Received: 5 July 2022 | Revised: 16 September 2023 | Accepted: 4 October 2023

DOI: 10.1111/jfr3.12959

SPECIAL ISSUE

CIWEM Chartered Institution of Water and Environmental Management | Journal of Flood Risk Management | WILEY

A decision-led evaluation approach for flood forecasting system developments: An application to the Global Flood Awareness System in Bangladesh

Sazzad Hossain[1,2] | **Hannah L. Cloke**[1,3] | **Andrea Ficchì**[4] | **Harshita Gupta**[3] | **Linda Speight**[5] | **Ahmadul Hassan**[6] | **Elisabeth M. Stephens**[3,6]

[1]Department of Geography and Environmental Science, University of Reading, Reading, UK

[2]Flood Forecasting and Warning Centre, BWDB, Dhaka, Bangladesh

[3]Department of Meteorology, University of Reading, Reading, UK

[4]Department of Electronics, Information, and Bioengineering, Politecnico di Milano, Milano, Italy

[5]School of Geography and the Environment, University of Oxford, Oxford, UK

[6]Red Cross Red Crescent Climate Centre, The Hague, The Netherlands

Correspondence
Sazzad Hossain, Department of Geography and Environmental Science, University of Reading, Reading, UK.
Email: sazzad176@gmail.com

Funding information
UK Research and Innovation, Grant/Award Number: NE/P000525/1; SHEAR Studentship Cohort programme, Grant/Award Number: NE/R007799/1

Abstract

Scientific and technical changes to flood forecasting models are implemented to improve forecasts. However, responses to such changes are complex, particularly in global models, and evaluation of improvements remains focussed on generalised skill assessments and not on the most relevant outcomes for those taking decisions. Recently, the Global Flood Awareness System (GloFAS) flood forecasting model has been upgraded from version 2.1 to 3.1 with a significant change to its hydrological model structure. In the updated version 3.1, a single fully configured hydrological model (LISFLOOD) has been adopted, including ground water and river routing processes, instead of two coupled models, a land surface and a simplified hydrological model, of the previous version 2.1. This study aims to evaluate changes in the simulated behaviour of floods and the forecast skill of the two GloFAS versions based on different decision criteria for early action. We evaluate GloFAS reforecasts for the Brahmaputra and the Ganges Rivers in Bangladesh for the period 1999–2018. For the Brahmaputra River, the old GloFAS 2.1 version performs better than the 3.1 version, especially in predicting low- (90th percentile) and medium-level (95th percentile) floods. For the Ganges, GloFAS 3.1 shows improved probability of detection of low- to medium-level floods compared to version 2.1, especially for lead times longer than 10 days. Both versions show limited skill for more extreme floods (99th percentile) but results are less robust for these less frequent floods given the lower number of events. Using lead-time dependent thresholds improves the false alarm ratio while reducing the probability of detection. The changes in model structures influence the model performance in a complex and varied way and forecast skill needs further investigation across regions and decision-making criteria. Understanding the skill changes between different model versions is important for decision-makers; however, focused case studies such as this should also be used by model developers to

guide future changes to the system to ensure that they lead to improvements in decision-making ability.

KEYWORDS

Brahmaputra, decision-making, flood forecast, forecast skill, Ganges, GloFAS

1 | INTRODUCTION

Global flood forecasts are now available from a few days ahead to the seasonal scale due to improvement of numerical weather prediction (NWP) models (Emerton et al., 2016) and can support a variety of anticipatory actions such as evacuation of vulnerable people, management of flood defence structures, crop planning decisions and emergency aid distributions (e.g., Coughlan de Perez et al., 2016; Emerton et al., 2020). These operational flood forecasting models use ensembles of NWPs, known as ensemble prediction systems (EPS), instead of single value deterministic forecasts, to provide flood forecasting with longer lead time (Cloke & Pappenberger, 2009). In EPS, a number of possible scenarios are generated by changing model initial conditions or using output from multiple models (Jain et al., 2018; Weigel, 2011). Ensembles of river discharge forecasts are generated using EPS of weather forecast information as input (Thiemig et al., 2015; Figure 1). Ensemble forecasts provide advantages over single deterministic forecasts by considering uncertainty, and allowing to provide predictions in the form of probability (Cloke & Pappenberger, 2009; WMO, 2012). Scientific and technical changes to the operational flood forecasting models are regularly implemented in order to improve forecasts. The impacts of these changes on model performance and forecast quality are evaluated with skill assessments to ensure robust system developments and ensure the trust of general forecast users. However, in computationally expensive global systems, which require substantial resources to evaluate fully, forecast responses to model changes and evaluation of model performance often necessarily remains focused on generalised forecast skill at the global scale (Cloke et al., 2017). Therefore, it is also essential to consider the impact of model changes on the different anticipatory decisions being made by those organisations using the forecasts including humanitarian, national and local level disaster management agencies. More user-oriented decision-led evaluation is important for operational global scale models that are used for forecast-based action by humanitarian organisations ahead of and during floods, for preparedness and real-time emergency operations (e.g., Emerton et al., 2020).

Methods of forecast verification of ensemble probabilistic forecasts differ from the deterministic forecast case

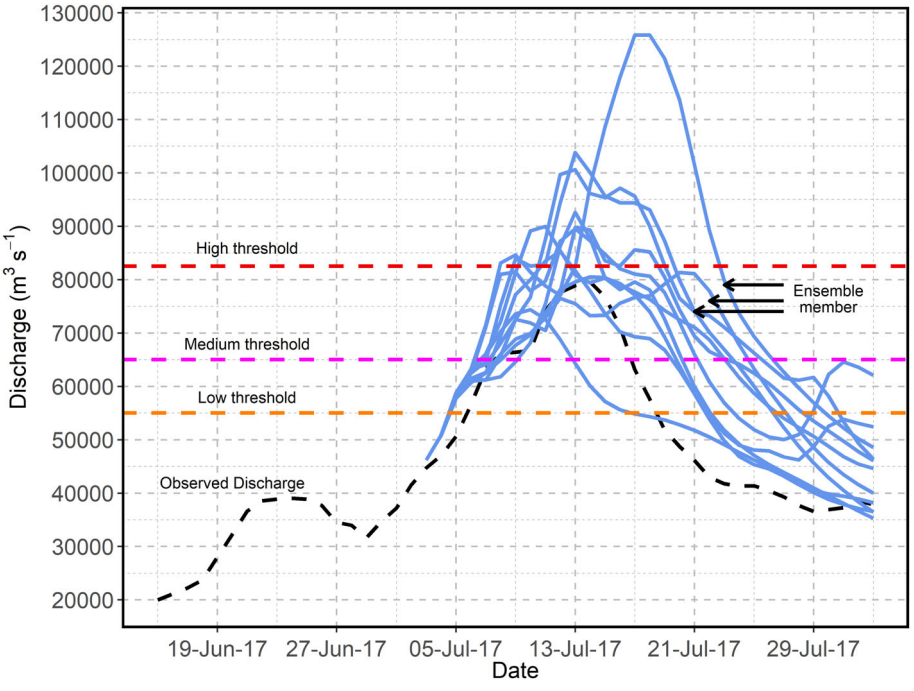

FIGURE 1 GloFAS ensemble reforecast 'spaghetti' plot for a flood event in July 2017 in the Brahmaputra River in Bangladesh. The plot presents ensemble members in solid blue lines with observed discharge in black dashed line. The horizontal dashed lines show three flood threshold levels—low, medium and high thresholds.

(Bartholmes et al., 2009) and a larger variety of accuracy metrics and skill scores are available to assess probabilistic forecasts (Weigel, 2011; Wilks, 2011). Therefore, the quality of ensemble forecasts can be assessed in different ways, depending on how they are used and interpreted and on user's criteria and perspectives (Weigel, 2011). Some common approaches to assess flood forecast skill are based on exceedance probabilities of different flood thresholds that can have operational significance (Alfieri et al., 2014; Bartholmes et al., 2009). The selection of appropriate flood thresholds and forecast trigger probabilities is important for specific early actions such as aid distribution by humanitarians or warnings to alert communities and national disaster management agencies (Coughlan de Perez et al., 2016). For instance, the European Flood Awareness System (EFAS) has two important flood thresholds, that is, the severe alert level indicating severe flooding and high alert level at bankfull conditions, which are estimated based on selected return periods (Thielen et al., 2009). Similarly, single-value (static) flood thresholds are traditionally used in global hydrological forecasting models, such as the Global Flood Awareness System (GloFAS) model (Alfieri et al., 2013) using return periods computed from reanalysis data of river discharge, which help remove the impact of biases that are often substantial in a global system as GloFAS (Harrigan, Zsoter, et al., 2023). The single value thresholds in GloFAS are based on reanalysis river discharge data and do not, account for differences of forecast climatology across lead times. However, a recent study shows that lead time dependent thresholds (with values changing across the forecast horizon) provide better results than static thresholds that are generated from reanalysis or observed river discharge (Zsoter et al., 2020).

GloFAS has been developed jointly by the Joint Research Centre of the European Commission and the European Centre for Medium-Range Weather Forecasts (ECMWF) as part of the Copernicus Emergency Management Service (CEMS), to anticipate upcoming floods for river basins all over the world with 30-day lead-time (now extended to 46 days), and the system has been operational since 2011 (Alfieri et al., 2013). Global scale evaluations suggest that GloFAS has skill in forecasting large-scale floods for the large river basins in the world (Alfieri et al., 2013; Harrigan, Zsoter, et al., 2020), for example the Pakistan flood of 2010 for lead-times 1 to 15 days (Bischiniotis, van den Hurk, Zsoter, et al., 2019).

The GloFAS forecasts have been used to support humanitarian decision-making in countries such as Uganda and Mozambique (Coughlan de Perez et al., 2016; Emerton et al., 2020). In Bangladesh, the Flood Forecasting and Warning Centre (FFWC) has been using the GloFAS extended-range forecast to predict flood events during the monsoon since 2016. Humanitarian agencies such as the Bangladesh Red Crescent, United Nations Food and Agriculture Organisation, United Nations Office for the Coordination of Humanitarian Affairs, United Nations Population Fund also use the GloFAS forecasts to support their aid distribution decision at lead times of about 10 days before the onset of floods.

There have been several model versions implemented in the GloFAS flood forecasting system since 2011, as the system is being continuously refined (Harrigan, Zsoter, et al., 2023). The recent upgrade to version 3.1 included a major change to the modelling approach, driven by the need of having a single hydrological model (for runoff production and routing) that would be easier to calibrate and consequently to improve forecasts overall at the global scale, at least for most catchments where calibration data are available (Alfieri et al., 2020). GloFAS 3.1 is based on the full configuration of the LISFLOOD hydrological model (Alfieri et al., 2013; Thielen et al., 2012), while version 2.1 was based on the coupling of ECMWF's land surface model HTESSEL (now known as ECLAND) with the channel routing component of LISFLOOD (GloFAS Wiki, 2021). The upgraded version was evaluated against the previous one by comparing the overall performance using generalist scores such as the Kling-Gupta Efficiency (KGE; Gupta et al., 2009) and looking at their distributions for thousands of catchments (Harrigan, Zsoter, et al., 2023). This synthetic comparative evaluation is necessary when developing a global system that can be used for several applications across the world, but there is a need to understand the implications of each model transition for specific users at the local scale as results may change across catchments and metrics.

For example, for the many different agencies using GloFAS in Bangladesh it is important to understand how the implementation of GloFAS 3.1 affected the ability to predict floods in the country's major flood-prone river basins like the Brahmaputra and Ganges, and how changes in the forecasting model might impact the robustness of any early action plans required for flood preparedness. For instance, the FFWC needs to improve early warning with longer lead-times to anticipate extreme flood events as well as annual floods (bankfull condition), whereas humanitarian agencies are piloting forecast-based early actions for extreme floods (for example, aid distribution) some days before the flood event. Considering these needs, the present study aims to evaluate changes in forecast performance with the recent model transition for two different river basins in Bangladesh, as exemplary case studies, and for a set of use cases to drive the evaluation.

Therefore, our study addresses the following objectives:

i. Understand any differences in simulated flood behaviour between GloFAS version 2.1 and 3.1 in terms of flood magnitude and the rise and decay of the annual flood wave.
ii. Evaluate forecast skill against lead-time for GloFAS2.1 and GloFAS3.1 taking into account (and removing the impact of) lead-time dependent biases (Zsoter et al., 2020), against the operational annual flood threshold (90th percentile).
iii. Evaluate forecast skill for different flood preparedness decisions considering several thresholds of impact, based on observed river discharge and water level.

Carrying out such a multi-criteria decision-led evaluation is crucial for any operational forecasting model transitions, not only to gain a comprehensive understanding of the multifaceted model performance changes but also for tailoring flood preparedness strategies to the local capacity of global flood forecasting systems. Hydrological model performance is known to vary with local catchment characteristics, given the variety of hydrometeorological conditions and the uniqueness of place and environmental features (Beven, 2000). On the other hand, early action plans and decision-making criteria also change with the varying risk profile and local capacities of each area of interest (Bischiniotis, van den Hurk, Coughlan de Perez, et al., 2019). Thus, local detailed evaluations of global hydrological models are necessary as a decision-led multi-criteria model performance assessment can be made more relevant only if local end-users criteria are considered (Lopez et al., 2020).

We have selected two river basins in Bangladesh, the Ganges and the Brahmaputra (Figure 2), which are affected by floods annually, from transboundary flows, that cause damages to agricultural crops, physical infrastructure and affect livelihoods (Islam et al., 2010; Mirza, 2003). To provide forecasts on an extended range is a challenge for forecasters for such transboundary basins with sufficient skilful lead-times for early action in Bangladesh. Different users such as disaster and flood managers, community people and humanitarian agencies need longer forecast lead-times for better flood preparedness decisions. The first objective of our skill assessment provides an analysis of model capabilities to simulate the evolution of high river discharges, particularly peak flows, which occur during the monsoon season. The second objective provides a skill assessment of the two model versions considering lead time dependent thresholds to find whether forecast skill improves. Finally, the third objective assesses model skill of the two model versions from different decision perspectives relevant to early action. The reforecast data of the two model versions are compared with observed floods from two river gauges: from Bahadurabad gauging station on the Brahmaputra and the Hardinge Bridge River gauging station on the Ganges River, both in Bangladesh (Figure 2). We calculate the KGE to study model performance as well as the false alarm ratio (FAR) and probability of detection (POD) scores to investigate forecast skill considering different early action decision criteria such as lead-time, flood thresholds, forecast probability and acceptable margin of error. This decision-led forecast analysis will give

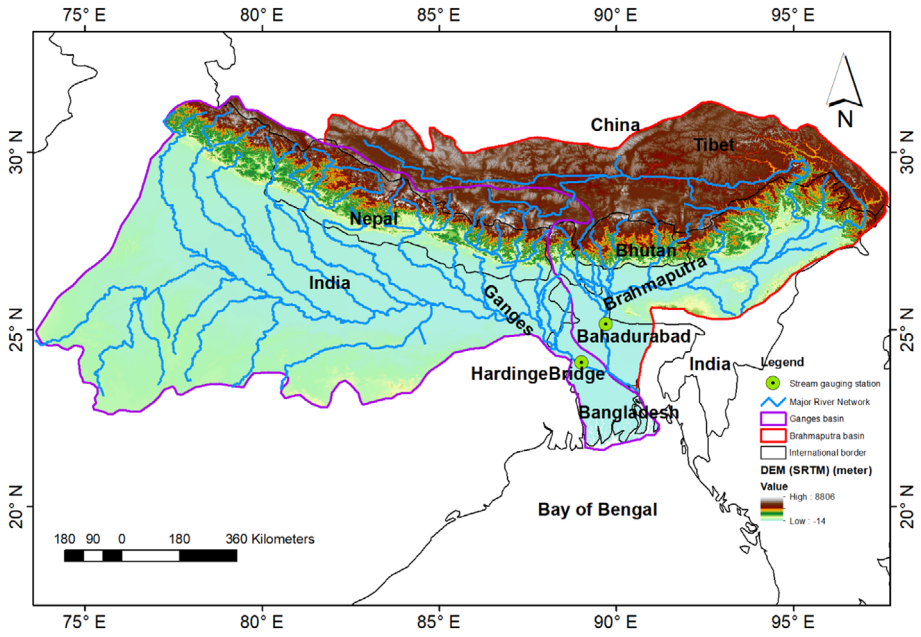

FIGURE 2 Ganges and Brahmaputra basin with stream gauging stations at Bahadurabad and Hardinge Bridge in Bangladesh.

the opportunity to understand how forecast skill changes in two different flood-prone river basins with major implications for decision-making and anticipatory actions aimed at saving the lives of millions of people and reducing impacts to vulnerable livelihoods.

2 | DESCRIPTION OF THE STUDY BASINS

Floods occur annually in the Brahmaputra and the Ganges basin in Bangladesh due to South Asian monsoon rainfall between June and September. Both rivers originate from the high altitude of the Himalayan mountain range in the Tibet region in China (Frenken, 2012). The drainage area of the Brahmaputra is estimated as 580,000 km^2, of which China, India, Bangladesh and Bhutan share 50.5%, 33.6%, 8.1% and 7.8%, respectively (Bora, 2004). The basin area of the Ganges is about 1,089,370 km^2 which extends over an area of Tibet (China) (3.67%), Nepal (12.85%), India (79.2%) and Bangladesh (4.28%; Rajmohan & Prathapar, 2013). Both basins consist of diverse landscapes, from high altitude Himalayan Mountain to floodplain delta (Figure 2) and despite being close and connected (the Brahmaputra flows downstream into the Ganges), they present different characteristics, with different river topologies and diversified climatic patterns (Mirza et al., 1998). In both basins, around 60% to 70% of annual precipitation falls during the monsoon (Bhattachaiyya & Bora, 1997; Immerzeel, 2008), and there is strong spatial and temporal variation in monsoon rainfall (Immerzeel, 2008), with higher precipitation in the Brahmaputra basin, especially in upstream areas, with respect to the Ganges (Mirza et al., 1998). The average river discharge at Bahadurabad on the Brahmaputra and Hardinge Bridge on the Ganges river during the monsoon season (June–September) is 41,000 and 23,314 $m^3\ s^{-1}$, respectively (BWDB, 2017). The flood characteristics of the two basins vary remarkably in terms of magnitude, timing and duration (Figure 3), with large interannual variability for both basins. Flood records show the Ganges River experiences floods in August and September, whereas the Brahmaputra can experience multiple flood pulses from June to September.

Flooding in Bangladesh is defined by the FFWC as when the river level exceeds the 'danger level' threshold (usually 90th percentile). At the danger level, floodwater starts to cause damage to property, crops or infrastructures. A 'severe flood' is defined when river discharge reaches the 99th percentile. The flood characteristics in terms of duration can vary from a few days or more than a month. The most impactful floods along the Brahmaputra in Bangladesh occurred during the 1998 monsoon when both rivers experienced synchronisation of flood events and the flood duration was more than 2 months (Islam et al., 2010; Siddique & Chowdhury, 2000). The flooding situation becomes devastating in Bangladesh when floods peaks of the river basins synchronise (Islam et al., 2010; Mirza, 2003). Recent successive extreme flood events in 2016, 2017 and 2019 are characterised by exceedance of the previous maximum high water level in conjunction with an unusually rapid rise in water levels in the Brahmaputra basin (Hossain et al., 2019; Hossain et al., 2021). Interestingly, the Ganges River in Bangladesh experiences comparatively less flood events after 2004 (Figure 3b). The trend analysis of annual maxima of both water level and discharge show a positive trend for the Brahmaputra and a negative trend in the Ganges at the two stream gauging stations located in Bangladesh (Figure S1). This may be due to the human alterations to the natural flow of the Ganges which is affected by the anthropogenic interventions such as construction of dam and reservoirs to divert flow (Khan et al., 2014). For instance, diversions occur at Farakka barrage just 50 km from the international border of Bangladesh (Paura, 2004).

3 | GLOFAS FLOOD FORECAST SYSTEM

GloFAS provides operational (real-time) forecasts daily for the whole world at 0.1° resolution (in versions 2.1 and 3.1; recently further increased to 0.05° in the latest version 4.0 release) and with 51 ensemble members (Harrigan, Zsoter, et al., 2023). Forecasts are made freely available through a web interface, with different forecast layers available including the probability of exceeding different return period flows out to 30 days. GloFAS uses global scale NWPs and a hydrological model to generate flood forecast information. Daily meteorological forcing is provided from the Integrated Forecasting System ensemble of the ECMWF weather forecasts, with 51 ensemble members (50 perturbed and 1 control simulation) with variable grid resolution; ∼18 and ∼36 km horizontal resolution for up to 15 days and 16 to 30 days, respectively (Alfieri et al., 2013; Harrigan, Zsoter, et al., 2023). To initialise the hydrological simulation, GloFAS uses ERA5 reanalysis river flow data along with the first day of the control member of the ECMWF ensemble forecasts (Figure 4).

In GloFAS2.1 (and previous versions) the hydrological component consists of ECMWF's land-surface scheme ECLAND (previously known as HTESSEL) (Balsamo et al., 2009; Pappenberger et al., 2010) coupled to a

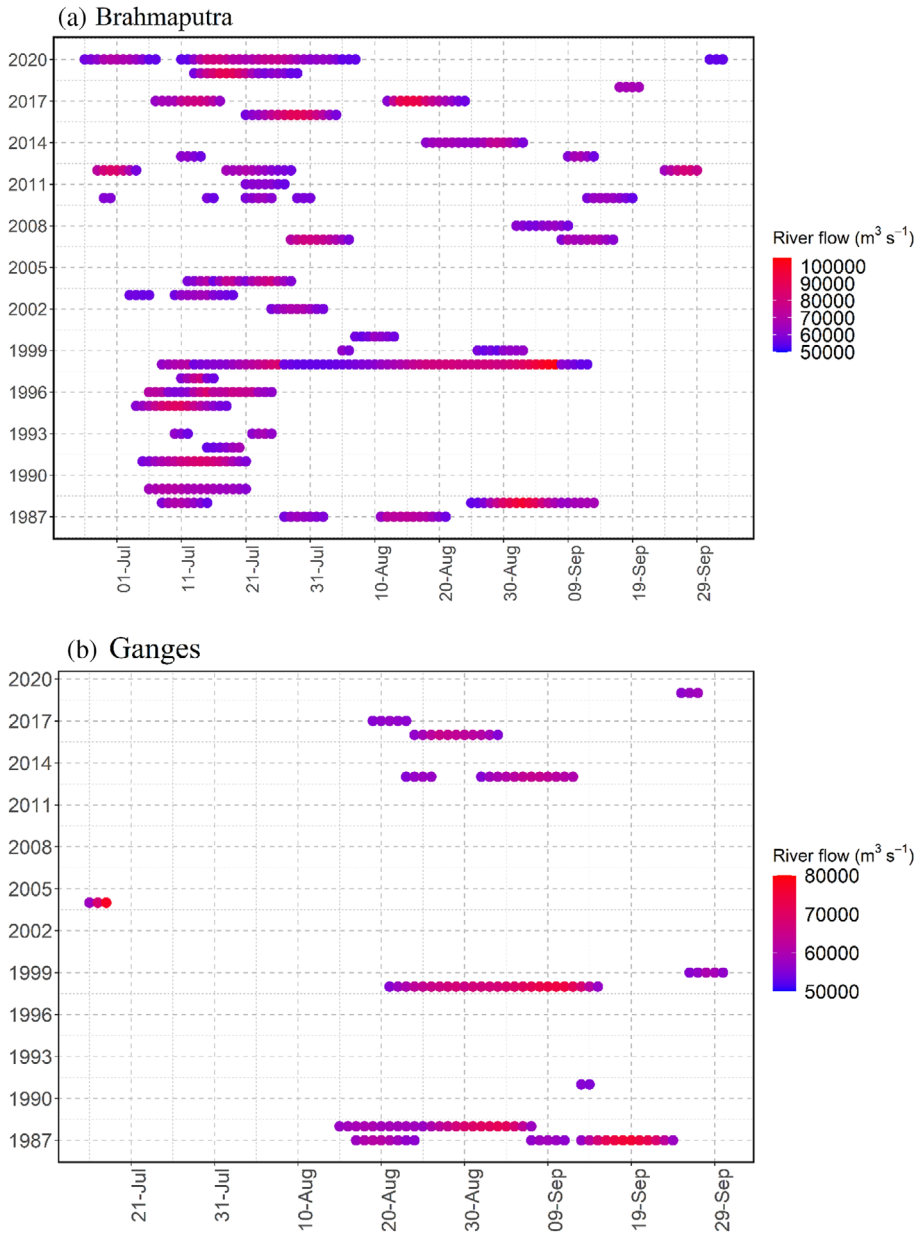

FIGURE 3 Floods at the Bahadurabad station on the Brahmaputra River (a) and at Hardinge Bridge on the Ganges River (b). Dates indicated by a coloured dot show when river discharge exceeded the flood threshold (90th percentile). The colour indicates the river flow (from low, blue, to high, red). The figure is inspired by a similar one for the Brahmaputra River in Hossain et al. (2019).

spatially distributed hydrological river routing model (Van Der Knijff et al., 2010; Figure 4). The surface and sub-surface runoff are generated from ECLAND while LISFLOOD performs ground water mass balance and river flow routing. On 26 May 2021 the GloFAS operational system was upgraded to version 3.1, using a fully configured LISFLOOD model (both surface and routing components) instead of the coupled ECLAND/LISFLOOD approach (GloFAS Wiki, 2021; Figure 4). The LISFLOOD routing and groundwater model parameters were calibrated in GloFAS2.1 using daily observed streamflow data from 1287 stations over 795 catchments worldwide (Hirpa et al., 2018). However, the calibration did not cover the Indus, Ganges or Brahmaputra River basins in South Asia. The LISFLOOD hydrological model adopted for GloFAS3.1 is calibrated across more catchments (1226 river basins), including the Ganges and Brahmaputra) (Alfieri et al., 2020).

4 | DATA

Forecast datasets from two different model versions were compared against observed river discharge (both direct and derived) and water level observations at Bahadurabad and Hardinge Bridge on the Brahmaputra and the Ganges River, respectively. The observed data is available at www.hydrology.bwdb.gov.bd under FFWC data dissemination policy, whereas reforecast data is available from the Copernicus Climate Data Store (CDS) through https://cds.climate.copernicus.eu/cdsapp#!/dataset/cems-glofas-reforecast?tab=overview.

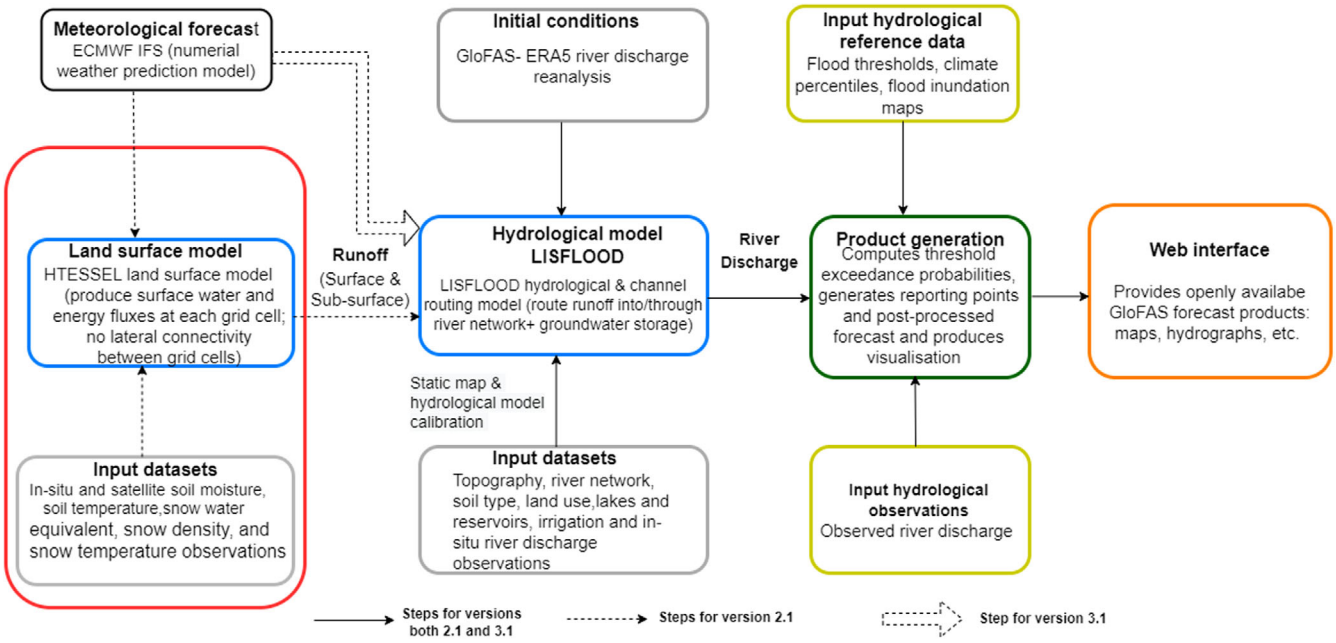

FIGURE 4 Schematic of GloFAS 30-day flood forecast system. The diagram is simplified based on Harrigan, Zsoter, et al. (2023) and GloFAS (2021). Here, steps in solid arrow lines are associated with both GloFAS versions (2.1 and 3.1), single dotted lines with version 2.1 and double dotted lines with version 3.1. A schematic of LISFLOOD can be found in Figure S2.

4.1 | Observed river discharge and water level data

River discharge and daily water level were collected from the hydrological division of the Bangladesh Water Development Board (BWDB) at the Bahadurabad and Hardinge Bridge gauging station on the Brahmaputra and the Ganges River in Bangladesh. These two stations are the only discharge measurement stream gauging stations of the two river inside Bangladesh and have a long record of water level and discharge data. The location of the river gauge can be seen in Figure 2. The water level is collected manually every 3 h interval five times in a day, starting from morning 06.00 AM to 06.00 PM with no data at night. River velocities and cross-sections are measured by BWDB usually twice in a month. Daily discharge is calculated using a rating curve. Prior to 2016 river discharge was calculated using observations from a current metre, but since this date velocity observations are collected with an Acoustic Doppler Current Profiler.

4.2 | Flood forecast data

Reforecasts for each model version (GloFAS 2.1 and GloFAS 3.1) are available for a common period from 1999 to 2018. These provide forecast data from a consistent model version and a long time period which is required for robust system evaluation (Harrigan, Zsoter, et al., 2023). GloFAS reforecasts are produced twice per week on Monday and Thursday respectively, using ECMWF weather reforecasts and initialised by ERA5 reanalysis flow (Harrigan, Zsoter, et al., 2020). Reforecasts are available for a long time period (20 years) but with only 11 ensemble members instead of 51 up to 30 days lead-time at daily time step due to computational constraints (Harrigan, Zsoter, et al., 2023).

5 | METHODS

Our approaches are divided into three parts. First, we evaluate changes in the overall accuracy of predicted river flows between GloFAS2.1 and GloFAS3.1 computing general scores (KGE and components) with respect to observed river flows. Then, forecast skill is assessed by calculating action-relevant scores (FAR and POD) with and without applying a lead-time dependent correction to remove the impact of lead-time dependent biases. Finally, the forecast skill is evaluated against decision-making criteria for GloFAS2.1 and GloFAS3.1. To evaluate the general overall model accuracy of GloFAS in simulating river flow, we use the mean of the ensemble forecasts. On the other hand, for our decision-led evaluation, we calculate action-relevant scores based on the probability of exceeding thresholds calculated from the whole ensemble, that is, percentage of forecast members above a threshold, considering the ensemble spread.

5.1 | Comparison between observed and simulated floods

The overall ability of GloFAS2.1 and GloFAS3.1 to simulate flows in the Brahmaputra and Ganges Rivers is assessed using the ensemble mean. We use the KGE (Equation 1) which includes three components: correlation, variability, and bias (Gupta et al., 2009; Kling et al., 2012) and is widely used as objective function for hydrological model calibration and evaluation (Knoben et al., 2019; Liu, 2020).

$$\mathrm{KGE} = 1 - \sqrt{(r-1)^2 + (\alpha-1)^2 + (\beta-1)^2}, \qquad (1)$$

$$\beta = \frac{\mu_{sim}}{\mu_{obs}}, \alpha = \frac{\sigma_{sim}}{\sigma_{obs.}},$$

where r represents the correlation between observation and simulation, β is the bias ratio and α is the flow variability ratio. The optimal value for these components is 1. Here, σ_{obs} and σ_{sim} are the standard deviations in observation and simulation, whereas μ_{sim} and μ_{obs} are simulation and observation mean, respectively. A KGE equal to 1 indicates perfect agreement between simulation and observation, while several authors considered that KGE <0 indicates that the mean of observation provides a better estimate than the simulation (Castaneda-Gonzalez et al., 2018; Koskinen et al., 2017) and that the model can be considered as 'not satisfactory' (Schönfelder et al., 2017). However, Knoben et al. (2019) shows that negative KGE values do not essentially indicate that a model performs worse than the mean flow benchmark, as this would rather correspond to KGE values < -0.41.

The KGE is decomposed into its three components to assess linear correlation, model bias and variability error of the reforecasts datasets against observed discharge for lead-time 1 to 30 days. The linear correlation (r) identifies any linear relationship between observed and simulated discharge (Moriasi et al., 2007) but is sensitive to outliers (Legates & McCabe Jr., 1999). The Bias ratio (β) of the KGE can be converted to percent bias (Pbias) by $(\beta-1)*100$ (Harrigan, Zsoter, et al., 2020) and provides information on model overestimation or underestimation. The variability ratio (α) is used to measure relative variation of simulated and observed flow (Gupta et al., 2009), and α >1 indicates more variability in the simulated results than in the observed data. Assessing all these three components together is important to understand how effectively the model represents the real world.

Flow duration curves are used to explore the temporal distributions of river flow in the reforecasts, and examine the duration of extreme flow above 90th, 95th and 99th percentiles. We selected three different lead-times that are relevant for early action: shorter lead-time (5 days), medium-range lead-time (10 days) and extended-range lead-time (15 days). The annual cycle of the observed and simulated floods are examined through the long-term mean for each of these three lead-times to provide a comparison of how the simulations capture the rise and decay of the annual flood wave.

5.2 | Forecasting skill for observed flood events

The objective of GloFAS is to provide early information on flood events and bias in the magnitude of flows is acknowledged but not considered to be problematic for this objective. This is because GloFAS adopts an approach whereby flood threshold exceedance probabilities within GloFAS are calculated based on thresholds obtained from the simulated flows climatology, with the assumption that when a GloFAS forecast indicates that the simulated 1 in 5 year flow (20% annual exceedance probability) threshold will be exceeded this corresponds to a real-world 1 in 5 year flow threshold exceedance. For instance, for the Brahmaputra River at the Bahadurabad gauging station the 1 in 5 year observed flow is 75,000 $m^3\ s^{-1}$ whereas the 1 in 5 year GloFAS threshold is 93,000 $m^3\ s^{-1}$. Given this, any evaluation of the GloFAS forecasting skill needs to also follow this approach. Any evaluation of model performance should consider specific thresholds which classify flood events, that is, flow above a threshold. In this study we assess GloFAS skill using a threshold of bankfull discharge (90th percentile) as defined by the FFWC. For annual floods, the bankfull condition is the flow at which water fills the channel to the top of the banks and is the threshold that indicates the onset of flooding (Ahilan et al., 2013; Wu et al., 2008). Bankfull discharge is typically estimated rather than measured and the FFWC in Bangladesh uses the 90th percentile daily flow threshold to define the onset of flooding. To define severe floods, the FFWC uses the 99th percentile of daily river flows. These percentile thresholds are calculated based on observed historical data which spans the most recent 30 years period.

Operationally, FFWC declares floods if a forecast exceeds this threshold at the Bahadurabad gauge. This provides a decision-relevant threshold while also giving a large enough sample to conduct a robust evaluation.

TABLE 1 Forecast contingency table for yes/no dichotomous method.

		Observed		
		Yes	**No**	**Total**
Forecast	Yes	Hits	False alarms	Forecast yes
	No	Misses	Correct Negative	Forecast no
Total		Observed yes	Observed no	Total

To undertake the evaluation river discharge is considered a dichotomous variable (yes or no events) and a 2×2 contingency table is calculated depending on whether the above threshold is met (Table 1). We evaluated GloFAS skill for lead-times from 1 to 30 days from 1999 to 2018 using a 50% forecast trigger probability that river discharge exceeded the 90th percentile on a particular day. Based on the contingency table, we estimated the POD (Equation 2) and false alarm ratio (FAR) (Equation 3), similar to previous authors (Bischiniotis, van den Hurk, Zsoter, et al., 2019; Passerotti et al., 2020). POD gives the percentage of flood events that are forecasted, whereas FAR provides the percentage of forecasted floods where no flood is observed.

$$POD = \frac{hits}{hits + misses} \tag{2}$$

$$FAR = \frac{false\ alarms}{hits + false\ alarms} \tag{3}$$

Two alternative approaches can be used to calculate flood thresholds from ensemble forecasts across lead times: (i) a constant threshold can be calculated using the climatology of forecasted flows at the first lead-time, and then kept static throughout the forecast time horizon; (ii) thresholds varying across lead times can be calculated. In this approach, thresholds are estimated from ensemble reforecasts data that vary with lead-time. In the operational uses of both GloFAS 2.1 and 3.1 the thresholds are kept static across all lead times, but Zsoter et al. (2020) found an improvement in forecast skill by taking into account how the thresholds should change for each lead time due to changing forecast biases. The climatology of the GloFAS model varies by lead-time, for example the 1 in 5 year return period is different at a lead time of 1 day compared to a lead time of 10 days. By using thresholds that vary by lead time (lead-time dependent) it accounts for this variation. Within this study we also calculated forecast skill for each model version using lead-time dependent thresholds similar to Zsoter et al. (2020). Here, 11 ensemble members of reforecasts are used to derive different thresholds (90th, 95th, 99th percentile) across all lead-times instead of constant thresholds. Hence, we refer to 'lead-time dependent correction' when FAR and POD are calculated using lead-time dependent thresholds while FAR and POD that are calculated based on constant threshold are referred to as 'not corrected'.

5.3 | Decision-led flood forecast evaluation

Different flood preparedness decisions have different requirements for acceptable forecast skill. We performed a decision-led evaluation which takes into account the following requirements for different decisions: (i) flood threshold (both for discharge and water level), (ii) lead-time required, (iii) acceptable margin of flood timing error (how much later the flood can arrive, and it still count as a 'hit'), (iv) acceptable frequency of false alarms and (v) acceptable hit rate (corrected forecasted floods). Different acceptable thresholds were selected for each criterion based on consultation with the key stakeholders (Table 2), including humanitarian organisations, government agencies working for disaster and flood management and the local community in flood vulnerable areas. Information was gathered via interviews with seven categories of stakeholders, including flood forecasters, humanitarians, disaster managers, vulnerable communities (Hossain et al., 2023), asking their requirements for forecast horizons, errors in flood timing, false alarms and hit rates, to be expressed in quantitative terms. The flow thresholds were chosen from the set of operational flood thresholds used by FFWC for early warning (90th, 95th and 99th percentiles) and they were associated to the relevant decisions by the decision-makers. Different decisions are made based on the different flood thresholds; for instance, the evacuation of vulnerable people to shelters is activated by national disaster managers for medium-flood thresholds (i.e., 95th percentile) that affect low lying areas (including small islands on the Brahmaputra); on the other hand, for aid distributions humanitarian organisations are interested in more extreme events (i.e., >99th percentile) and longer lead times (Table 2). There is a wide range of required lead-times for the decisions, from 3 days for the evacuation of people to a flood shelter through to 15 days for humanitarian organisations and 18 days for agriculture planning decisions. Short margins of error for the evacuation of people and livestock reflect the impact that extended time away from their livelihoods will have, whereas larger margins of error are acceptable for aid distribution by humanitarian

TABLE 2 Decision-led criteria used for forecast evaluation classified by type of decision.

Decision number	Decisions	Lead-time (days)	Flow threshold (percentile)	Acceptable delay in flood timing (days)	Acceptable false alarm ratio (FAR)	Acceptable probability of detection (POD)
d1	Evacuation of flood vulnerable people to flood shelter (people on relatively higher land)	3	99th	2	0.30	0.60
d2	Evacuation of livestock to safe place (livestock on relatively higher land)	4	99th	2	0.30	0.60
d3	Evacuation of flood vulnerable people to flood shelter (people on low lying *Char Island*)	3	95th	2	0.30	0.60
d4	Evacuation of livestock to safe place (livestock on low lying *Char Island*)	4	95th	2	0.30	0.60
d5	Household level preparedness (protecting household goods-wrapping and shifting to a safer place, storing dry foods, collecting, and saving money, storing cooking wood)	2	95th	2	0.30	0.50
d6	Pre-activation trigger for aid distribution by humanitarian agencies	15	99th	10	0.50	0.50
d7	Communication flood preparedness and response decisions of the Government agencies such as disaster management, flood management, agriculture extension at national and sub-national levels	10	95th	3	0.50	0.50
d8	Agriculture (aman rice) planning decisions by farmers (seedbed preparation to transplantation in field)	18	95th	7	0.50	0.50

agencies because the aid will still have a benefit even if the flood arrives later. Likewise, the acceptable FAR for evacuations is lower than for humanitarian operations because of concerns over maintaining trust in the forecast in the long-term. For all decisions, the 50% exceedance probability was selected in this study, based on the consultation with the flood forecasters in Bangladesh. Forecasters in Bangladesh expect that a model should be capable to predict an event with 50% forecast probability otherwise it is difficult to take relevant action. However, we also performed a sensitivity analysis of the results to the forecast probability (varying from 20% to 90%) and these will be presented in the paper (Figure 9).

These requirements (Table 2) are then used as the key parameters for our decision-led forecast evaluation framework (as Figure S3).

6 | RESULTS

First, we present a comparison of simulated flood behaviour in the GloFAS model versions at different lead-times compared to the observed river flows. We then evaluate the forecast skill for different lead-times and for specific decision perspectives. From this point, for ease of reading, we use v2.1 and v3.1 for GloFAS2.1 and GloFAS3.1 respectively.

6.1 | Comparison between GloFAS reforecast version 2.1 and 3.1 with observed river flow

GloFAS v2.1 shows better performance than v3.1 for the KGE and all its three components (Figure 5) for both the Brahmaputra and Ganges, apart from a slight improvement of correlation for the Ganges with v3.1. For the Brahmaputra, KGE values range from 0.68 to 0.81 for v2.1, and 0.52 to 0.64 for v3.1 over lead-times of 1 to 30 days (Figure 5a). For the Ganges, KGE values were higher (0.68 to 0.72) in v2.1, but have worsened for v3.1 (0.10 to 0.12) due to higher variability and bias in v3.1 (Figure 5b). Interestingly, the KGE values demonstrate improvements in performance at longer lead times, especially in v2.1, driven by improvements to the variability and bias components. The probable reasons for this are the large biases in the hydrological model initial conditions and in rainfall forecasts for extreme events at short lead times (with low ensemble spread), while the return to climatological conditions at longer lead times is better represented by the hydrological forecasts (also thanks to larger ensemble spread) and given the lower impact and frequency of extreme rainfall forecasts at longer lead times.

The bias component of the KGE shows that the bias changes with lead-time, with a very high average positive percent bias also at short lead times, that is, 16% and 21% in v2.1 and v3.1, respectively, for the Brahmaputra, and 27% and 74% in v2.1 and v3.1, respectively, for the Ganges, meaning that flow is being overestimated even in the initial conditions. Bias in v3.1 increases with lead-time before dropping from lead-time 18 days onwards, whereas in v2.1 the bias reduces with lead-time for the Brahmaputra. However, the bias is almost constant up to 18 days lead-time and then slightly increases in both v3.1 and v2.1 with a slight drop after lead time 23 days in v2.1 for the Ganges. For a global model, bias ±20% is considered acceptable (Lin et al., 2019) and ±25% is also measured as a good model simulation (Khoshchehreh et al., 2020). Here, the maximum positive bias across lead times in v2.1 and v3.1 is 19% and 24%, respectively for the Brahmaputra whereas it is 31% and 77%, respectively, for the Ganges.

The variability component of the KGE in both v2.1 and v3.1 is higher (variability ratio >1) than the observed discharge for all lead-times (1 to 30 days). In both v2.1 and v3.1 the variability increases from day 7 and falls again from day 17 (v2.1) and day 19 (v3.1) for the Brahmaputra. The variability component of the Ganges is similar to the Brahmaputra for v2.1, however the variability

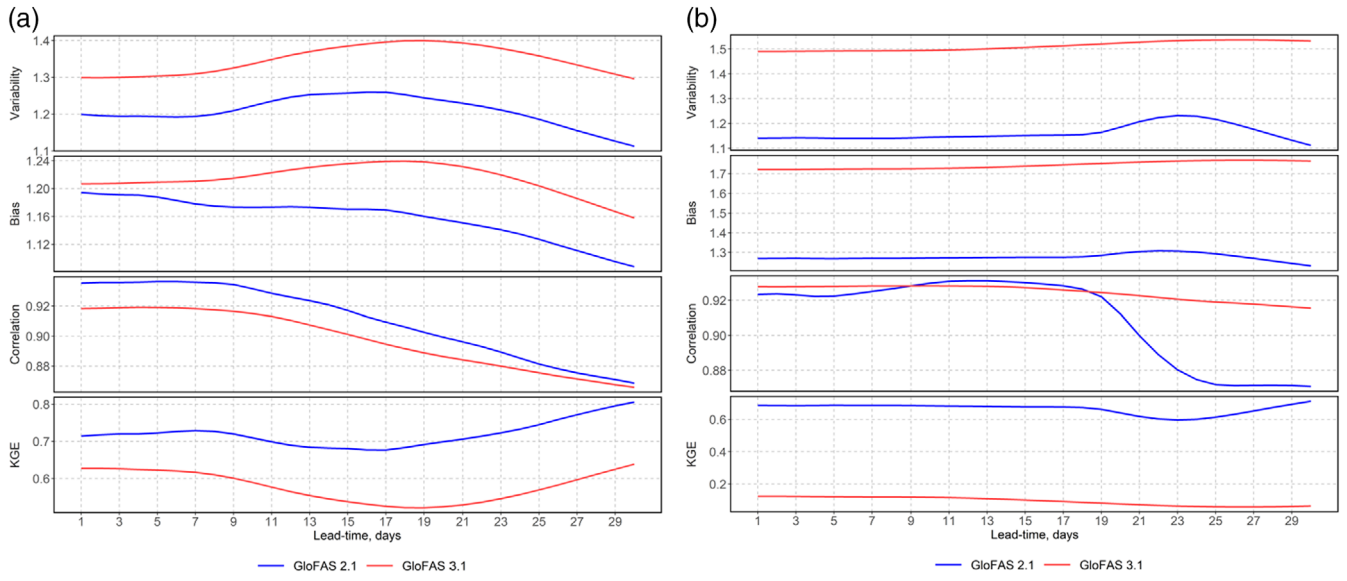

FIGURE 5 Evolution of KGE and its three components: variability, bias and correlation across lead-times 1 to 30 for GloFAS2.1 and GloFAS3.1 compared to the observed data for (a) Bahadurabad gauging station on the Brahmaputra River and (b) Hardinge Bridge gauging station on the Ganges River.

TABLE 3 Summary matrices of the model performance.

River name			Version 2.1	Version 3.1
Brahmaputra	KGE	Min	0.68	0.52
		Max	0.81	0.64
	Variability	Min	1.11	1.30
		Max	1.26	1.40
	Bias	Min	1.09	1.16
		Max	1.19	1.24
	Correlation	Min	0.87	0.87
		Max	0.93	0.92
Ganges	KGE	Min	0.60	0.06
		Max	0.71	0.12
	Variability	Min	1.11	1.49
		Max	1.23	1.54
	Bias	Min	1.23	1.72
		Max	1.31	1.77
	Correlation	Min	0.87	0.92
		Max	0.93	0.93

Abbreviation: KGE, Kling-Gupta Efficiency.

is higher in v3.1 for the Ganges compared to the Brahmaputra (Table 3). The percent variability has greater differences between versions than the bias, with mean values in v2.1 and v3.1 of 21% and 35%, respectively, for the Brahmaputra and 16% and 51% in v2.1 and v3.1, respectively, for the Ganges. The correlation component of the KGE score shows that simulated and observed discharge are strongly correlated for both versions of GloFAS ($r > 0.87$) for both rivers. This is perhaps unsurprising given the strong seasonal cycle. The correlation varies from 0.94 to 0.87 in v2.1 (above 0.9 for 19 days), while it ranges from 0.92 to 0.87 in v3.1 (above 0.9 for 15 days) for the Brahmaputra. Correlation of Ganges is slightly higher (0.92 to 0.93) in v3.1 than v2.1 (0.87 to 0.92) for the Ganges and there is a drop in correlation value from 0.92 to 0.87 in v2.1 at lead time 19 day. A summary of KGE value and its components for both the rivers are shown in Table 3. Whereas the percent bias describes an overall bias in the model, the flow duration curves (Figure 6a–c) can indicate where in the flow regime any bias is located. For the Brahmaputra, the flow duration curve shows that the wet biases are connected to the upper tail of the flow duration curve. For the behaviour in the annual cycle (Figure 6d–f), the rising-limb and peak flow are overestimated while flows in the decaying phase of the annual cycle are underestimated. Also for the Ganges, the wet biases are related to the high flows distribution with increasing overestimation from the median flow upwards (Figure 7a–c), while both rising and falling limbs are overestimated for both GloFAS versions (Figure 7d–f).

6.2 | GloFAS flood prediction skill with lead-time

For forecasts of bankfull discharge threshold (90th percentile) the FAR and POD increase and decrease, respectively, with increasing lead time, in line with an expected deterioration in forecast skill with longer lead times. The FAR in v2.1 is lower than for v3.1 across all lead times and in both uncorrected and lead time corrected versions for the Brahmaputra (Table 4; Figure 8a,b). For POD, all model versions show a similar result out to about 8–10 days, but beyond that the POD is highest for v3.1 (uncorrected). The lead time correction provides an improvement (reduction) in the FAR (0.43 to 0.40) for v3.1 from a lead time of 11 days onwards (see Table 4), but this is mirrored by a degradation (reduction) in the POD (0.67–0.60) for the Brahmaputra. For the Ganges, the lead time correction improves FAR values remarkably in v3.1 (Figure 8c) with an average FAR value decrease of about 20% (see Table 5). In comparison with the Brahmaputra, v3.1 provides better POD for the Ganges (see Figure 8d, Tables 4 and 5). The lead-time correction accounts for changes in the bias of the models at different lead times (as shown in Figure 8), with the increase in FAR and POD for longer lead times in v3.1

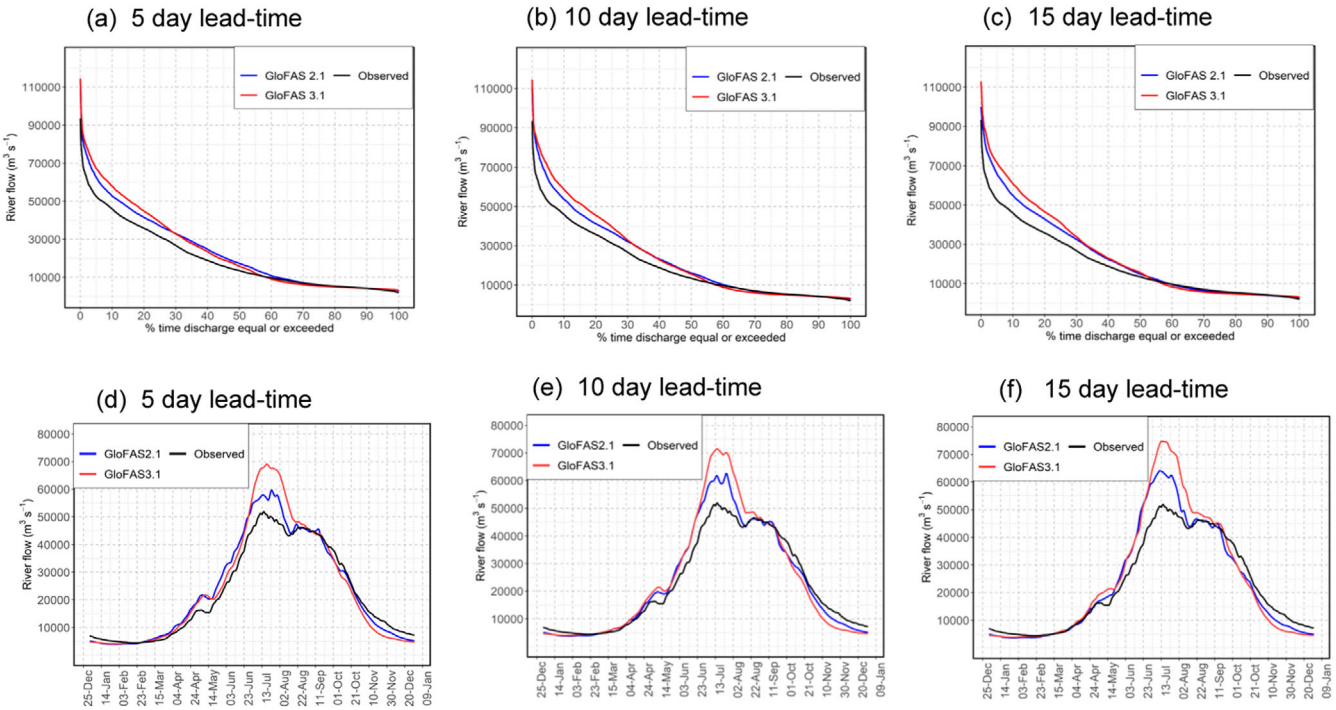

FIGURE 6 Flow duration curves of GloFAS reforecasts v2.1 and v3.1 and observed river flows for the Brahmaputra river at: (a) 5-day lead-time, (b) 10-day lead-time and (c) 15-day lead-time. Annual cycle of observed flow and the GloFAS reforecasts (v2.1 and v3.1) for different lead-times: (d) 5-day lead-time, (e) 10-day lead-time and (d) 15-day lead-time based on the long-term mean over 20 years (1999–2018). Blue, red and black lines are GloFAS v2.1, v3.1 and observed respectively.

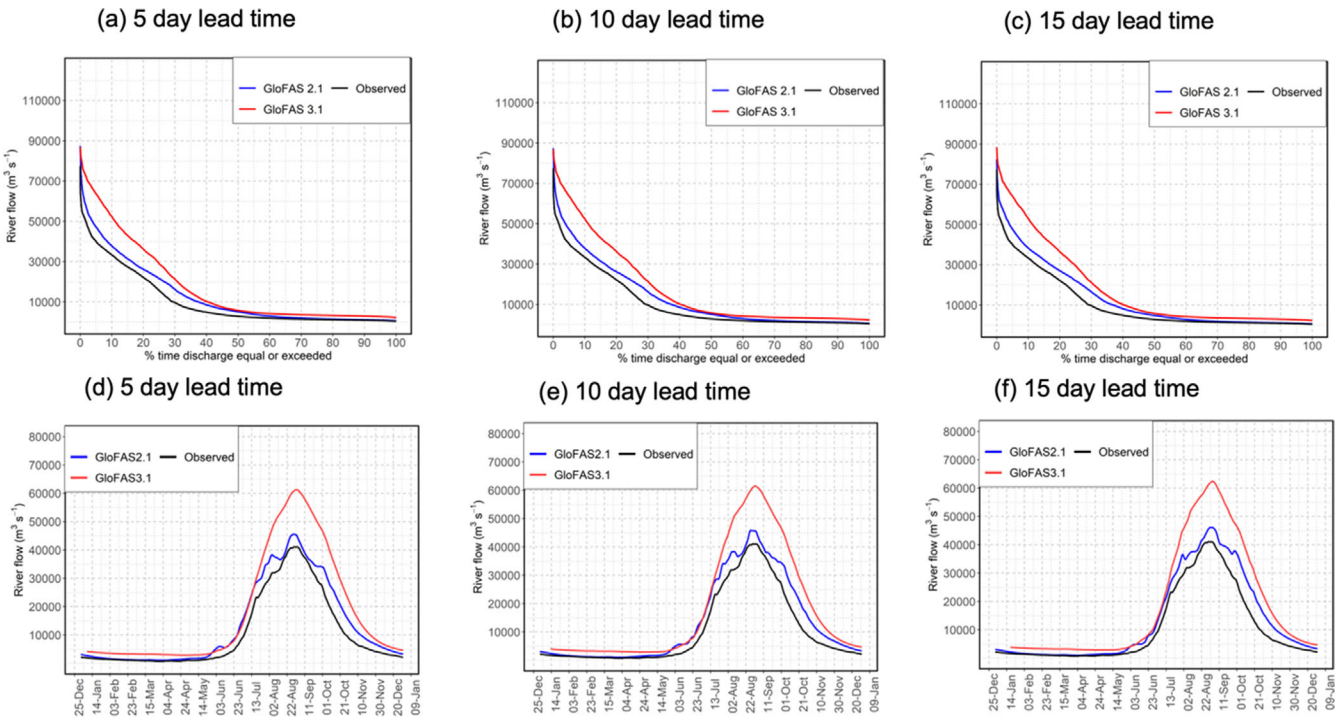

FIGURE 7 Flow duration curves of GloFAS reforecasts v2.1 and v3.1 and observed river flows for the Ganges River at: (a) 5-day lead-time, (b) 10-day lead-time and (c) 15-day lead-time. Annual cycle of observed flow and GloFAS reforecasts v2.1 and v3.1 at: (d) 5-day lead-time, (e) 10-day lead-time and (d) 15-day lead-time, based on the long-term mean over 20 years (1999–2018). Blue, red and black lines are GloFAS v2.1, v3.1 and observed respectively.

TABLE 4 Average FAR and POD for different lead-time clusters at threshold 90th percentile for the Brahmaputra River.

	Lead-time—not corrected v2.1		Lead-time corrected v2.1		Lead-time—not corrected v3.1		Lead-time corrected v3.1	
Leadtime (days)	**FAR**	**POD**	**FAR**	**POD**	**FAR**	**POD**	**FAR**	**POD**
1–10	0.27	0.68	0.28	0.70	0.34	0.68	0.34	0.70
1–15	0.29	0.67	0.29	0.68	0.36	0.68	0.35	0.70
11–20	0.36	0.63	0.34	0.59	0.43	0.67	0.40	0.60
21–30	0.44	0.44	0.44	0.45	0.49	0.57	0.46	0.50

Abbreviations: FAR, false alarm ratio; POD, probability of detection.

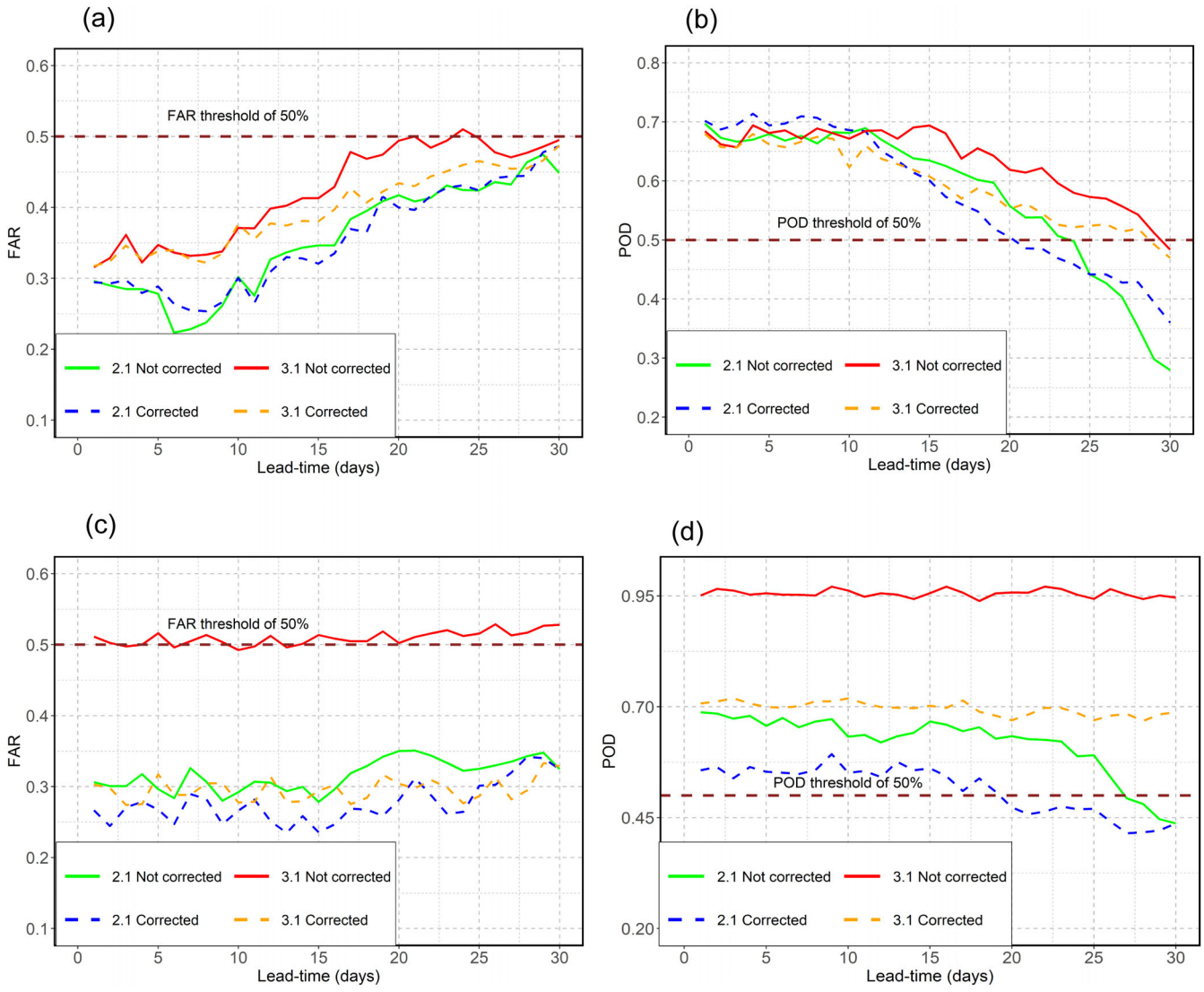

FIGURE 8 (a) FAR of GloFAS v2.1 and v3.1 reforecasts and (b) POD of GloFAS v2.1 and v3.1 reforecasts for the Brahmaputra River, (c) FAR of GloFAS v2.1 and v3.1 reforecasts and (d) POD of GloFAS v2.1 and v3.1 reforecasts for the Ganges River for lead-times of 1–30 days, with 90th percentile threshold discharge and 50% forecast trigger probability. Green and blue colour lines are for v2.1 and red and orange colours are for v3.1 for both FAR and POD. Dashed lines represent FAR and POD with lead-time dependent threshold correction, while solid lines are scores with static thresholds (not corrected). The horizontal brown dashed line shows the 50% threshold for FAR and POD. FAR, false alarm ratio; POD, probability of detection.

TABLE 5 Average FAR and POD for different lead-time clusters at threshold 90th percentile for the Ganges River.

	Lead-time—not corrected v2.1		Lead-time corrected v2.1		Lead-time—not corrected v3.1		Lead-time corrected v3.1	
Leadtime (days)	**FAR**	**POD**	**FAR**	**POD**	**FAR**	**POD**	**FAR**	**POD**
1–10	0.30	0.67	0.27	0.56	0.50	0.96	0.29	0.70
1–15	0.30	0.66	0.26	0.56	0.50	0.96	0.29	0.70
11–20	0.31	0.64	0.26	0.54	0.50	0.96	0.29	0.70
21–30	0.34	0.55	0.31	0.47	0.52	0.95	0.30	0.68

Abbreviations: FAR, false alarm ratio; POD, probability of detection.

likely to be due to the larger bias in the model at these longer lead times. The lead-time corrected models show similar skill to uncorrected for shorter lead-times, for example, 5 days as suggested by Zsoter et al. (2020).

With increasing lead times, ensemble spread increases and we found that GloFAS v2.1 has a larger spread than version 3.1 for both rivers (see Figure S4). With longer lead times, higher FAR are due to higher spread in the ensemble particularly for the Brahmaputra.

6.3 | Forecast skill for preparedness decisions

In this final section of the analysis, we present a decision-led analysis of forecast skill in the two GloFAS model versions against the criteria for each decision as presented in Table 2. In Figure 8, in cases where FAR and POD values are both located inside the shaded box then the skill is sufficient for the considered decision. The same analysis but against water level rather than discharge is provided in Figure S5. Note that for these decisions we also include evaluation against the 95th and 99th percentile river flows and water levels. Results show that using GloFAS for decisions d1 and d2, which support evacuation of people and their livestock living in areas of relatively higher land is not feasible with either v2.1 or v3.1. Similarly, the pre-activation of humanitarian action before floods at 99th percentile threshold and 15 days lead time is also not feasible with either version for both rivers (d6, Figure 9a,b). The decisions for the low-lying *char* Islands (d3 and d4) are feasible using both v2.1 and v3.1, as are the longer lead time actions of communication to government agencies and agricultural planning decisions (d7 and d8). Household level preparedness actions (d5) are feasible in v2.1, but not in v3.1 due to a slightly higher FAR than acceptable for the Brahmaputra. The results show underperformance for the Ganges, where it is found only d7 and d8 decisions are feasible in both v2.1 and v3.1. These results suggest that the Ganges is more suitable for decisions relevant to the medium thresholds floods with relatively longer margin of error.

The evaluation against water level (Figure S5) provides a more relevant assessment of whether the system can forecast impact. For v2.1 the results are similar to the evaluation against discharge (Figure S5a), but there are differences for v3.1, with d3 and d4 no longer feasible (Figure S5b) for the Brahmaputra. However, for the Ganges GloFAS shows similar decision-based performance for water level as well as discharge. The decision-maker could choose any forecast probability as a threshold provided that the FAR/POD score lies within the shaded box. For longer lead-times a larger spread in the ensemble (see Figure S4) means that the choice of the probability threshold makes a larger difference than for shorter lead times when the spread is small.

7 | DISCUSSION AND RECOMMENDATIONS

The GloFAS flood forecasting system has been changed from a coupled land surface scheme and hydrological model (v2.1) to a single fully configured global hydrological model (v3.1), that is, LISFLOOD, which is used to simulate all the hydrological processes (groundwater, rainfall-runoff processes and river routing). This study provides a new decision-led evaluation of the forecast skill of these two recent GloFAS model versions, v2.1 and v3.1, for the Brahmaputra and the Ganges Rivers in Bangladesh, based on observations for the period from 1999 to 2018. We evaluate (i) the model's capability to simulate observed river discharge, (ii) the predicting skill for flood events across forecast horizons (accounting for lead time dependent biases), and (iii) the forecasting skill for different early action decisions based on user-oriented criteria. Both GloFAS v2.1 and v3.1 perform relatively well in that they simulate the hydrological behaviour of the two large rivers, far better than simple average benchmarks. The model performance in terms of KGE shows

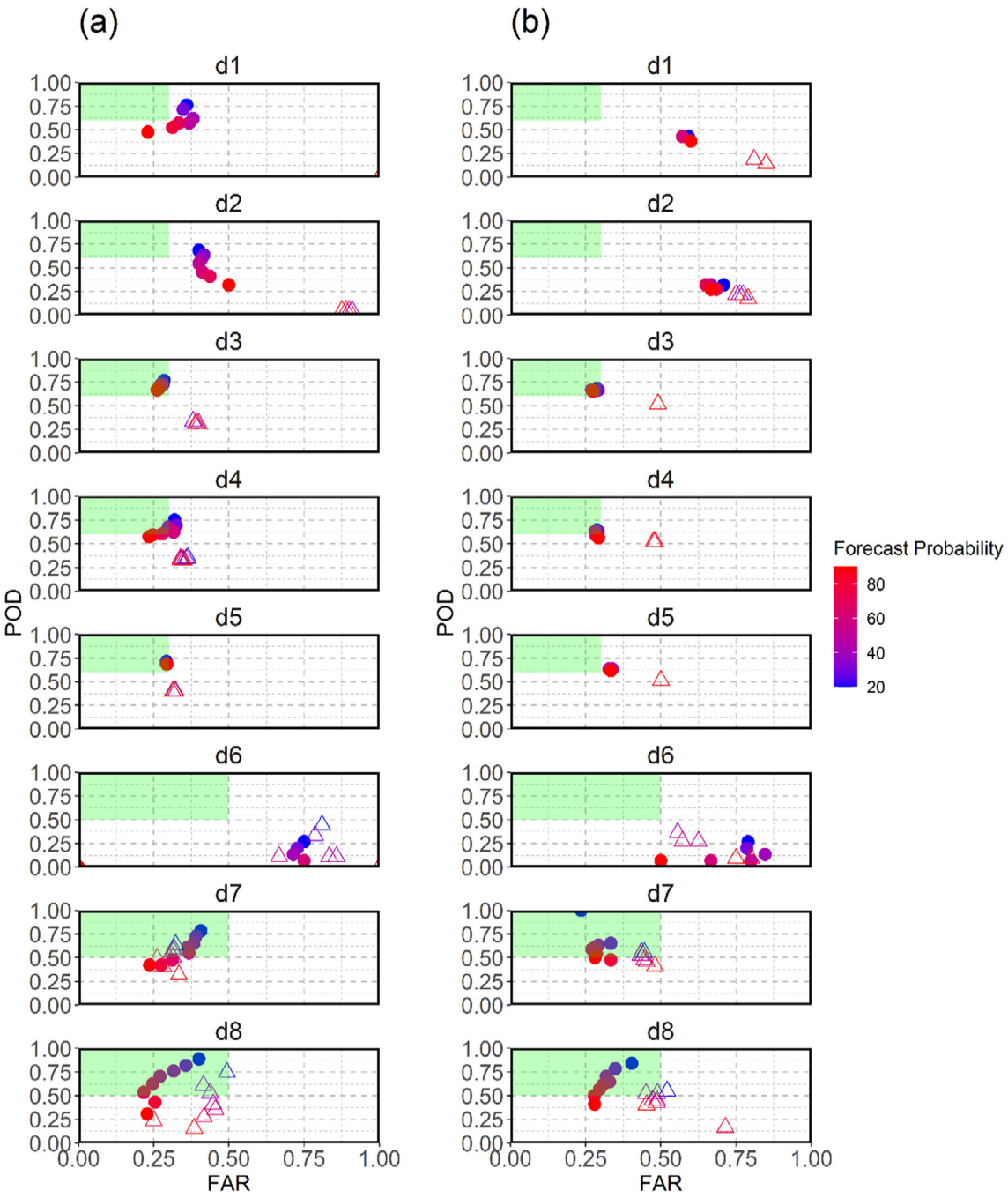

FIGURE 9 Forecast skill (FAR and POD) for different forecast probabilities (from 20% to 90%) and for each decision d1 to d8 (Table 2), based on discharge thresholds for (a) GloFAS v2.1 and (b) GloFAS v3.1 (both using lead-time dependent thresholds). Where the skill score point in the graph (FAR/POD) falls inside the shaded box, the forecast meets the requirements for each decision d1 to d8 (details of the decisions and criteria are provided in Table 2). Here, points are used for the Brahmaputra and empty triangles for the Ganges. FAR, false alarm ratio; POD, probability of detection.

values higher than 0.5 for all lead times up to 30 days for both versions for the Brahmaputra and for v2.1 for the Ganges, which can be considered acceptable following previous studies (Franco et al., 2020; Khoshchehreh et al., 2020). Though v3.1 shows low KGE (0.10 to 0.12) for the Ganges, low positive KGE values can still be associated to a relatively skilful model with respect to simple benchmarks like average conditions (Knoben et al., 2019; Towner et al., 2019). GloFAS v2.1 performs better than v3.1 with respect to all three components of KGE such as bias, variability and correlation for the Brahmaputra River while similar results were found for the Ganges apart from a slightly higher correlation at most lead times for v3.1. Analysis of the bias shows that flows are

overestimated in both v2.1 and v3.1, but biases have worsened with the model upgrade to v3.1. This is in line with the global scale model performance assessment of GloFAS which shows that v3.1 produces 43% higher river discharge than v2.1, based on the global average of all analysed catchments (GloFAS Wiki, 2021). These biases can potentially originate, at least partially, from the meteorological forcing, mainly through precipitation and marginally also temperature, humidity, wind and radiation particularly for around the first 2 weeks of the forecasts (Zsótér et al., 2016). However, in our study we noted the presence of a large systematic positive bias also in the shortest lead times (i.e., 1–3 days) of GloFAS forecasts, ranging from about 20% to 70% (of average flow) for the two rivers in Bangladesh. This bias at short lead-times is substantially larger for GloFAS 3.1 with respect to 2.1 suggesting that the hydrological model structure plays a key role in producing the bias, more than rainfall forecasts, as these have not changed between the two systems and as the bias is steady or decreases with lead time. This means that the production components of the LISFLOOD and HTESSEL (to a lesser extent) models are producing overestimated baseflow and direct runoff from precipitation, and/or underestimated evaporation from land and vegetation (including interception). The differences in bias between GloFAS version 3.1 and 2.1 are then likely to be caused by differences in treatment of runoff production processes (soil water storage, evaporation, overland runoff, etc.). Our results on the bias are in line with previous global evaluation of the GloFAS v3.1 reanalysis (Alfieri et al., 2020; ECMWF Wiki, 2021) which showed that in South Asia, including Bangladesh, the model overestimates substantially river flows, with a worse performance in terms of bias with respect to GloFAS v2.1. These previous studies have discussed that this general feature of the new model in v3.1 of generating more water than in the previous version (v2.1) could be linked to the changes in the hydrological modelling approach, with a different partitioning of the available precipitation into groundwater, evaporation and runoff. However, the exact source of increased bias across the different model components has not been identified in previous studies yet. Further work on the model structure should investigate the sources of model biases and inadequacies with a thorough analysis of GloFAS model structural components following two paths: (i) using multiple hydrologically relevant 'signature' metrics to quantify the performance in terms of catchment behavioural functions, including overall water balance and temporal redistribution (Yilmaz et al., 2008); and (ii) analysing the internal model fluxes changes across model versions, similarly to that undertaken by previous studies on other models (Ficchì et al., 2019).

Similarly, the variability component of the KGE, which measures the differences in standard deviations between simulated and observed streamflow, is higher in v3.1. However, both versions have a strong correlation ($r > 0.8$) with the observed discharge, though this is largely indicative of the model being able to simulate the annual cycle. This aligns with previous work which found that GloFAS shows higher correlation for catchment areas greater than 10,000 square kilometres, with this correlation increasing with the upstream area (Alfieri et al., 2013). The flow duration curves and annual cycle show an overestimated discharge with potential timing error on the rising limb of the flood wave, with higher bias in v3.1 than v2.1. These errors are linked with challenges for the structural development, calibration and validation of global hydrological models, as various uncertainties are associated to the model calibration and validation processes such as uncertainties in the observed streamflow measurements (due to errors, data collection process, human intervention, etc.), as well as uncertainty in the model forcing data and model parameterisation and structure (Hirpa et al., 2018) and lack of data on human influences. In particular, it is difficult to get detailed information on upstream water usages, storage and transfers, to include them in the models (and properly calibrate models) for a transboundary river basin like the Ganges, where streamflow is regulated by a number of water control structures, including dams, barrages, reservoirs, and so forth (Bharati et al., 2011; Nepal et al., 2018). Therefore, the development of global hydrological models remains challenging because of the wide variety of processes and human influences in various regions of the world with different climatic conditions (Werth & Güntner, 2010). In addition, model results vary due to different model structures and approaches, for example standalone hydrological models or coupled land surface and hydrological models. Therefore, it is recommended to study the implications of the structural changes in models considering their effects on different processes representations. However, a detailed investigation of the processes representation in the two model structures (GloFAS v2.1 and v.3.1) and its effects are beyond the scope of this paper but could be studied in the future. Similarly, improving the global runoff estimation through reducing the errors in the meteorological forecasts (such as bias reduction) is also suggested by Hirpa et al. (2018).

Through the use of a threshold-based approach, biases are accounted for when it comes to decision-making in GloFAS (i.e., by taking the 95th percentile of simulated flows and comparing to the 95th percentile of observed flows). However, variations in the timing of the flood wave will impact forecast skill scores. Our

analysis shows that GloFAS is able to predict flood events above the FFWC-defined danger level (90th percentile) for the onset of floods out to 30 days ahead, with a FAR of less than 50%. Out to around 10 days lead time the GloFAS performance is consistent with little change over lead-time in POD and FAR forecast skill scores for both v2.1 and v3.1. This is expected, since for a large river basin changes in discharge occurs at a slow rate and stream flow prediction does not change substantially for lead-time 1–10 days from a forecast of persistence (Alfieri et al., 2013).

The use of lead-time dependent thresholds improves FAR while decreasing POD, especially at longer lead times. The use of lead-time dependent thresholds is a method of accounting for the change in bias with lead-time, therefore is expected that a positive change in FAR would lead to a negative change in POD, and vice versa.

Different stakeholder decisions for early action require different flood magnitude thresholds, lead-times and criteria for acceptable forecast skill. Of the 8 decisions evaluated in this study, five were feasible when using GloFAS v2.1 and 4 for GloFAS v3.1 for the Brahmaputra whereas only two were for the Ganges. Counterintuitively, because the decisions made at longer lead times had less stringent criteria for forecast skill, these were the decisions that were feasible, and not the shorter lead time decisions such as evacuation. We acknowledge that the specific results on the relative performance of GloFAS 2.1 vs 3.1 in the Brahmaputra and Ganges may not apply to other basins, as proven by the diversity of our results for these two close, but diverse catchments in Bangladesh. However, our decision-led evaluation approach demonstrates the need for more in user-oriented local-scale forecast analysis, particularly the use of more case studies to support global flood forecasting systems developments. In other words, our results suggest that multi-criteria local analyses of model transitions are necessary, at least until resources become available to scale up these case studies to the global scale and to the different key model users sectors.

We propose the following recommendations for model developers and users to improve further development and application for better decisions.

7.1 | Model development perspective

There are clear differences between versions 2.1 and 3.1 in terms of simulated discharge as well as flood forecast skill. Though both model versions are predicting the normal flood thresholds (90th percentile) with acceptable skill, for extreme floods there are more false alarms in v3.1 which limits the usability of the forecast. Therefore, more model development effort is required to improve forecast skill for extreme floods at shorter lead-times, for example to support evacuation. The web interface of GloFAS v3.1 includes model performance (KGE) and forecast skill (CRPSS), and this evaluation is against benchmarks of persistence and climatology forecasts (ERA5 reanalysis discharge 1979–2018) (Harrigan, Zsoter, et al., 2023). This is less useful for decision-makers, and so we recommend that also other user-oriented metrics, like FAR and POD, should be added so that potential users have the opportunity to look at relevant performance indicators. To develop the confidence of users in new versions, it is necessary to test models with historical observed floods before a new version is implemented operationally. Developers of global forecast models should collaborate with national hydro-meteorological organisations and in-country partners to ensure that decisions that are being made following established protocols are still robust when using new model versions or at least the implications of model transitions are properly assessed.

7.2 | Application for improving forecast lead-time in Bangladesh

Increasing the lead-time of skilful forecasts is a challenge in a transboundary basin where major flows come from upstream catchments where there may be limited access to upstream hydro-meteorological information. Skilful global models provide a source of forecast information for such transboundary river basins. GloFAS is able to predict normal floods (90th percentile) at short to medium range (1 to 10 days) and extended range (11 to 30 days) time scale with a FAR <0.5. There is also an indication of improvement in skill by applying lead-time dependent thresholds. However, our results suggest that FFWC should use their own short-range forecast for short lead times due to the skill issue in v3.1 at the shorter lead times. Bridging between the lead-times of the FFWC short-range forecast and skilful GloFAS forecasts is required to support decisions before extreme flooding.

7.3 | Recommendation for decision makers

Decision makers must find a balance between the acceptable number of missed events and false alarms. For a global forecasting system, a new model version poses a challenge because overall improvements in skill do not necessarily mean that the same skill changes are seen in

individual river catchments. In this case, the update from v2.1 to v3.1 limits the number of decisions that can be made from GloFAS in the Brahmaputra basin, especially at shorter lead times for extreme flood events. In this case, FFWC forecasts are available at these lead-times, but working closer with GloFAS developers allows issues with the model to be dealt with more efficiently.

8 | CONCLUSION

Model upgrades take place as part of a continuous cycle of developments and updates. The upgrade to GloFAS 3.1 included the use of a stand-alone hydrological model instead of previous coupled land-surface and hydrological routing components, that is a significant change. The relevance of changes in forecast skill following this upgrade becomes clear when evaluating against specific decisions. Decision-based evaluation of a forecast model aims to look at forecast performance from the perspective of different early action decisions that are being taken. In this study the forecast skill of two versions of GloFAS is assessed by using 20 years of reforecast data, and with and without the use of lead-time dependent thresholds. The study provides a methodology to evaluate forecast skill for different early action decisions which can be applied for different river basins, each with its own decision criteria.

Although the new v3.1 has been calibrated for the Brahmaputra and the Ganges River by GloFAS developers, there is no improvement in the skill of simulated flow compared to the previous version 2.1, and even a deterioration in some components of skill. Between the two rivers, GloFAS perform better for the Brahmaputra River. GloFAS forecasts skill can vary based on basin and flood characteristics, as well as performance metrics. By assessing forecast skill against the criteria for different anticipatory action decisions, we find that, the new GloFAS version v3.1 shows acceptable skill for fewer decisions than v2.1 (no longer operational), and this means the loss of the ability to provide information that could be used to prepare households for flooding. The new version is capable to predict relatively well low to medium level floods with acceptable margin of error (3–7 days) and longer lead time (10–20 days). However, for more extreme events skill is lower and users may need to be cautious in some cases in applying anticipation actions as there is possibility of having higher FAR in the new versions. The decision-led evaluation has provided counter-intuitive results, showing acceptable skill for decisions such as aid distribution and communication which are made at longer lead times, but not for the short-lead time evacuation decisions, because of the different margins of acceptable errors for these different actions. This underlines the importance of decision-led evaluation, not only to give confidence in forecast use, but also to ensure that forecast 'upgrades' do in fact lead to gains in decision-making ability. More recently, a new GloFAS version upgrade (4.0) came out, introducing increased spatial resolution and other improvements to the model. However, we believe that our results and recommendations are still valid as the underlying main hydrological model remains LISFLOOD as in version 3.1 and its calibration and evaluation are still following a global approach based on generalist scores.

ACKNOWLEDGEMENTS

This study has been carried out as part of a PhD research (by Sazzad Hossain) supported by the UK Research and Innovation (UKRI); the Natural Environment Research Council (NERC) and the Foreign, Commonwealth and Development Office (FCDO) under the Forecasts for AnTicipatory HUManitarian action (FATHUM) project (grant number NE/P000525/1) and SHEAR Studentship Cohort programme (grant number NE/R007799/1). The authors are also grateful to the European Centre for Medium-Range Weather Forecasts (ECMWF) for providing GloFAS reforecast data and supporting the authors as visiting scientists to carry out the research work. The authors are thankful to the Bangladesh Meteorological Department and Bangladesh Water Development Board (BWDB) for providing observed river discharge and water level data for the Brahmaputra and Ganges Rivers in Bangladesh. Andrea Ficchì also acknowledges support from the AXA Research Fund Fellowship on Coastal Livelihoods.

CONFLICT OF INTEREST STATEMENT

The authors declare that they have no conflict of interest.

DATA AVAILABILITY STATEMENT

Data in the paper are available in public repositories: (a) the GloFAS reforecast archive from https://cds.climate.copernicus.eu/cdsapp#!/dataset/cems-glofas-reforecast?tab=form, and (b) the observed river discharge dataset from http://www.hydrology.bwdb.gov.bd/ under FFWC data dissemination policy.

ORCID

Sazzad Hossain https://orcid.org/0000-0003-4960-4529
Hannah L. Cloke https://orcid.org/0000-0002-1472-868X
Andrea Ficchì https://orcid.org/0000-0001-5630-7069
Linda Speight https://orcid.org/0000-0002-8700-157X
Elisabeth M. Stephens https://orcid.org/0000-0002-5439-7563

REFERENCES

Ahilan, S., O'Sullivan, J. J., Bruen, M., Brauders, N., & Healy, D. (2013). Bankfull discharge and recurrence intervals in Irish rivers. *Proceedings of the Institution of Civil Engineers: Water Management*, *166*(7), 381–393. https://doi.org/10.1680/wama.11.00078

Alfieri, L., Burek, P., Dutra, E., Krzeminski, B., Muraro, D., Thielen, J., & Pappenberger, F. (2013). GloFAS—Global ensemble streamflow forecasting and flood early warning. *Hydrology and Earth System Sciences*, *17*(3), 1161–1175. https://doi.org/10.5194/hess-17-1161-2013

Alfieri, L., Lorini, V., Hirpa, F. A., Harrigan, S., Zsoter, E., Prudhomme, C., & Salamon, P. (2020). A global streamflow reanalysis for 1980–2018. *Journal of Hydrology X*, *6*(100), 049. https://doi.org/10.1016/j.hydroa.2019.100049

Alfieri, L., Pappenberger, F., Wetterhall, F., Haiden, T., Richardson, D., & Salamon, P. (2014). Evaluation of ensemble streamflow predictions in Europe. *Journal of Hydrology*, *517*, 913–922. https://doi.org/10.1016/j.jhydrol.2014.06.035

Balsamo, G., Beljaars, A., Scipal, K., Viterbo, P., van den Hurk, B., Hirschi, M., & Betts, A. K. (2009). A revised hydrology for the ECMWF model: Verification from field site to terrestrial water storage and impact in the integrated forecast system. *Journal of Hydrometeorology*, *10*(3), 623–643. https://doi.org/10.1175/2008jhm1068.1

Bartholmes, J. C., Thielen, J., Ramos, M. H., & Gentilini, S. (2009). The European flood alert system EFAS—Part 2: Statistical skill assessment of probabilistic and deterministic operational forecasts. *Hydrology and Earth System Sciences*, *13*(2), 141–153. https://doi.org/10.5194/hess-13-141-2009

Beven, K. J. (2000). Uniqueness of place and process representations in hydrological modelling. *Hydrology and Earth System Sciences*, *4*(2), 203–213. https://doi.org/10.5194/hess-4-203-2000

Bhattachaiyya, N. N., & Bora, A. K. (1997). Floods of the Brahmaputra River in India. *Water International*, *22*(4), 222–229. https://doi.org/10.1080/02508069708686709

Bharati, L., Lacombe, G., Gurung, P., Jayakody, P., Hoanh, C. T., & Smakhtin, V. (2011). *The impacts of water infrastructure and climate change on the hydrology of the Upper Ganges River Basin; IWMI Research Report 142* (p. 36). International Water Management Institute. https://doi.org/10.5337/2011.210

Bora, A. (2004). Fluvial geomorphology. In V. P. Singh, N. Sharma, & C. S. Ojha (Eds.), *The Brahmaputra Basin Water Resources* (pp. 88–112). Springer.

Bischiniotis, K., van den Hurk, B., Coughlan de Perez, E., Veldkamp, T., Nobre, G. G., & Aerts, J. (2019). Assessing time, cost and quality trade-offs in forecast-based action for floods. *International Journal of Disaster Risk Reduction*, *40*(101), 252. https://doi.org/10.1016/j.ijdrr.2019.101252

Bischiniotis, K., van den Hurk, B., Zsoter, E., Coughlan de Perez, E., Grillakis, M., & Aerts, J. C. J. H. (2019). Evaluation of a global ensemble flood prediction system in Peru. *Hydrological Sciences Journal*, *64*(10), 1171–1189. https://doi.org/10.1080/02626667.2019.1617868

BWDB. (2017). Daily river discharge of the Brahmnapura river (internal database). In *Daily*. BWBD.

Castaneda-Gonzalez, M., Poulin, A., Romero-Lopez, R., Arsenault, R., Brissette, F., Chaumont, D., & Paquin, D. (2018). Impacts of regional climate model spatial resolution on summer flood simulation. *EPiC Series in Engineering*, *3*, 372–380.

Cloke, H. L., & Pappenberger, F. (2009). Ensemble flood forecasting: A review. *Journal of Hydrology*, *375*(3), 613–626. https://doi.org/10.1016/j.jhydrol.2009.06.005

Cloke, H. L., Pappenberger, F., Smith, P. J., & Wetterhall, F. (2017). How do I know if I've improved my continental scale flood early warning system? *Environmental Research Letters*, *12*(4), 044006. https://doi.org/10.1088/1748-9326/aa625a

Coughlan de Perez, E., van den Hurk, B., van Aalst, M. K., Amuron, I., Bamanya, D., Hauser, T., Jongma, B., Lopez, A., Mason, S., Mendler de Suarez, J., Pappenberger, F., Rueth, A., Stephens, E., Suarez, P., Wagemaker, J., & Zsoter, E. (2016). Action-based flood forecasting for triggering humanitarian action. *Hydrology and Earth System Sciences*, *20*(9), 3549–3560. https://doi.org/10.5194/hess-20-3549-2016

Emerton, R., Cloke, H., Ficchi, A., Hawker, L., de Wit, S., Speight, L., Prudhomme, C., Rundell, P., West, R., Neal, J., Cuna, J., Harrigan, S., Titley, H., Magnusson, L., Pappenberger, F., Klingaman, N., & Stephens, E. (2020). Emergency flood bulletins for cyclones Idai and Kenneth: A critical evaluation of the use of global flood forecasts for international humanitarian preparedness and response. *International Journal of Disaster Risk Reduction*, *50*(101), 811. https://doi.org/10.1016/j.ijdrr.2020.101811

Emerton, R. E., Stephens, E. M., Pappenberger, F., Pagano, T. C., Weerts, A. H., Wood, A. W., Salamon, P., Brown, J. D., Hjerdt, N., Donnelly, C., Baugh, C. A., & Cloke, H. L. (2016). Continental and global scale flood forecasting systems. *WIREs Water*, *3*(3), 391–418. https://doi.org/10.1002/wat2.1137

Frenken, K. (2012). *Irrigation in southern and eastern Asia in figures AQUASTAT survey–2011*. Food and Agriculture Organization of the United Nations.

Ficchì, A., Perrin, C., & Andréassian, V. (2019). Hydrological modelling at multiple sub-daily time steps: Model improvement via flux-matching. *Journal of Hydrology*, *575*, 1308–1327. https://doi.org/10.1016/j.jhydrol.2019.05.084

Franco, A. C. L., Oliveira, D. Y. D., & Bonumá, N. B. (2020). Comparison of single-site, multi-site and multi-variable SWAT calibration strategies. *Hydrological Sciences Journal*, *65*(14), 2376–2389. https://doi.org/10.1080/02626667.2020.1810252

GloFAS. (2021). Medium-range flood forecasts (GloFAS). https://www.globalfloods.eu/technical-information/glofas-30, day/

GloFAS Wiki. (2021). Global Flood Awareness System. https://confluence.ecmwf.int/display/COPSRV/GloFAS+v3.1

Gupta, H. V., Kling, H., Yilmaz, K. K., & Martinez, G. F. (2009). Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. *Journal of Hydrology*, *377*(1), 80–91. https://doi.org/10.1016/j.jhydrol.2009.08.003

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 System Science Data*, *12*(3), 2043–2060. https://doi.org/10.5194/essd-12-2043-2020

Harrigan, S., Zsoter, E., Cloke, H., Salamon, P., & Prudhomme, C. (2023). Daily ensemble river discharge reforecasts and real-time forecasts from the operational Global Flood Awareness System. *Hydrology and Earth System Sciences*, *27*(1), 1–19. https://doi.org/10.5194/hess-27-1-2023

Hirpa, F. A., Salamon, P., Beck, H. E., Lorini, V., Alfieri, L., Zsoter, E., & Dadson, S. J. (2018). Calibration of the global flood awareness system (GloFAS) using daily streamflow data.

Journal of Hydrology, *566*, 595–606. https://doi.org/10.1016/j.jhydrol.2018.09.052

Hossain, S., Cloke, H. L., Ficchì, A., Turner, A. G., & Stephens, E. (2019). Hydrometeorological drivers of the 2017 flood in the Brahmaputra basin in Bangladesh. *Hydrology and Earth System Sciences*. https://doi.org/10.5194/hess-2019-286

Hossain, S., Cloke, H. L., Ficchì, A., Turner, A. G., & Stephens, E. M. (2021). Hydrometeorological drivers of flood characteristics in the Brahmaputra river basin in Bangladesh. *Hydrology and Earth System Sciences Discussions*, *2021*, 1–28. https://doi.org/10.5194/hess-2021-97

Hossain, S. (2023). Hydro meteoroglocal drivers and extended range flood forecasting for the Brahmaputra River Basin in Bangladesh [Unpublished PhD Thesis]. University of Reading.

Immerzeel, W. (2008). Historical trends and future predictions of climate variability in the Brahmaputra basin. *International Journal of Climatology*, *28*(2), 243–254. https://doi.org/10.1002/joc.1528

Islam, A. S., Haque, A., & Bala, S. K. (2010). Hydrologic characteristics of floods in Ganges–Brahmaputra–Meghna (GBM) delta. *Natural Hazards*, *54*(3), 797–811. https://doi.org/10.1007/s11069-010-9504-y

Jain, S. K., Mani, P., Jain, S. K., Prakash, P., Singh, V. P., Tullos, D., Kumar, S., Agarwal, S. P., & Dimri, A. P. (2018). A brief review of flood forecasting techniques and their applications. *International Journal of River Basin Management*, *16*(3), 329–344. https://doi.org/10.1080/15715124.2017.1411920

Khan, M. R., Voss, C. I., Yu, W., & Michael, H. A. (2014). Water resources Management in the Ganges Basin: A comparison of three strategies for conjunctive use of groundwater and surface water. *Water Resources Management*, *28*(5), 1235–1250. https://doi.org/10.1007/s11269-014-0537-y

Kling, H., Fuchs, M., & Paulin, M. (2012). Runoff conditions in the upper Danube basin under an ensemble of climate change scenarios. *Journal of Hydrology*, *424–425*, 264–277. https://doi.org/10.1016/j.jhydrol.2012.01.011

Khoshchehreh, M., Ghomeshi, M., & Shahbazi, A. (2020). Hydrological evaluation of global gridded precipitation datasets in a heterogeneous and data-scarce basin in Iran. *Journal of Earth System Science*, *129*(1), 201. https://doi.org/10.1007/s12040-020-01462-5

Knoben, W. J. M., Freer, J. E., & Woods, R. A. (2019). Technical note: Inherent benchmark or not? Comparing Nash–Sutcliffe and Kling–Gupta efficiency scores. *Hydrology and Earth System Sciences*, *23*(10), 4323–4331. https://doi.org/10.5194/hess-23-4323-2019

Koskinen, M., Tahvanainen, T., Sarkkola, S., Menberu, M. W., Laurén, A., Sallantaus, T., Marttila, H., Ronkanen, A.-K., Parviainen, M., Tolvanen, A., Koivusalo, H., & Nieminen, M. (2017). Restoration of nutrient-rich forestry-drained peatlands poses a risk for high exports of dissolved organic carbon, nitrogen, and phosphorus. *Science of the Total Environment*, *586*, 858–869. https://doi.org/10.1016/j.scitotenv.2017.02.065

Legates, D. R., & McCabe, G. J., Jr. (1999). Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation. *Water Resources Research*, *35*(1), 233–241. https://doi.org/10.1029/1998WR900018

Lin, P., Pan, M., Beck, H. E., Yang, Y., Yamazaki, D., Frasson, R., David, C. H., Durand, M., Pavelsky, T. M., Allen, G. H., Gleason, C. J., & Wood, E. F. (2019). Global reconstruction of naturalized river flows at 2.94 million reaches. *Water Resources Research*, *55*(8), 6499–6516. https://doi.org/10.1029/2019WR025287

Liu, D. (2020). A rational performance criterion for hydrological model. *Journal of Hydrology*, *590*(125), 488. https://doi.org/10.1016/j.jhydrol.2020.125488

Lopez, A., Coughlan de Perez, E., Bazo, J., Suarez, P., van den Hurk, B., & van Aalst, M. (2020). Bridging forecast verification and humanitarian decisions: A valuation approach for setting up action-oriented early warnings. *Weather and Climate Extremes*, *27*(100), 167. https://doi.org/10.1016/j.wace.2018.03.006

Mirza, M. M. Q. (2003). Three recent extreme floods in Bangladesh: A hydro-meteorological analysis. *Natural Hazards*, *28*(1), 35–64. https://doi.org/10.1023/A:102116973

Mirza, M. Q., Warrick, R. A., Ericksen, N. J., & Kenny, G. J. (1998). Trends and persistence in precipitation in the Ganges, Brahmaputra and Meghna river basins. *Hydrological Sciences Journal*, *43*(6), 845–858. https://doi.org/10.1080/02626669809492182

Moriasi, D. N., Arnold, J. G., Van Liew, M. W., Bingner, R. L., Harmel, R. D., & Veith, T. L. (2007). Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. *Transactions of the ASABE*, *50*(3), 885–900.

Nepal, S., Pandey, A., Shrestha, A. B., & Mukherji, A. (2018). *Revisiting key questions regarding upstream–downstream linkages of land and water management in the Hindu Kush Himalaya (HKH) region. HI-AWARE Working Paper 21.* HI-AWAR. https://lib.icimod.org/record/34354

Pappenberger, F., Cloke, H. L., Balsamo, G., Ngo-Duc, T., & Oki, T. (2010). Global runoff routing with the hydrological component of the ECMWF NWP system. *International Journal of Climatology*, *30*(14), 2155–2174. https://doi.org/10.1002/joc.2028

Passerotti, G., Massazza, G., Pezzoli, A., Bigi, V., Zsótér, E., & Rosso, M. (2020). Hydrological model application in the Sirba River: Early warning system and GloFAS improvements. *Water*, *12*(3), 620. https://www.mdpi.com/2073–4441/12/3/620

Paura, P. K. (2004). Flood Management in Ganga-Brahmaputra-Meghna Basin: Some aspects of regional cooperation. Workshop on flood and drought management, New Delhi.

Rajmohan, N., & Prathapar, S. (2013). Hydrogeology of the eastern Ganges Basin: An overview.

Schönfelder, L., Bakken, T., Alfredsen, K., & Adera, A. (2017). Application of HYPE in Norway, Assessment of the hydrological model HYPE as a tool to support the implementation of EU Water Framework Directive in Norway.

Siddique, Q. I., & Chowdhury, M. M. H. (2000). Flood '98: Losses and damages. In A. K. A. C. Q. K. Ahmed, S. H. Imam, & M. Sarker (Eds.), *Perspectives on flood 1998* (pp. 1–13). University Press Limited.

Thielen, J., Alfieri, L., Burek, P., Kalas, M., Salamon, P., Thiemig, V., de Roo, A., Muraro, D., Pappenberger, F., & Dutra, E. (2012). In F. Klijn & T. Schweckendiek (Eds.), *The global flood awareness system (GloFAS). Comprehensive flood risk management*. Taylor and Francis, CRC Press.

Thielen, J., Bartholmes, J., Ramos, M. H., & de Roo, A. (2009). The European flood alert system—Part 1: Concept and development. *Hydrology and Earth System Sciences*, *13*(2), 125–140. https://doi.org/10.5194/hess-13-125-2009

Thiemig, V., Bisselink, B., Pappenberger, F., & Thielen, J. (2015). A pan-African medium-range ensemble flood forecast system. *Hydrology and Earth System Sciences*, *19*(8), 3365–3385. https://doi.org/10.5194/hess-19-3365-2015

Towner, J., Cloke, H. L., Zsoter, E., Flamig, Z., Hoch, J. M., Bazo, J., Coughlan de Perez, E., & Stephens, E. M. (2019). Assessing the performance of global hydrological models for capturing peak river flows in the Amazon basin. *Hydrology and Earth System Sciences*, *23*(7), 3057–3080.

Van Der Knijff, J. M., Younis, J., & 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. https://doi.org/10.1080/13658810802549154

Weigel, A. P. (2011). Ensemble forecasts. In *Forecast Verification: A Practitioner's Guide in Atmospheric Science* (pp. 141–166). John Wiley and Sons.

Werth, S., & Güntner, A. (2010). Calibration analysis for water storage variability of the global hydrological model WGHM. *Hydrology and Earth System Sciences*, *14*(1), 59–78. https://doi.org/10.5194/hess-14-59-2010

Wilks, D. S. (2011). Forecast verification. In *International geophysics* (Vol. 100, pp. 301–394). Elsevier.

WMO. (2012). Guidelines on ensemble prediction systems and forecasting. *World Meteorological Organization Weather Climate and Water*, *1091*, 1–32.

Wu, B., Wang, G., Xia, J., Fu, X., & Zhang, Y. (2008). Response of bankfull discharge to discharge and sediment load in the Lower Yellow River. *Geomorphology*, *100*(3), 366–376. https://doi.org/10.1016/j.geomorph.2008.01.007

Yilmaz, K. K., Gupta, H. V., & Wagener, T. (2008). A process-based diagnostic approach to model evaluation: Application to the NWS distributed hydrologic model. *Water Resources Research*, *44*(9), 1–18. https://doi.org/10.1029/2007WR006716

Zsótér, E., Pappenberger, F., Smith, P., Emerton, R. E., Dutra, E., Wetterhall, F., Richardson, D., Bogner, K., & Balsamo, G. (2016). Building a multimodel flood prediction system with the TIGGE archive. *Journal of Hydrometeorology*, *17*(11), 2923–2940.

Zsoter, E., Prudhomme, C., Stephens, E., Pappenberger, F., & Cloke, H. (2020). Using ensemble reforecasts to generate flood thresholds for improved global flood forecasting. *Journal of Flood Risk Management*, *13*(4), e12658. https://doi.org/10.1111/jfr3.12658

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Hossain, S., Cloke, H. L., Ficchì, A., Gupta, H., Speight, L., Hassan, A., & Stephens, E. M. (2023). A decision-led evaluation approach for flood forecasting system developments: An application to the Global Flood Awareness System in Bangladesh. *Journal of Flood Risk Management*, e12959. https://doi.org/10.1111/jfr3.12959