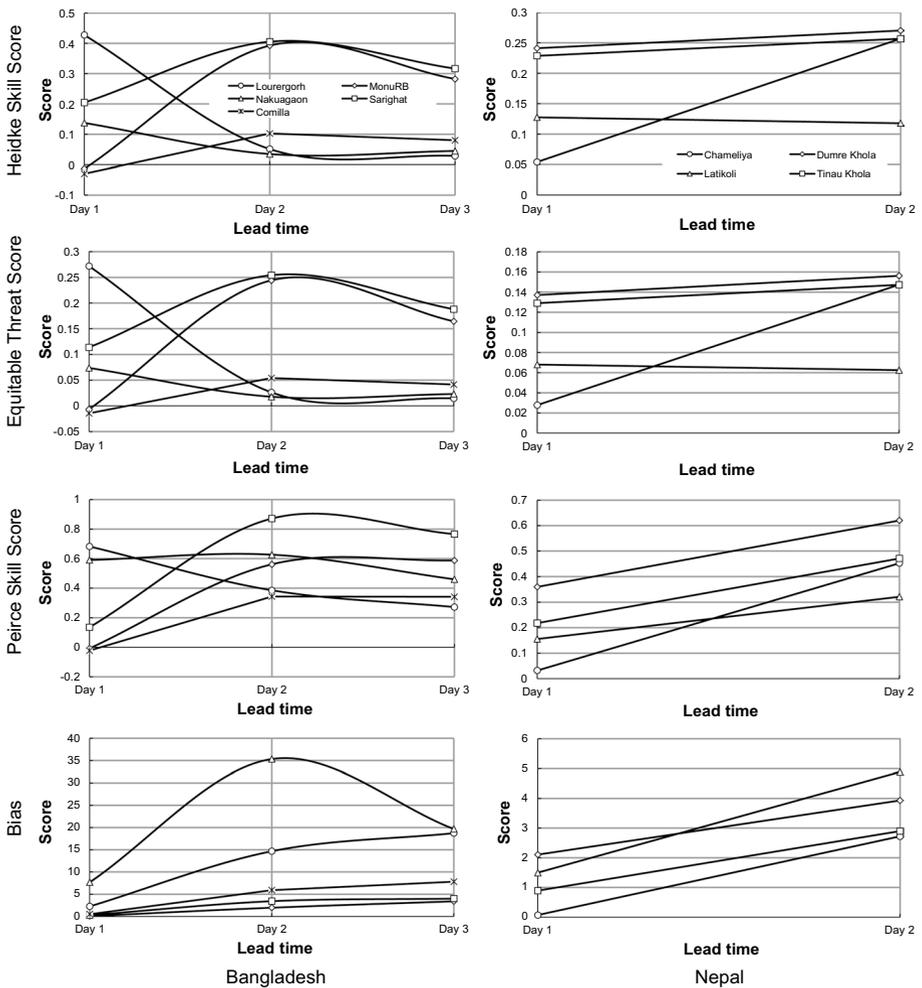


Fig. 12 Bias value and Skill score of different stations in Bangladesh and Nepal

Lourergh, Sarighat, and Nakuagaon show a value greater than one suggesting the tool has skill over a random chance. However, for stations at MonuRB and Comilla, the skill scores are less than 1, for day-1 lead and increase as the lead days increase. Figure 12 shows the results of categorical validation for stations in Nepal. As seen in the Bangladesh stations, the Bias score in the case of Nepal is also above 0 indicating overestimation of high floods. However, the bias values are much less compared to Bangladesh. The Heidke skill score, equitable threat score, and Peirce skill score for all three stations in Nepal are higher than 1, meaning the tools have skill in detecting higher discharge in these four stations. This analysis is based on forecast data for two monsoon seasons 2018–2019, and hence it is quite difficult to come to a concert conclusion. Therefore it is recommended that further analysis should be performed with multiple stations and a longer duration of data in future. This will provide us with a robust validation of the HIWAT-RAPID tool.

## 5 Discussion and conclusion

The GLoFAS-SPT is the frontend web interface to GLoFAS-RAPID model outputs and processing them into simple statistical and graphical outputs that users can easily understand and interpret. The predictions sourced by the GLoFAS-SPT application are an ensemble of 51 members, to capture the level of uncertainty in modeled forecasts based on perturbed initial conditions. The HIWAT-SPT tool disseminates and delivers the information with a deterministic outlook. To demonstrate the full benefit of these tools, it is crucial that the flood forecasting and warning services are assessed not only in the measurement space but also in the probability space to quantify uncertainties for better decision-making. For this validation exercise, the 90th percentile of the observed discharge dataset is selected as the threshold to distinguish flood and non-flood situations for evaluation. Depending on which tools are being verified, various error statistics and performance metrics are used for probabilistic, deterministic, and categorical validation. GLoFAS-SPT forecasts were verified on river reaches within which reporting and verifying observational stations are located. HIWAT-RAPID model was validated and forecasts verified at sites selected by partners on rivers and tributaries prone to flash flooding almost every year, with significant downstream settlement and development infrastructure that were impacted resulting in loss of lives, livelihood, and economic assets.

The results of our forecast evaluation provide users with an overview of the HKH-level quality of the dataset. The key limitation of the dataset is the large biases identified at several locations. The attribution of such biases in the forecast is outside the scope of this paper, but ongoing investigations like Zsoter et al. (2019) have shown biases can be introduced by the real-time land data assimilation within the HTESSEL land surface model of ECMWF-IFS setup. Other possible causes of differences between forecast and observations could result from built channel structures (dams, reservoirs, embankments, levees, diversions, etc.) within the watersheds, riverbed, and stream channels (Harrigan et al. 2020). The models do not, or if at all in a simplified way, account for flow modifying structures in the modeled river network, or operating schedules for such instream structures. Given the fundamental dependence of the forecasts on GLoFAS runoff inputs in GLoFAS-SPT and the whole of rainfall amount in HIWAT-SPT, it is worth being aware of the limitations and known issues. Also, the sample discharge climatology estimated from limited available observed data may not truly represent the long-term climatology.

The process of verifying bias and errors in forecasting systems on a probabilistic scale is a complex exercise. Hydrologic modeling systems are structured and systematized to provide complementary flood forecast information aimed at improving the FEWS through enhanced flood forecasting skills. The GLoFAS-SPT and HIWAT-SPT hydrologic modeling system performance varied spatially and over time in terms of accuracy and forecast lead time. The consecutive days of forecasts with an acceptable level of reliability, skill, and discriminatory power in flood dichotomy at medium range tend to vary across sites. Further interrogation into the paired forecasted-observed time series for each lead day provided valuable clues on the extent of forecast horizon over which the tools yield serviceable information in setting the limits of predictability. All performance metrics and quality criteria selected for the evaluation framework indicated that GLoFAS-SPT forecasts are reliable out to 10 days, and HIWAT-SPT forecasts out to at least 48 h into the future. Then again, our analyses also showed that the models failed to produce the best forecasts on day-1 lead time at all the sites. This contradicts the mainstream interpretation that forecasts are more accurate and reliable closer to runtime and deteriorate with increasing lead time. Such discrepancies observed in this study were rarely reported in the past studies with similar medium-range hydrological ensemble prediction systems such as GLoFAS. We can only stipulate at this stage that the contrasting score and skill results at day-1 lead time might indicate the influence of the initial state of hydrological conditions over precipitation (HIWAT-SPT) or runoff (GLoFAS-SPT) forcings, or could equally be attributable to the presence of systematic bias at this lead time. This aberration is also quite prominent for the basins in Nepal, and as explained earlier in the Results section, another potential cause could be the basin lag time before runoff input to the model actually shows up as streamflow as it is still dominated by the initial flow condition.

However, it is not fair to look at model performance from a forecast error perspective alone whether forecasts are capable of reproducing the observed streamflow/discharge. The GLoFAS-SPT and HIWAT-SPT predictions are particularly perceptive in the context of flood detection critical to the development of effective flood early warning systems from large to small local scales. In general, the forecasts were considered hydrologically skillful in the majority of catchments assessed, although the strength of skill varies considerably depending on location. Through the careful dissection of all important aspects of forecast, we have managed to uncover few not so obvious values in the modeled forecasts. It is notably good at producing a sharp probabilistic forecast of flood dichotomy. Reliability of probability forecasts did not emerge favorably, but it is a relative attribute assessed against observations of possibly questionable quality. The results from both systems also demonstrated a very good ability to discriminate floods from non-flood events. We also verified the forecasts to be hydrologically skillful in the majority of catchments studied, although the strength of skill varied considerably depending on location. The skill scores used in verifying both GLoFAS-SPT and HIWAT-SPT show a value greater than 0, which indicates the tools have skill over a reference forecast or random chance. The models were able to capture flow peaks and seasonality in the flow pattern with high temporal correlations. The strong influence of initial states over hydrometeorological forcing, and the systematic bias within, that drive the models did manifest strongly in results. Positive and negative biases are presented, but with discernible patterns at the regional level across Bangladesh, Bhutan, and Nepal where validation points are located. The deterministic statistics were not the best due to significant systematic bias inflating the error terms and throttling the efficiency scores. Notwithstanding the actual skill inherent in the forecasting system, it is recommended that future quantification should be based on a large enough sample size to achieve reliable and stable skill scores especially for rare events like floods. It gets

extremely difficult to verify probabilities on the tail end of the PDF as the sample size becomes too small to get robust statistics, and ensemble size may not resolve the tail of the forecast PDF.

The deterministic forecasts from the HIWAT-SPT model were evaluated during the height of monsoon season from April end to the beginning of September 2019. Categorical statistics computed from the elements in the contingency table confirmed the validation results obtained using the continuous and probabilistic forecasts. The dimensionless categorical metrics are also useful in comparing locations with different hydrologic responses. Considering the model structure and configuration, HIWAT-SPT is limited in its predictive capability up to 48–54 h ahead of event occurrence, and applicable to pluvial floods in small rivers and streams.

The results utilizing the verification approach adopted in our study point to the fact that the performance of the models, the GLoFAS-SPT in particular, could be improved further with the recalibration of RAPID parameters to more closely representing the local situation. Both models were able to capture the basin response to meteorological forcing despite the uncertainties associated with the observed dataset. The correspondence between the forecast probability and the relative observed frequency of the events they predict was particularly conclusive as indicated by consistently higher Brier score (BS). The deterministic scores verifying the correspondence, agreement, and association between forecast and observation pairs are relatively stable out to at least 10 days (GLoFAS-SPT) and effectively two days (HIWAT-SPT) of the forecast horizon. The models were able to capture flow peaks in the observed hydrographs and seasonality in the flow pattern with high temporal correlations. We verified the skill in the forecasting systems compared to reference forecasts, and far more reliable than just depending on random chance. Although the GLoFAS-SPT modeling system can predict streamflow out to 15 days lead time, the forecasts were reliable and skillful out to a maximum of 10 days in the majority of the cases, after which the performance deteriorated variably depending on the climatic regime and drainage characteristics.

National agencies collaborating in this validation work have their hydrological modeling facilities to produce streamflow forecasts and disseminate them to the end users. We believe that a comparative assessment of the SERVIR tools with the national systems could provide further insight into the contextual application and value proposition. This can help raise the prospect for integrated forecast capability in the provision of reliable and accurate information to managing floods in the region. This will be an area of focus in future for continued cooperation in research and operation in scaling up hydrological modeling as a service concept.

Further, no post-processing like bias correction of the forecasted streamflow was considered, which could be implemented in future studies with the partner countries. We hope that the verification work reported in this paper stimulates further interest in hydrologic modeling on a regional scale and transferring the results to the local level where flood impacts are most visible and consequential. The demand for model-based forecasts to inform early warning services is growing rapidly. The proliferation of modeling solutions can be confusing to end users not knowing which system fit their decision context in the absence of comprehensive evaluation information. Our work emphasizes the need to assess the performance and quality of forecasts not only in measurement space using metrics, but also in the probability space to quantify uncertainties for better decision-making. It is also hoped that the results and our interpretation provide enough evidence to deepen trust and confidence in the SERVIR flood forecasting systems as decision support tools and to integrate within the existing scheme of operational guidance systems. The ultimate integrated

FEWS system will increase the lead time from what is presently 2–3 days to 10–15 days for riverine flood, and at least out to 48–54 h ahead of an extreme convection-driven flash flooding.

In conclusion, the hydrological modeling systems that guide the riverine and flash flood forecasting and early warning tools, i.e., GloFAS-SPT and HIWAT-SPT, have the potential to enhance better decision-making for prevention and preparedness against flood disasters. The model performance and forecast quality vary over the forecast horizon defined by the respective modeling system capability. However, usable predictions still have adequate lead time to warn impact constituencies of impending floods and take appropriate action. We can confidently state here that the two systems are a major improvement over the current state of forecasts and warnings currently produced by national hydrological service agencies in terms of lead time. The limits of predictability taken as the period ahead of the forecast issuance time are 10 days for GloFAS-SPT and 48 h (2 days) in the case of HIWAT-SPT.

**Acknowledgements** This paper and material used are based upon the work supported under the SERVIR-HKH program of ICIMOD. The authors thank the reviewers whose comments helped to improve the quality of the manuscript. The hydrological observation datasets from Bangladesh, Bhutan, and Nepal are greatly appreciated.

**Author contributions** KT led this work and coordinated the verification study and writing of this article. MS contributed to the verification, methodology, and writing of this article. BB, MM, and JN provided intellectual and scientific advisory. JLSL and KS provided the forecasts and real-time forecasts from the two systems, respectively. TW, BP, and MAB provided respective national hydrological observations for the validation stations.

**Funding** Funding was provided by USAID Bureau for Resilience and Food Security (Grant No. 720BFS20IO00003).

## Declarations

**Conflict of interest** The authors declare that they have no conflict of interest.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

- Alfieri L, Pappenberger F, Wetterhall F, Haiden T, Salamon DRP (2014) Evaluation of ensemble streamflow predictions in Europe. *J Hydrol* 517(2014):913–922
- Bartholmes JC, Thielen J, Ramos MH, Gentilini S (2009) The European flood alert system EFAS—part 2: statistical skill assessment of probabilistic and deterministic operational forecasts. *Hydrol Earth Syst Sci* 13:141–153
- Basher R (2006) Global early warning systems for natural hazards: systematic and people-centred. *Phil Trans R Soc A* 364:2167–2182

- Bauer P, Thorpe A, Brunet G (2015) The quiet revolution of numerical weather prediction. *Nature* 525(7567):47–55
- Beck HE, Vergopolan N, Pan M, Levizzani V, van Dijk AIJM, Weedon GP, Brocca L, Pappenberger F, Huffman GJ, Wood EF (2017) Global-scale evaluation of 22 precipitation datasets using gauge observations and hydrological modeling. *Hydrol Earth Syst Sci* 21:6201–6217. <https://doi.org/10.5194/hess-21-6201-2017>
- Bogner K, Pappenberger F (2011) Multiscale error analysis, correction, and predictive uncertainty estimation in a flood forecasting system. *Water Resour Res* 47:W07524. <https://doi.org/10.1029/2010WR009137>
- Brier GW, Allen RA (1951) Verification of weather forecasts. In: Malone TF (ed) *Compendium of meteorology*. American Meteorological Society, pp 841–848
- CEGIS (2006) Early warning system. Final Report to the Asian Development Bank. Centre for Environmental and Geographical Service (CEGIS)
- Cloke HL, Pappenberger F (2009) Ensemble flood forecasting: a review. *J Hydrol* 375:613–626
- David CH, Maidment DR, Niu GY, Yang ZL, Habets F, Eijkhout V (2011) River network routing on the NHDPlus dataset. *J Hydrometeorol* 12:913–934. <https://doi.org/10.1175/2011JHM1345.1>
- Harrigan S, Zsoter E, Alfieri L, Prudhomme C, Salamon P, Wetterhall F, Barnard C, Cloke H, Pappenberger F (2020) GloFAS-ERA5 operational global river discharge reanalysis 1979–present. *Earth Syst Sci Data*. <https://doi.org/10.5194/essd-2019-232>
- Jolliffe IT, Stephenson DB (2003) *Forecast Verification: a practitioner's guide in atmospheric science*. Wiley, Chichester, ISBN 0-471-49759-2
- Khandu FE, Schumacher M, Awange JL, Schmied HM (2016) Exploring the influence of precipitation extremes and human water use on total water storage (TWS) changes in the Ganges-Brahmaputra-Meghna river basin. *Water Resour Res* 52:2240–2258. <https://doi.org/10.1002/2015WR018113>
- Lin P, Pan M, Beck HE, Yang Y, Yamazaki D, Frasson R, David CH, Durand M, Pavelsky TM, Allen GH, Gleason CJ, Wood EF (2019) Global reconstruction of naturalized river flows at 2.94 million reaches. *Water Resour Res*. <https://doi.org/10.1029/2019WR025287>
- Mirza MMQ (2011) Climate change, flooding in South Asia and implications. *Region Environ Change* 11:95–107. <https://doi.org/10.1007/s10113-010-0184-7>
- Mullen SL, Buizza R (2001) Quantitative precipitation forecasts over the United States by the ECMWF ensemble prediction system. *Mon Weather Rev* 129(4):638–663
- Murphy AH (1991) Forecast verification: Its complexity and dimensionality. *Mon Weather Rev* 119(7):1590–1601
- Murphy AH (1993) What is a good forecast? An essay on nature of goodness in weather forecasting. *Weather Forecast* 8:281–293
- Pappenberger F, Ramos MH, Cloke HL, Wetterhall F, Alfieri L, Bogner K, Mueller A, Salamon P, submitted (2015) How do I know if my forecasts are better? Using benchmarks in hydrological ensemble predictions. *J Hydrol* 522:697–713
- Priya S, Young W, Hopson T, Avasthi A (2017) Flood risk assessment and forecasting for the Ganges-Brahmaputra-Meghna River Basins (English). Washington, DC: World Bank Group. Available online at: <http://documents.worldbank.org/curated/en/272611508255892547/Flood-risk-assessment-and-forecasting-for-the-Ganges-Brahmaputra-Meghna-River-basins>. Accessed 21 May 2019
- Sikder MS, Hossain F (2018) Improving operational flood forecasting in monsoon climates with bias-corrected quantitative forecasting of precipitation. *Int J River Basin Manag*. <https://doi.org/10.1080/15715124.2018.14.76368>
- Snow AD (2015) A new global forecasting model to produce high-resolution stream forecasts [Master's thesis]. Brigham Young University, Provo, UT, United States
- Snow AD, Christensen SD, Swain NR, James Nelson E, Ames DP, Jones NL, Ding D, Noman NS, David CH, Pappenberger F, Zsoter E (2016) A high-resolution national-scale hydrologic forecast system from a global ensemble land surface model. *J Am Water Resour as (JAWRA)* 52(4):950–964. <https://doi.org/10.1111/1752-1688.12434>
- Souffront Alcantara MA, Nelson EJ, Shakya K, Edwards C, Roberts W, Krewson C, Ames DP, Jones NL, Gutierrez A (2019) Hydrologic modeling as a service (HMaaS): a new approach to address hydroinformatic challenges in developing countries. *Front Environ Sci* 7:158. <https://doi.org/10.3389/fenvs.2019.00158>
- Swain NR, Christensen SD, Snow AD, Dolder H, Espinoza-Dávalos G, Goharian E, Jones NL, Nelson EJ, Ames DP, Burian SJ (2016) A new open source platform for lowering the barrier for environmental web app development. <https://www.sciencedirect.com/science/article/pii/S136481521630462583d339ef534741da754fcb8f9345b826>

- Verkade JS, Brown JD, Reggiani P, Weerts AH (2013) Post-processing ECMWF precipitation and temperature ensemble reforecasts for operational hydrologic forecasting at various spatial scales. *J Hydrol* 501:73–91. <https://doi.org/10.1016/j.jhydrol.2013.07.039>
- Wilks DS (2006) *Statistical methods in the atmospheric sciences*, 2nd edn. Academic Press
- Zsoter E, Cloke H, Stephens E, de Rosnay P, MuñozSabater J, Prudhomme C, Pappenberger F (2019) How well do operational numerical weather prediction setups represent hydrology. *J Hydrometeorol* 14:1533–1552. <https://doi.org/10.1175/JHM-D-18-0086.1>

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.