https://doi.org/10.5194/hess-2022-53
Preprint. Discussion started: 8 February 2022
© Author(s) 2022. CC BY 4.0 License.

Hydrology and Earth System Sciences
Discussions

Using a long short-term memory (LSTM) neural network to boost river streamflow forecasts over the western United States

Kieran M. R. Hunt[1,2], Gwyneth R. Matthews[1], Florian Pappenberger[3], and Christel Prudhomme[3,4,5]

[1]Department of Meteorology, University of Reading, UK
[2]National Centre for Atmospheric Sciences, University of Reading, UK
[3]European Centre for Medium-Range Weather Forecasts, Reading, UK
[4]Department of Geography and Environment, Loughborough University, UK
[5]UK Centre for Ecology and Hydrology, Wallingford, UK

Correspondence: Kieran M. R. Hunt (k.m.r.hunt@reading.ac.uk)

Abstract. Accurate river streamflow forecasts are a vital tool in the fields of water security, flood preparation and agriculture, as well as in industry more generally. Over the last century, physics-based models traditionally used to produce streamflow forecasts have become increasingly sophisticated, with forecasts improving accordingly. However, the development of such models is often bound by two soft limits: empiricism – many physical relationships are represented by computationally efficient empirical formulae; and data sparsity – the calibration of these models often requires long time-series of high-resolution observational data at both surface and subsurface levels.

Artificial neural networks have previously been shown to be highly effective at simulating nonlinear systems where knowledge of the underlying physical relationships is incomplete. However, they also suffer from issues related to data sparsity. Recently, hybrid forecasting systems, which combine the traditional physics-based approach with statistical forecasting techniques, have been investigated for use in hydrological applications. In this study, we test the efficacy of a type of neural network, the long-short term memory (LSTM), at predicting streamflow at ten river gauge stations across various climatic regions of the western United States. The LSTM is trained on the catchment-mean meteorological and hydrological variables from the ERA5 and GloFAS-ERA5 reanalysis as well as historical streamflow observations. The performance of these hybrid forecasts is evaluated and compared to the performance of both raw and bias-corrected output from the Copernicus Emergency Management Service (CEMS) physics-based Global Flood Awareness System (GloFAS).

Two periods are considered, a testing phase (June 2019 to June 2020), during which the models were fed with ERA5 data to investigate how well they simulated streamflow at the ten stations; and an operational phase (September 2020 to October 2021), during which the models were fed forecast variables from ECMWF's Integrated Forecast System (IFS), to investigate how well they could predict streamflow at lead times of up to ten days.

All three models performed well in the testing phase, with the LSTM performing the best (skilful at nine stations, of which six were highly skilful). Similarly, the LSTM forecasts beat the raw and bias-corrected GloFAS forecasts during the operational phase, with skilful 5-day forecasts at nine stations, of which five were highly skilful. Implications and potential improvements to this work are discussed. In summary, this is the first time an LSTM has been used in a hybrid system to create a medium-

https://doi.org/10.5194/hess-2022-53
Preprint. Discussion started: 8 February 2022
© Author(s) 2022. CC BY 4.0 License.

Hydrology and Earth System Sciences
Discussions

range streamflow forecast, and in beating established physics-based models, shows promise for the future of neural networks in hydrological forecasting.

1 Introduction

Accurate forecasts of river streamflow are vital across a range of sectors, including, but not limited to, agriculture, water security, recreation, disaster management and heavy industry. As such, modelling streamflow as a function of observable hydrological and meteorological variables has been the subject of focused study for nearly 200 years, and has intensified considerably over the past few decades as demands on water resources continue to increase dramatically (Beven, 2011).

The earliest attempt (Mulvaney, 1851) comprised a simple linear relationship between streamflow and catchment rainfall, derived using linear regression. Key early developments then split the catchment into regions based on estimated travel time to the gauge (Imbeaux, 1892; Ross, 1921) and included more variables in the regression model (Linsley et al., 1949). One of the first physics-based streamflow models was developed by Horton (1933), who considered the role of excess soil filtration in runoff. Since then, physics-based models have largely dominated, particularly with continued improvements to process understanding (e.g. Freeze and Harlan, 1969), computing power (e.g. Kollet et al., 2010; Schiemann et al., 2018), observation systems – including discharge data vital for model calibration (e.g. Newman et al., 2015), and remote sensing (e.g. Huffman et al., 1995; Robock et al., 2000).

Physics-based models are still currently limited by lack of process understanding, lack of – particularly subsurface – data, and inadequate grid resolution (Wood et al., 2011). Such problems can be overcome by the application of artificial neural networks, which can produce highly accurate simulations of physical systems even if the underlying physical relationships are not known. In hydrology, a particular type of artificial neural network, known as a Long Short-Term Memory network (LSTM; Schmidhuber et al., 1997; Hochreiter and Schmidhuber, 1997; Gers et al., 2000), has become increasingly popular due to its ability to process sequential data (Shen and Lawson, 2021).

LSTMs are a special case of so-called recurrent neural networks (RNNs), i.e., artificial neural networks that are capable of processing data containing temporal sequences. The basic feature of RNNs is a feedback loop that allows the network to retain information over time. However, due to their relatively simple construction, RNNs do not readily retain long-term temporal dependencies (Chung et al., 2014). This is overcome in LSTMs by the addition of a special unit, known as a memory cell or the forget gate, that can preserve information indefinitely, allowing LSTMs to learn long-term dependencies that other RNNs cannot. In modelling river discharge over the United States, Kratzert et al. (2018) showed that LSTMs vastly outperform conventional RNNs.

For the reasons outlined above, it is unsurprising that LSTMs are being increasingly used to explore complicated hydrological problems including modelling river streamflow. In this regard, studies fall into two categories – either seeking to create a model capable of replicating existing streamflow observations, or seeking to create a model capable of forecasting streamflow at some future time. Several highly illustrative studies approach the former topic. In particular, a series of papers published by researchers at Johannes Kepler University Linz (Kratzert et al., 2018, 2019a, b; Klotz et al., 2021) demonstrated the remarkable

https://doi.org/10.5194/hess-2022-53
Preprint. Discussion started: 8 February 2022
© Author(s) 2022. CC BY 4.0 License.

Hydrology and Earth System Sciences
Discussions

ability of LSTMs to simulate daily streamflow in catchments across the United States, even if the models were trained on multiple basins at once, and showed how these results allowed for the modelling of ungauged basins and quantification of streamflow uncertainty. Extending this work, Gauch et al. (2021) used a reanalysis framework to demonstrate the predictive power of LSTMs in streamflow modelling.

The latter topic has proven more challenging, with only a handful of studies trying (to the authors' knowledge) to use LSTMs to predict streamflow (Slater et al., 2021). The most basic of these rely on rivers where streamflow has a strong annual cycle and large lagged autocorrelation (i.e. high persistence), using only antecedent streamflow data from the same site. Such studies have mixed, although promising results (de Melo et al., 2019; Sahoo et al., 2019; Sudriani et al., 2019; Zhu et al., 2020). More advanced models also incorporate upstream data, either just streamflow at different sites (Silva et al., 2021), or additionally precipitation (Le et al., 2019; Hu et al., 2020) with improved results. Le et al. (2019), for a case study in Vietnam, and Silva et al. (2021), for a case study in Brazil, achieved good results at 3- and 5-day lead times respectively. Ding et al. (2019) produced perhaps the most sophisticated LSTM-based hydrological forecast model to date. A hybrid forecast that ingested ECMWF forecasts of precipitation, soil moisture, and other variables to produce a runoff forecast around the confluence region of the Lech and Danube Rivers. They verified forecasts up to lead times of nine hours, finding a Nash-Sutcliffe efficiency of 0.71, rising to 0.77 after the inclusion of an attention mechanism.

In this study, we seek to expand on these previous efforts and develop an LSTM capable of providing skilful river streamflow forecasts at lead times of up to ten days at ten stations across the western United States. We will train the model mainly using meteorological variables from the ERA5 reanalysis (Sec. 3.1) and hydrological variables from GloFAS-ERA5 (Sec. 3.3, meaning that we are not at the mercy of potentially sparse observational data, but will use official gauge observations (Sec. 2) as the target to give optimal calibration. Once trained, the LSTM will be used to produce forecasts by replacing ERA5 inputs with forecast variables from the ECMWF Integrated Forecast System (IFS; Sec. 3.2). Again the use of the IFS means that the LSTM forecasts are not vulnerable to data latency of observations in an operational setting. Additionally, since the IFS is used to drive the ERA5 reanalysis, any significant climatological biases in one are likely to be present in the other. Since we are training the model with ERA5 data, such biases – so long as they are consistent between the two products – will be mitigated as the LSTM either applies an internal bias correction, or gives the field a low weighting in the input layer. We train the model on publicly-available ERA5, rather than IFS hindcasts, so that our methods can be completely reproduced by any interested reader. The forecasts will be made under operational time and data constraints for a thirteen month period (September 2020 to October 2021) and the results compared with GloFAS (Sec. 3.3), a physics-based streamflow forecast produced by ECMWF; a new, bias-corrected version of GloFAS (Sec. 4.2; and a simple persistence model. The core aims of this study are to determine (a) whether such an LSTM based hybrid system can provide skilful streamflow forecasts, (b) whether a hybrid system can perform better than existing state-of-the-art physics-based systems, and (c) whether advanced bias-correction techniques can improve the skill of physics-based models.

The study is laid out as follows: we discuss the study region and the climatological characteristics of the ten gauge stations in Sec. 2. We then describe the data used in Sec. 3 and methods – including the bias-correction algorithm and the LSTM setup – in Sec. 4. The results section is split into two parts, verification of a testing phase – where the models are driven with ERA5

https://doi.org/10.5194/hess-2022-53
Preprint. Discussion started: 8 February 2022
© Author(s) 2022. CC BY 4.0 License.

Hydrology and Earth System Sciences
Discussions
Open Access
EGU

– in Sec. 5.1, and verification of the operational phase – where the models are driven with IFS output – in Sec. 5.2. Finally, we discuss potential applications and improvements to our work in Sec. 6 and conclude with a summary in Sec. 7.

2 Study region and choice of stations

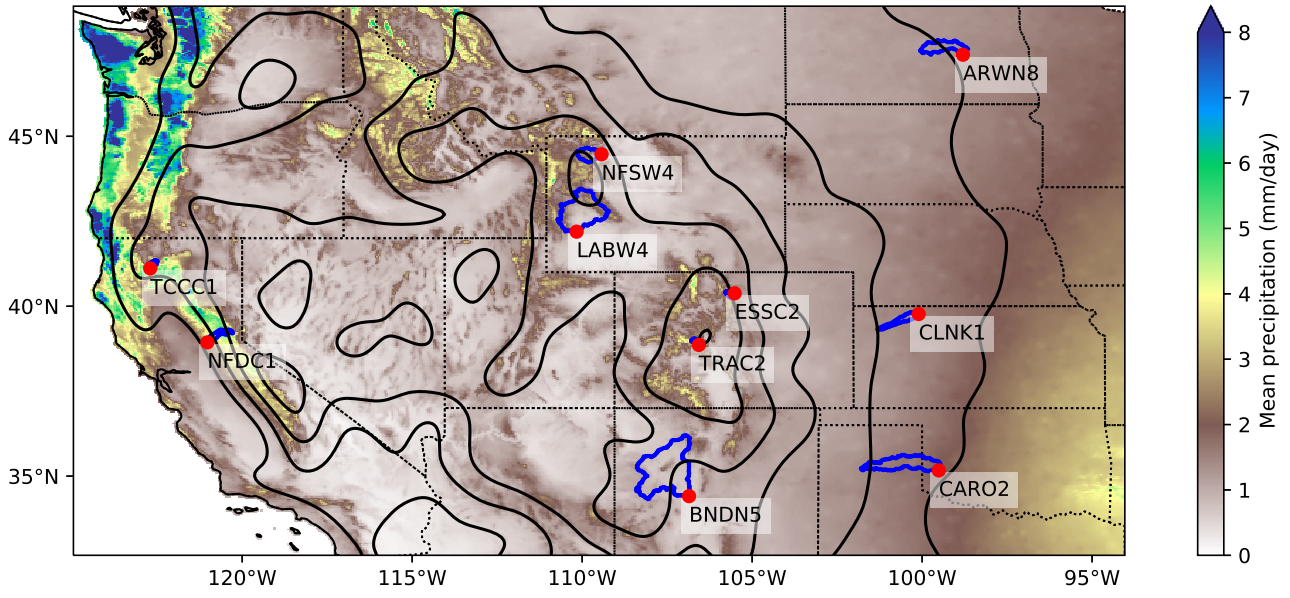


Figure 1. Locations of the ten streamflow gauges (red) and their catchment basins (dark blue). Overlaid are the climatological precipitation (filled contours, 1981–2010, from PRISM), smoothed orography (black line contours at 500 m intervals) from ETOPO, and state boundaries (dotted black lines).

The choice of study region and stations was dictated by the US Department of Reclamation, part of whose remit is water security over the western half of the contiguous United States. Between September 2020 and September 2021, they sponsored a competition (https://www.topcoder.com/community/streamflow) to predict streamflow at ten locations (Fig. 1 and Tab. 1) at lead times of up to ten days, into which we entered three forecasts, raw and bias-corrected GloFAS and an LSTM, which are discussed in greater detail in Sec. 4, and which form the basis of this study.

The ten gauge locations are shown in the context of climatological precipitation (from PRISM; Daly et al., 2008) and topography (from ETOPO1 Amante and Eakins, 2009) in Fig. 1, along with their respective catchment basins in blue. The ten locations represent a considerable diversity in climate and environment: two stations (TCCC1, NFDC1) are on the windward side of the Sierra Nevada, a region of high climatological rainfall; four (NFSW4, LABW4, ESSC2, TRAC2) are situated at high elevations along the Rockies; and four (ARWN8, CLNK1, CARO2, BNDN5) are in the relatively arid plains of the Midwest and south. These locations are summarised in Tab. 1.

https://doi.org/10.5194/hess-2022-53
Preprint. Discussion started: 8 February 2022
© Author(s) 2022. CC BY 4.0 License.

Hydrology and Earth System Sciences
Discussions
Open Access
EGU

Location ID	USGS Station Number	Description	Longitude	Latitude
BNDN5	8353000	Rio Puerco near Bernardo, NM	-106.85	34.41
ARWN8	6468250	James River above Arrowwood Lake near Kensal, ND	-98.80	47.40
TCCC1	11523200	Trinity River above Coffee Creek near Trinity Center, CA	-122.70	41.11
CARO2	7301500	North Fork Red River near Carter, OK	-99.51	35.17
ESSC2	6733000	Big Thompson River above Lake Estes, CO	-105.51	40.38
NFDC1	11427000	North Fork American River at North Fork Dam, CA	-121.02	38.94
LABW4	9209400	Green River near La Barge, WY	-110.16	42.19
CLNK1	6847900	Prairie Dog Creek above Keith Sebilius Lake, KS	-100.10	39.77
TRAC2	9107000	Taylor River above Taylor Park, CO	-106.57	38.86
NFSW4	6279940	North Fork Shoshone River at Wapiti, WY	-109.43	44.47

Table 1. Summary of the locations of the ten river streamflow gauges used in this study. See also Fig. 1.

Station ID	Basin area (km^2)	Elevation (m)	Discharge ($m^3\ s^{-1}$)				Missing (%)
			25%	median	mean	75%	
BNDN5	13739	1439	0.000	0.000	0.368	0.042	23.8
ARWN8	1165	439	0.195	1.44	8.24	9.34	50.6
TCCC1	386	773	12.1	17.2	20.5	24.1	1.6
CARO2	5367	508	0.821	1.95	3.28	3.82	9.0
ESSC2	357	2290	0.680	1.76	4.47	5.30	22.2
NFDC1	885	217	24.2	37.4	49.0	53.2	4.1
LABW4	10123	1987	18.3	21.2	26.0	28.6	37.1
CLNK1	1527	712	0.059	0.122	0.142	0.187	14.3
TRAC2	331	2847	0.963	1.05	1.19	1.39	33.3
NFSW4	1810	1700	4.76	6.17	8.55	10.1	18.5

Table 2. Summary of the climatological hydrological parameters of the ten river streamflow gauges used in this study. 'Missing' indicates the fraction of measurements since 01 Jan 1990 recorded either as 'Ice' or some other non-numeric value. Data from USGS and DWR, as outlined in Sec. 3.4.

Selected hydrological statistics are shown for each gauge in Tab. 2. These are computed over the entire measurement record for each gauge (minimum ∼30 years) and reflect the variance in river characteristics chosen by the Bureau of Reclamation. Notably, only three stations (BNDN5, CARO2, LABW4) have a drainage basin whose area exceeds 2000 km^2; this is the recommended threshold for GloFAS analysis, as basins smaller than this are not necessarily resolved by the underlying model. The remaining seven stations, therefore, provide us with an interesting forecast challenge. As Tab. 2 shows, some gauges are in extremely arid locations (e.g. BNDN5), and some have a great quantity of missing data (e.g. ARWN8), both of which present

https://doi.org/10.5194/hess-2022-53
Preprint. Discussion started: 8 February 2022
© Author(s) 2022. CC BY 4.0 License.

Hydrology and Earth System Sciences
Discussions

Station ID	Mean 2-m temp (°C)			Mean precip (mm d^{-1})		
	Annual	Jan	Jul	Annual	Jan	Jul
BNDN5	10.9	-1.6	22.7	1.0	0.8	1.8
ARWN8	4.9	-12.7	20.6	2.0	0.8	3.3
TCCC1	8.0	-0.4	20.2	3.4	6.6	0.5
CARO2	15.6	3.5	27.6	1.8	0.9	1.6
ESSC2	1.9	-8.4	14.8	2.1	1.3	3.1
NFDC1	10.2	2.2	21.4	4.3	8.6	0.1
LABW4	1.6	-9.9	16.4	1.5	1.7	0.7
CLNK1	12.0	-1.0	25.8	1.9	0.6	2.8
TRAC2	-0.9	-11.4	11.7	1.7	1.7	1.7
NFSW4	0.0	-10.6	14.2	2.0	2.0	1.2

Table 3. Summary of the climatological meteorological parameters of the ten river streamflow gauges used in the competition. Data from ERA5, as outlined in Sec. 3.1.

potential difficulties in the training and operational use of the LSTM. Overall (Tab. 3), the basin-average meteorology varies considerably between the gauges, reflecting their geographical diversity.

3 Data

3.1 ERA5

The Copernicus Climate Change Service (C3S) at ECMWF produces the five-generation ERA5 atmospheric reanalyses of global climate covering the period since January 1950 (Hersbach et al., 2020). Data from ERA5 cover the entire globe on a 30 km grid and resolve the atmosphere on 137 levels from the ground up to 80 km in altitude. At reduced spatial and temporal resolutions, ERA5 includes uncertainty information for all variables. We use catchment-mean ERA5 variables (near-surface, surface, and subsurface) as training data for the LSTM.

3.2 IFS

The study uses the ECMWF Integrated Forecasting System (IFS, version CY47R1). This IFS was run at full complexity with the configuration used for operational weather forecasts at ECMWF, as well as for the re-analysis (ERA5). The system is described in detail (https://www.ecmwf.int/en/publications/ifs-documentation) and has a turbulent diffusion and exchange with the surface represented by the Monin-Obukhov similarity theory in the surface layer and an Eddy-Diffusivity Mass-Flux (EDMF) framework above the surface layer and includes a mass-flux shallow-convection; a multilayer, multitiled land-surface scheme (HTESSEL); a five-species cloud microphysics model; and a shortwave and longwave radiation scheme including cloud radiation interactions. IFS data, up to a 10-day lead time, are used as input to the LSTM when it is run operationally.

https://doi.org/10.5194/hess-2022-53
Preprint. Discussion started: 8 February 2022
© Author(s) 2022. CC BY 4.0 License.

Hydrology and Earth System Sciences
Discussions

Each morning throughout the operational period (September 2020 to October 2021), the control member of the ensemble was downloaded using the Meteorological Archival and Retrieval System (MARS) API. This comprises more than 20 variables, globally, at a six-hourly frequency and resolution of $0.1\times0.1°$. Having to download and pre-process this large volume of data within the time constraints allowed limited us to using only one ensemble member. To retain more realistic variability at longer lead times, we chose to use the control member, rather than the ensemble mean. The full list of variables used is given in Sec. 4.3.

3.3 GloFAS

The worldwide Global Flood Awareness System (GloFAS; Harrigan et al., 2020), created collaboratively by the European Commission and the European Centre for Medium-Range Weather Forecasts (ECMWF), is a global hydrological forecasting and monitoring system that is not constrained by administrative or political boundaries. It combines cutting-edge meteorological predictions with a hydrological model, and due to its continental scale setup, it can deliver information on upstream river conditions as well as continental and global overviews to downstream nations. Since 2011, GloFAS has been producing daily ensemble flood predictions and monthly seasonal streamflow outlooks since November 2017 and is run operationally as a component of the Copernicus Emergency Management Service. For dates prior to May 25 2021 (i.e. all of the testing period and most of the operational period), we use GloFAS version 2.1 (Zsoter et al., 2019a), thereafter we use GloFAS version 3.1 (Zsoter et al., 2021). GloFAS products are freely available from its dedicated Information System, open to all following registration (www.globalfloods.eu) and its hydrological data from Copernicus Climate Data Store (https://cds.climate.copernicus.eu/#!/home). More detail on GloFAS service can be found on the dedicated wiki (https://confluence.ecmwf.int/display/COPSRV/Global+Flood+Awareness+System). Two sets of GloFAS data were used: GloFAS-ERA5 (Zsoter et al., 2019b), a global modelled daily data of river discharge from the Global Flood Awareness System (GloFAS) forced by ERA5, providing a simulation as close as observation as possible; and GloFAS forecasts (Zsoter et al., 2019a, 2021), an ensemble of global daily river discharge forecasts, forced from ECMWF ensemble forecasts from the IFS. As with the IFS above, for forecasts we use the control member up to a lead time of ten days, downloaded using the MARS API. These data are global, have daily frequency, and a resolution of $0.1\times0.1°$.

3.4 Observational station data

Observational gauge data were downloaded from https://dwr.state.co.us/Tools/Stations (for ESSC2) and https://waterdata.usgs.gov/nwis (all others). These data are available at three-hourly resolution, published in near real-time, and mostly with coverage from about 1990 onwards. Coverage and streamflow data are given in Tab. 2. These data are used to train the LSTM and calibrate the parameters for both stages of the bias-correction algorithm (see Sec. 4.2), and later for forecast verification.

https://doi.org/10.5194/hess-2022-53
Preprint. Discussion started: 8 February 2022
© Author(s) 2022. CC BY 4.0 License.

Hydrology and Earth System Sciences
Discussions

4 Methods

4.1 Verification metrics

The Kling-Gupta Efficiency (KGE, Gupta et al., 2009; Kling et al., 2012) is used to evaluate the performance of the forecasts. The KGE is defined as

$$\mathrm{KGE} = 1 - \sqrt{(r-1)^2 + (\beta-1)^2 + (\gamma-1)^2} \tag{1}$$

where r is the Pearson's correlation coefficient, β is the bias ratio, and γ is the variability ratio. The bias and variability ratios are defined as

$$\beta = \frac{\mu_{\mathrm{sim}}}{\mu_{\mathrm{obs}}}, \tag{2}$$

and

$$\gamma = \frac{\sigma_{\mathrm{sim}}}{\sigma_{\mathrm{obs}}}, \tag{3}$$

where μ_{sim} and σ_{sim} are the mean and standard deviation of the forecast discharge and μ_{obs} and σ_{obs} are the equivalent for the observed discharge. The KGE is widely used in hydrology as each component is a measure of a different type of error. The correlation coefficient, r, is a measure of temporal errors, the bias ratio, β indicates whether the discharge tends to be over- or under-predicted by the forecast, and the variability ratio, γ measures if the forecast captures the variability of the discharge magnitudes (Harrigan et al., 2020).

For a perfect forecast each component (r, β, and γ) has a value of 1, giving KGE=1. We also define two benchmarks. Following Knoben et al. (2019), we define a 'skilful' forecast as one where the KGE is higher than for a mean observed discharge benchmark (i.e., KGE$> 1-\sqrt{2} \sim -0.414$). We also arbitrarily define a 'highly skilful' forecast as one where KGE$> \sqrt{2}/2 \sim 0.707$. This benchmark corresponds to a mean relative error in the three coefficients of about 17%, or an error of about 30% in one coefficient if the other two have zero error.

As an additional upper benchmark, we will also use persistence forecasts. Persistence forecasts can help us to determine which stations are 'easy' or 'hard' to forecast. For each station, these are constructed by persisting its mean observed discharge from the previous 48 hours. For example, the 5-day persistence forecast for May 23 is given by the mean flow between May 16 and May 18. Evaluated at a fixed lead time, such persistence forecasts asymptotically approach the observed mean and standard deviation of the observations over long periods, typically giving them high KGEs. As such, we also use the Nash-Sutcliffe Efficiency Nash and Sutcliffe (NSE; 1970), which validates forecast or simulated flow based only on covariance with the observations, thus:

$$\mathrm{NSE} = 1 - \frac{\Sigma_t (Q_{\mathrm{sim}}(t) - Q_{\mathrm{obs}}(t))^2}{\Sigma_t (Q_{\mathrm{obs}}(t) - \overline{Q_{\mathrm{obs}}(t)})^2}, \tag{4}$$

where Q_{sim} is the simulated discharge, Q_{obs} is the observed discharge, and the overbar denotes a long-term average.

https://doi.org/10.5194/hess-2022-53
Preprint. Discussion started: 8 February 2022
© Author(s) 2022. CC BY 4.0 License.

Hydrology and Earth System Sciences
Discussions

4.2 Bias correction

When undertaking bias correction, we have a range of choices of complexity – ranging from the very simple (additive/multiplicative) through increasingly advanced methods (e.g. quantile mapping). Here, we have the advantage of a long timeseries of training data and we want to maximise the forecast skill under the single constraint that forecast output from GloFAS is the only input. To that end, we employ both quantile mapping and spatial fitting techniques, splitting the bias correction into two serial algorithms, which we outline below.

4.2.1 Quantile mapping

For the first stage of the bias correction we employ a basic quantile mapping method. The training period, January 2005 to June 2018, was extracted from the observational record. The start date was chosen because some stations have increasingly sparse and/or spurious records before 2005, and the end date is chosen so that a twelve-month testing period was available before the start of the operational forecasting period. GloFAS-ERA5 streamflow was extracted for the same period, not only for the grid point in which the gauge of interest is located, but also for surrounding points in a $0.6^\circ \times 0.6^\circ$ box centred on the gauge. This gives a total of 36 locations (given the 0.1° spacing of global GloFAS output).

Iteratively, these are then quantile-mapped to the observed streamflow, that is:

$$m_{\text{bc}}(i,t) = \tilde{q}_{\text{obs}}(q_{\text{raw}}(m_{\text{raw}}(i,t))), \tag{5}$$

where q is a function that maps streamflow to streamflow quantiles, $\tilde{q}$ is its inverse, m_{raw} and m_{bc} are the raw and bias-corrected modelled streamflows respectively, and i and t are spatiotemporal indices. The forms of q_{raw} and q_{obs} are both computed using data from the training period and used unchanged for the testing period and operational forecasts. For example, consider a forecast streamflow value of 38.7 $\text{m}^3\ \text{s}^{-1}$ for a grid point containing the NFSW4 gauge. This value, were it in the GloFAS-ERA5 training period, would have a quantile value of 0.88. The 0.88th quantile (or 88th percentile) for observed streamflow at NFSW4 in the training period is 77.0 $\text{m}^3\ \text{s}^{-1}$, and so this would be the value used for m_{bc} for that point at that forecast time. This makes sense, given that raw GloFAS underestimates high flow at NFSW4 by about 50% (cf. Fig. 3). This quantile mapping technique is then carried out independently for each of the 36 grid points in the neighbourhood of each gauge, in each case mapping the GloFAS output for the specific grid point to the observations taken at the gauge (as opposed to observations taken at the grid point itself).

4.2.2 Spatial optimisation

We must then convert these forecast values over the neighbourhood grid points into a single forecast value. This was achieved through a simple linear summation, i.e.,

$$m_{\text{bc}}(t) = \Sigma_i a_i m_{\text{bc}}(i,t), \tag{6}$$

where the coefficients a_i are to be determined.

https://doi.org/10.5194/hess-2022-53
Preprint. Discussion started: 8 February 2022
© Author(s) 2022. CC BY 4.0 License.

Hydrology and Earth System Sciences
Discussions

To compute a_i, eq. 6 was treated as an optimisation problem using the same training period as earlier (January 2005 to June 2018). Here, we seek to minimise

$$-\text{NSE}(m,o) - \text{KGE}(m,o)\,, \tag{7}$$

i.e. the negative sum of the Kling-Gupta and Nash-Sutcliffe efficiencies over the training period. This optimisation was carried out subject to the constraint that $0 \leq a_i \leq 1$, to prevent unphysical behaviour, computational noise, and/or overfitting. A sequential quadratic programming technique was then used to compute the optimal bias matrix, a_i, due to its ability to obey constraints through Lagrange multipliers (Nocedal and Wright, 2006). The resulting (static) 6×6 matrices were computed and stored for each of the ten gauges and are then used during each forecast to convert the quantile-mapped $n_t \times 6 \times 6$ forecasts into 1D vectors of length n_t.

As an example, the bias matrix for NFSW4 is:

$$\begin{bmatrix} 0.089 & 0 & 0 & 0 & 0.138 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.754 & 0.009 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.001 & 0 & 0 & 0.023 \end{bmatrix}. \tag{8}$$

It is clear that the gauge is located in the grid point associated with $a_i = 0.754$; however, there are some additional contributions from nearby river points. Aside from the direct spatial error described in the previous subsection, these uncentred contributions (both in NFSW4 and at other stations) largely work to correct two smaller biases. Firstly, any local spatial bias in the ERA5 precipitation used to drive GloFAS-ERA5 that, for example, incorrectly increases the streamflow in a nearby channel. Secondly, a temporal bias in the streamflow, for example an upstream point may receive nonzero weighting if the modelled streamflow at the gauge is occurring later than in observations.

Two final adjustments are made when the bias correction is run operationally. Firstly,the bias-corrected forecast is slightly relaxed towards the raw forecast ($q_{\text{bc}}^{\text{new}} = 0.25 q_{\text{raw}} + 0.75 q_{\text{bc}}^{\text{old}}$) to account for the different climatologies of GloFAS-ERA5 and GloFAS. Secondly, the whole forecast is then shifted by an additive δ, the difference between the mean observations and the mean day-1 forecasts over the two days prior to the forecast being issued.

4.3 LSTM

A number of hyperparameters are required for an LSTM, these were largely guided by previous literature (e.g. Kratzert et al., 2018) and then tuned using sensitivity tests. We decided to use basin-mean variables so that the LSTM architecture could be consistent for each of the ten gauges. That architecture is as follows:

There are three stacked LSTM layers, one input and two hidden, each with fifty neurons, and a single neuron dense output layer. We use 23 input variables covering the surface meteorology (10-m u and v winds, 2-m temperature, 2-m dewpoint, skin

https://doi.org/10.5194/hess-2022-53
Preprint. Discussion started: 8 February 2022
© Author(s) 2022. CC BY 4.0 License.

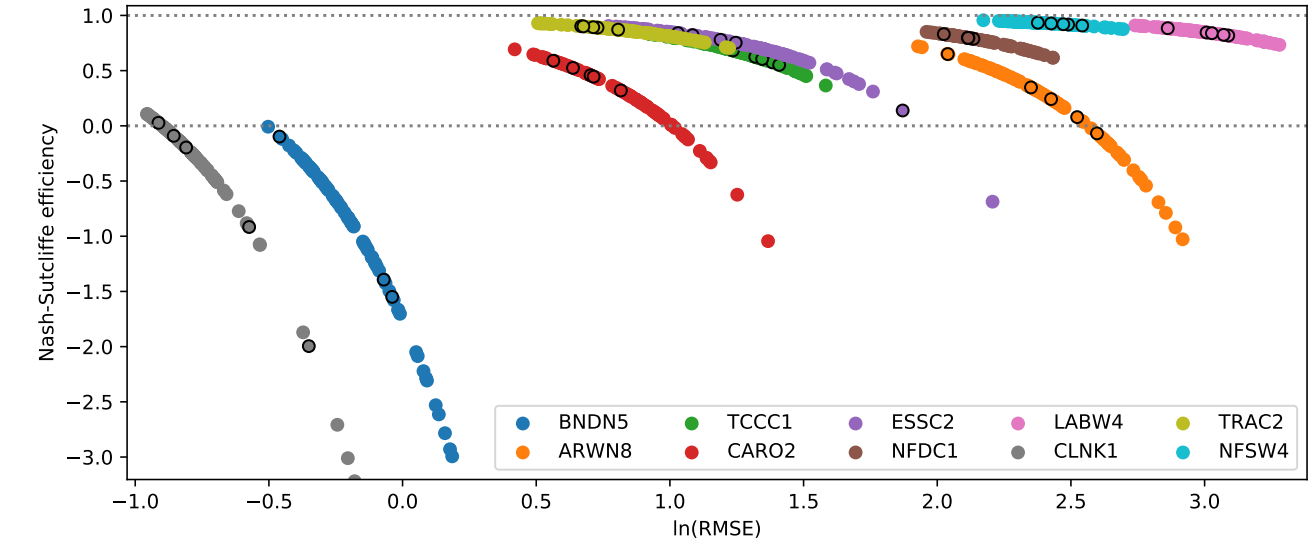


Figure 2. The Nash-Sutcliffe efficiency and logarithm of the root mean square error in streamflow ($m^3\ s^{-1}$) for each of the hundred ensemble members of each gauge-specific LSTM, computed over the June 2019 – June 2020 testing period. Each member was trained for ten epochs, except for five that were trained for 100 epochs. These five are marked with an additional black ring.

Layer	Input	Activation	Output
LSTM layer 1	(28,23)	ReLU	(28,50)
LSTM layer 2	(28,50)	ReLU	(28,50)
LSTM layer 3	(28,50)	tanh	(1,50)
dense	(1,50)	linear	(1)

temperature), surface hydrology (total precipitation, runoff, surface runoff, skin reservoir content, snowfall, snow depth, snow cover, surface latent heat flux, evaporation), subsurface hydrology (subsurface runoff, soil water volume over four layers, soil temperature), as well as GloFAS streamflow, and the mean historical discharge from gauge observations for the given day of year. For each of the ten gauges, all variables except streamflow and the historic mean (which are taken as point data at the gauge location) are averaged over the catchment (see Fig. 1) at a six-hourly frequency. We use a seven-day sequence for the input, giving the vector for each variable a length of 28 timesteps.

The LSTM was trained separately for each gauge over the training period (January 2005 to June 2018). For each gauge, the LSTM was trained 100 times using random starting weights over ten epochs each. Five additional ensemble members were also trained over 100 epochs to determine whether a extending the number of epochs provided a significant performance gain. Fig. 2 shows the performance of each ensemble member of the LSTM for each gauge, computed over the testing period (June 2019 to June 2020). We see that there is no significant advantage to extending the number of epochs beyond ten per member.

https://doi.org/10.5194/hess-2022-53
Preprint. Discussion started: 8 February 2022
© Author(s) 2022. CC BY 4.0 License.

Hydrology and Earth System Sciences
Discussions

For the operational product, the mean is taken from the five best-performing ensemble members (those with the highest NSE in Fig. 2). For eight out of the ten gauges, this yields a better KGE over the testing period than using the best-performing single member, as we will see in Sec. 5.1. The LSTM is trained using data from ERA5 and GloFAS-ERA5, which we also use for the testing phase. Although this leaves the operational LSTM vulnerable to biases in the IFS, these are mitigated to an extent with the IFS being the driving model for ERA5. On the positive side, this means that evaluation of the ability of the forecasts will be conservative and thus more informative to potential users who, in not having access to decades of archived forecasts with a consistent model version, must also resort to using reanalysis training data.

Note: snowc not available in IFS output

5 Results

5.1 Evaluation over test period: June 2019 to June 2020

5.1.1 Raw GloFAS-ERA5

Figure 3 compares raw (i.e. not bias corrected) GloFAS-ERA5 streamflow to observations at each of the ten stations over the test period (June 2019 to June 2020). The product is skilful (NSE>0 and KGE> $1-\sqrt{2}$) at six of the ten stations and highly skilful (KGE> $\sqrt{2}/2$) at two. The stations where raw GloFAS-ERA5 was not skilful (BNDN5, ARWN8, CARO2, and CLNK1) are characterised by low mean flows and highly intermittent peaks. Often, peaks appear at the wrong time – as exemplified at CLNK1 during the autumn months of 2019 – or respond too slowly or too smoothly to short precipitation stimuli.

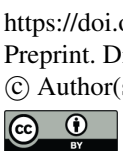
At stations where the raw GloFAS-ERA5 product is skilful but not highly skilful (TCCC1, ESSC2, NFDC1, and LABW4) are typically marked by it capturing the annual cycle well, but generally missing some or most intraseasonal variability. This is evident in ESSC2 and LABW4, where the summer maximum and winter minimum were well captured, but autumn and spring storms were not. Those stations where the product is highly skilful (TRAC2 and NFSW4) also capture intraseasonal variability well – see for example April and May 2020 at NFSW4, where the discharge associated with two spring storms is well simulated by the model.

5.1.2 Bias-corrected GloFAS-ERA5

Bias-corrected GloFAS-ERA5 is compared with observational streamflow for each of the ten gauges over the testing period in Fig. 4. The results are a substantial improvement over the raw GloFAS-ERA5 output: NSE is improved at seven of the ten stations (except BNDN5, ARWN8, TCCC1) and KGE is also improved at seven stations (except BNDN5, TCCC1, NFDC1). Following bias-correction, GloFAS-ERA5 is still skilful at seven stations but now highly skilful at four.

Failures at BNDN5 and ARWN8 are mostly due to the bias correction algorithm being unable to calibrate low and sporadic flow; although at ARWN8, the quantile mapping brings the simulated mean and variance closer to observations as desired, resulting in an improved KGE. Although the simulated flow at CLNK1 is still not skilful, it has been improved considerably by

https://doi.org/10.5194/hess-2022-53
Preprint. Discussion started: 8 February 2022
© Author(s) 2022. CC BY 4.0 License.



Hydrology and Earth System Sciences
Discussions

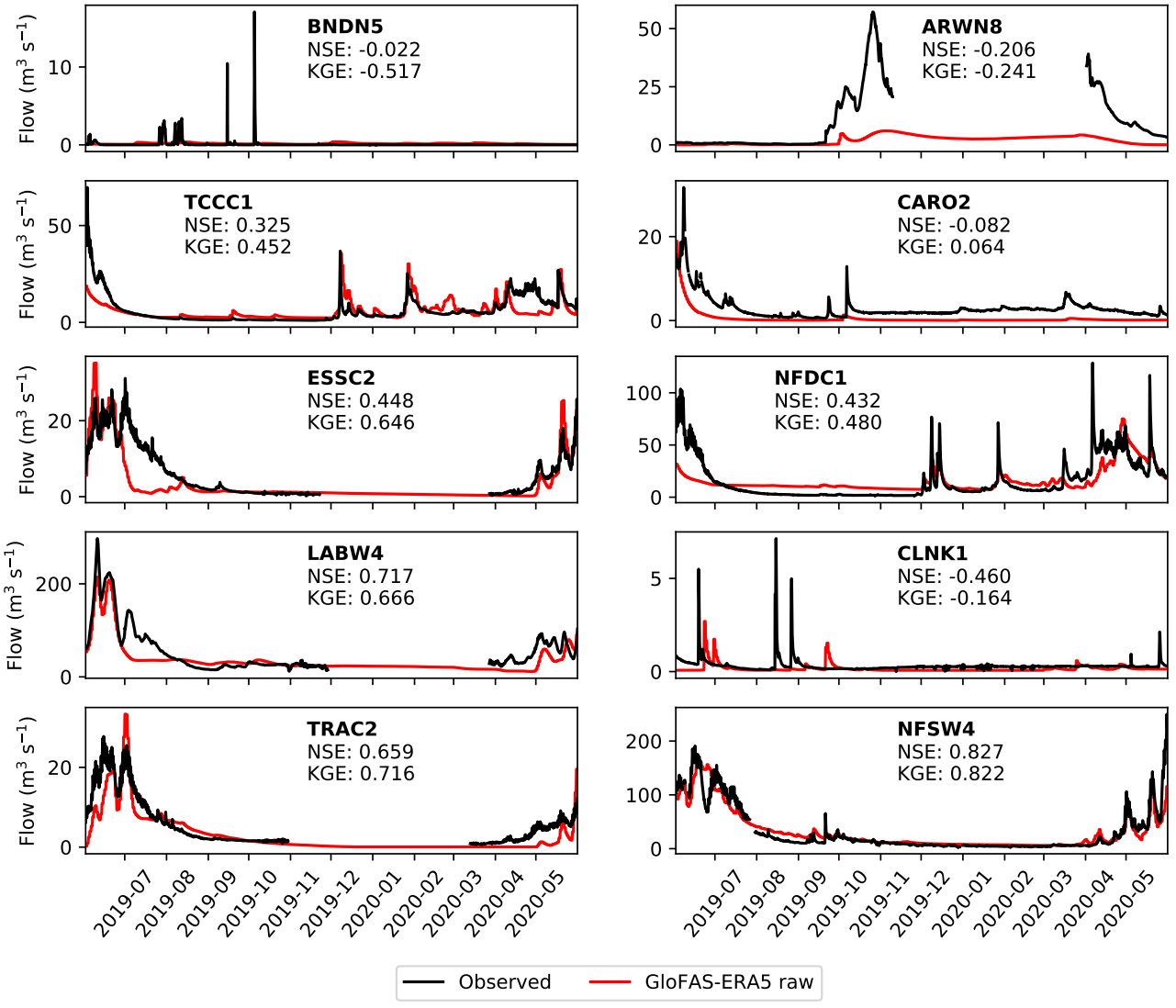


Figure 3. Comparison of observed streamflow [black] with raw GloFAS-ERA5 output [red] for each of the ten gauges over the testing period (June 2019 to June 2020). Gaps in the observational record over the winter are due to river freezing. Nash-Sutcliffe and Kling-Gupta efficiencies over the year-long period are given for each gauge.

285 the bias correction where the spatial optimisation routine has somewhat improved the timing of the peaks. Both CARO2 and TRAC2 also saw large increases in NSE and KGE due to improved representation of intraseasonal variability.

https://doi.org/10.5194/hess-2022-53
Preprint. Discussion started: 8 February 2022
© Author(s) 2022. CC BY 4.0 License.

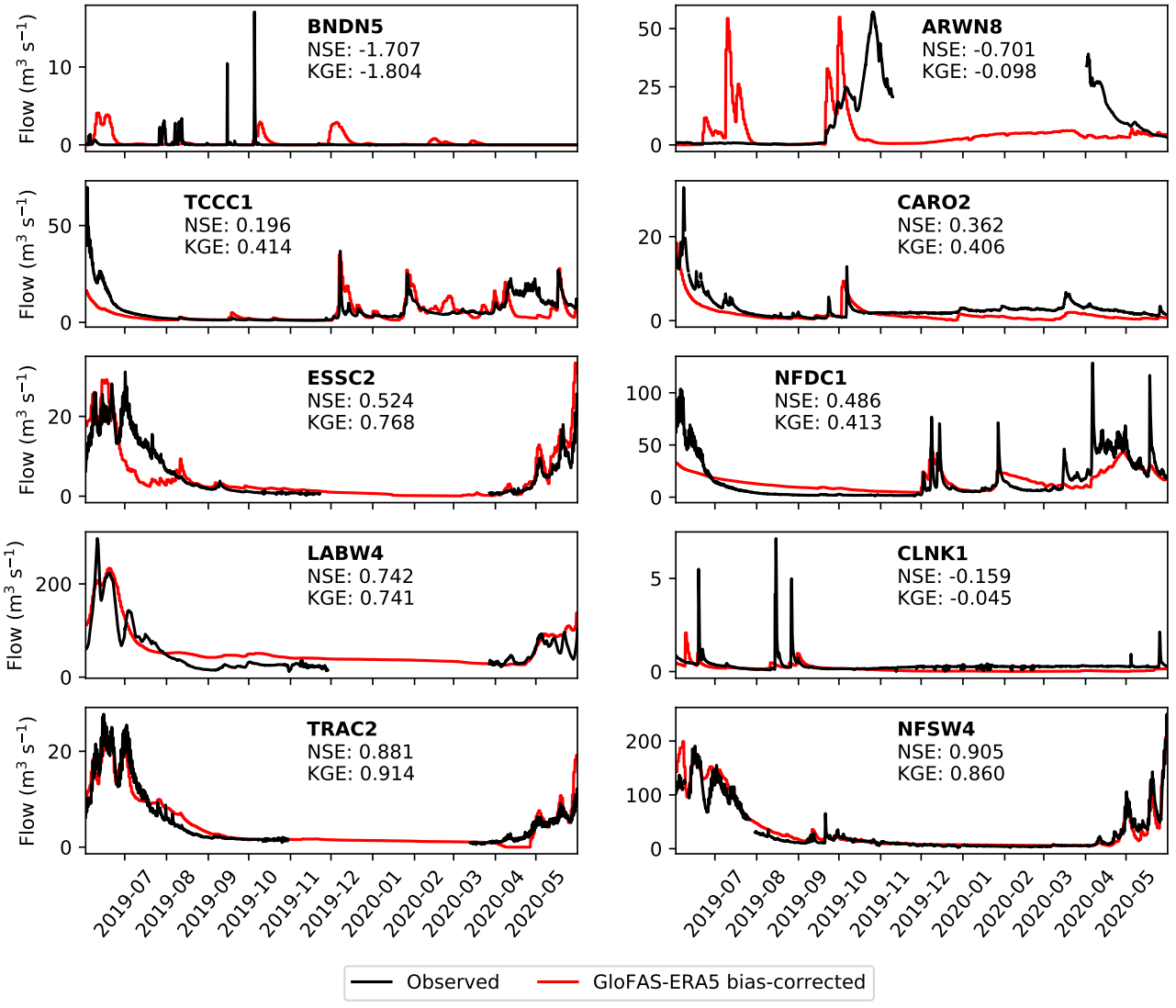


Figure 4. Comparison of observed streamflow [black] with bias-corrected GloFAS-ERA5 output [red] for each of the ten gauges over the testing period (June 2019 to June 2020). Gaps in the observational record over the winter are due to river freezing. Nash-Sutcliffe and Kling-Gupta efficiencies over the year-long period are given for each gauge.

5.1.3 LSTM

Fig. 5 shows the performance of the LSTM model, ingesting ERA5 and GloFAS-ERA5, at each gauge over the testing period. It represents a step-change in model efficiency over the bias-corrected GloFAS-ERA5 output, being skilful at nine gauges and highly skilful at six. The KGE is greater than 0.9 at three stations and the NSE is greater than 0.9 at four.

https://doi.org/10.5194/hess-2022-53
Preprint. Discussion started: 8 February 2022
© Author(s) 2022. CC BY 4.0 License.

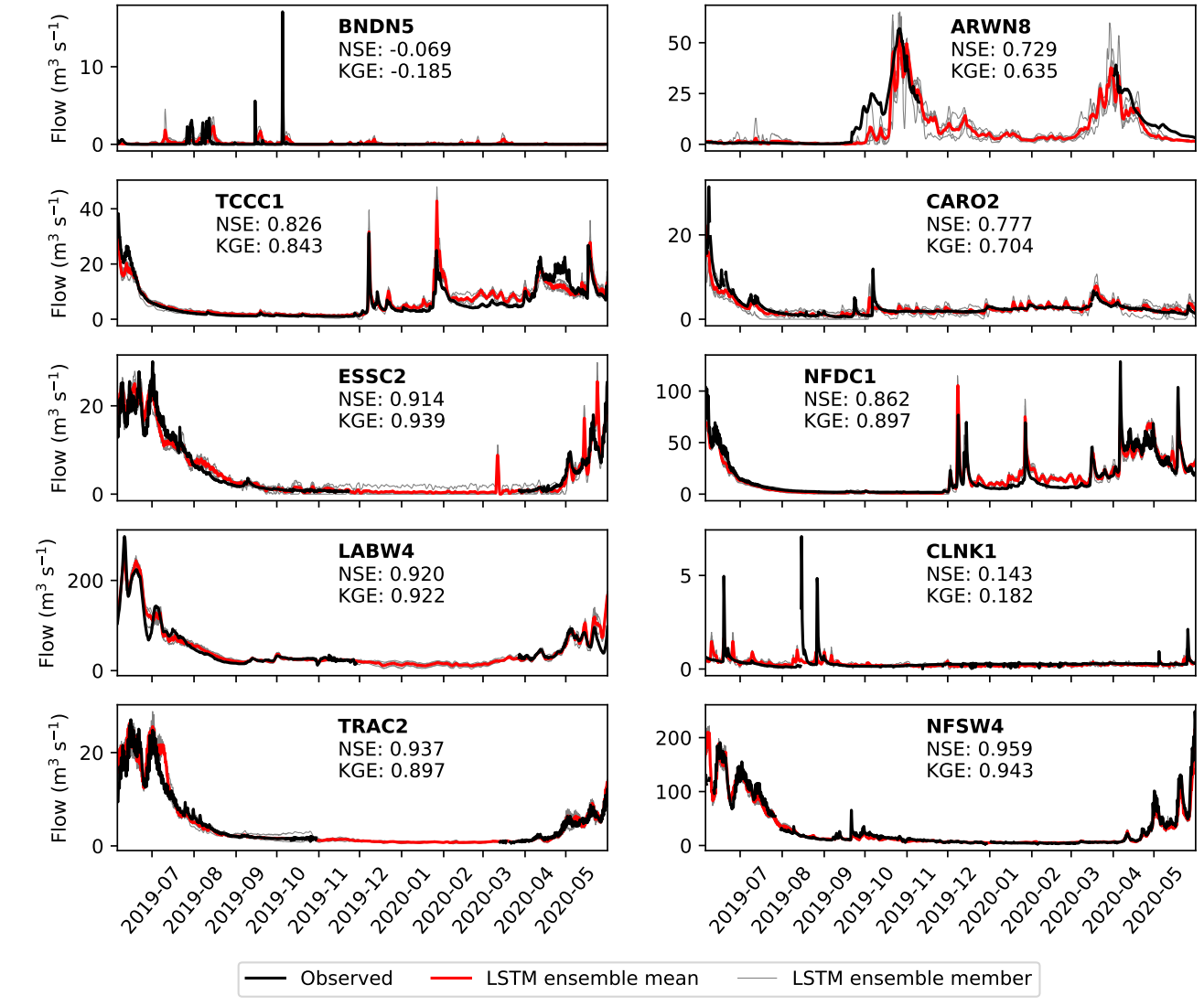
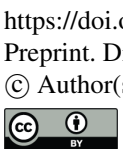


Figure 5. Comparison of observed streamflow [black] with the LSTM ensemble output [members:grey, mean:red] for each of the ten gauges over the testing period (June 2019 to June 2020). Gaps in the observational record over the winter are due to river freezing. Nash-Sutcliffe and Kling-Gupta efficiencies over the year-long period are given for each gauge.

Despite this, the BNDN5 and CLNK1 gauges remain relative poorly modelled (although the latter does qualify as skilful). In the observations, BNDN5 is characterised by long periods of no flow, with occasional short-lived (typically less than two days) peaks. The LSTM does manage to capture these peaks, but the timing is incorrect – usually several days late – and the magnitude is often far too small. The first of these issues, and perhaps the second, is likely due to the LSTM ingesting catchment-mean variables. The catchment basin for BNDN5 is large and arid, and therefore rain falling over it is probably not appropriately captured using this simplification. Future development of an LSTM model for streamflow at BNDN5 or similarly

https://doi.org/10.5194/hess-2022-53
Preprint. Discussion started: 8 February 2022
© Author(s) 2022. CC BY 4.0 License.



Hydrology and Earth System Sciences
Discussions

large and arid basins should thus consider an expanded input that ingests variables over all (or representative) grid points in the basin. At most of the other gauges, intraseasonal variability is captured very well. This is true both for individual storms – for example high-discharge events at TCCC1 in early December 2019 and NFDC1 in late January 2020 were almost perfectly
300 simulated – and broader flow patterns such at ESSC2 in summer 2019 and NFSW4 in spring 2020.

5.2 Verification over operational period: September 2020 to October 2021

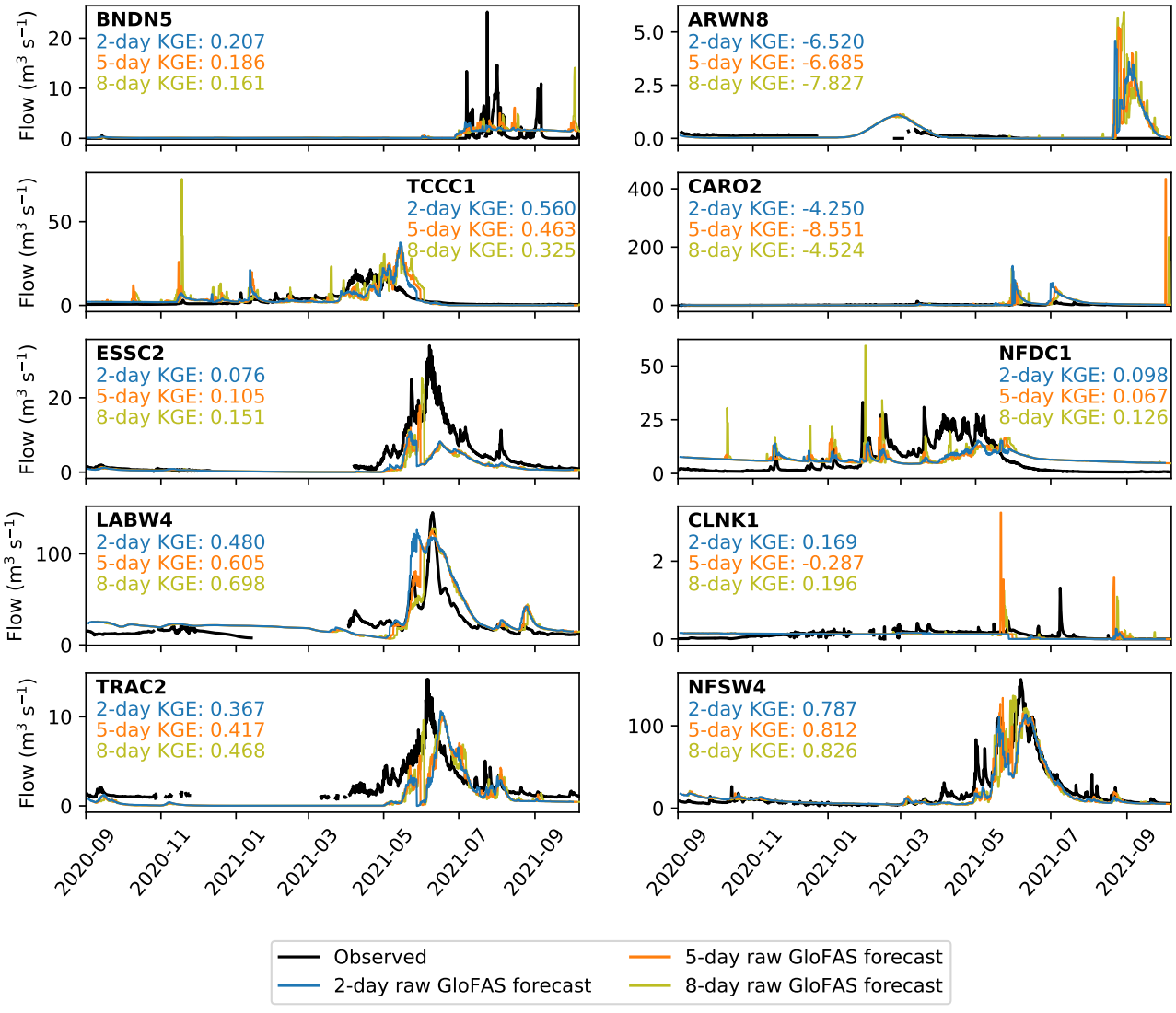


Figure 6. Verification of the 2-day (blue), 5-day (orange) and 8-day (yellow) raw GloFAS forecasts against observations (black) over the operational period (September 2020 to October 2021). Observations are plotted at six-hourly resolution to match the forecasts and are not plotted when the river is frozen. Kling-Gupta efficiencies for each lead time at each station are also given.

https://doi.org/10.5194/hess-2022-53
Preprint. Discussion started: 8 February 2022
© Author(s) 2022. CC BY 4.0 License.

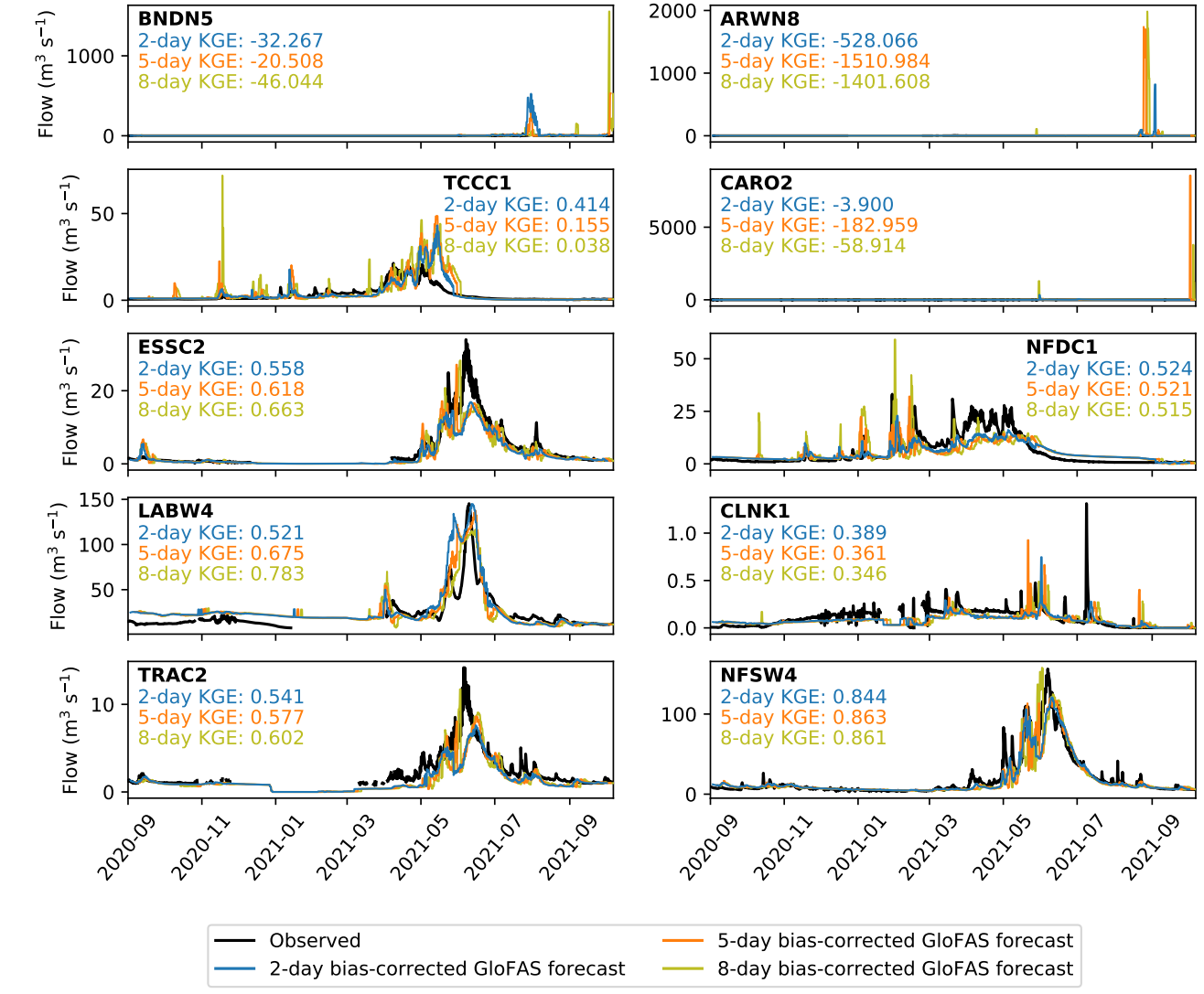


Figure 7. Verification of the 2-day (blue), 5-day (orange) and 8-day (yellow) bias-corrected GloFAS forecasts against observations (black) over the operational period (September 2020 to October 2021). Observations are plotted at six-hourly resolution to match the forecasts and are not plotted when the river is frozen. Kling-Gupta efficiencies for each lead time at each station are also given.

Following testing, we ran all three models (raw GloFAS, bias-corrected GloFAS, and the LSTM) operationally for thirteen months between September 1 2020 and October 1 2021, producing ten-day forecasts at daily frequency. As discussed in Sec. 4, the major difference between the testing period and the operational period is the switch from ERA5 to IFS for the variables ingested by the GloFAS prediction system and by the LSTM. The effects of this switch are somewhat mitigated by using the same underlying model (ERA5 uses the IFS), but there will be biases in IFS forecasts that are not present in ERA5, since the latter is nudged with simultaneous observations that are not available to the forecasts at nonzero lead times.

https://doi.org/10.5194/hess-2022-53
Preprint. Discussion started: 8 February 2022
© Author(s) 2022. CC BY 4.0 License.

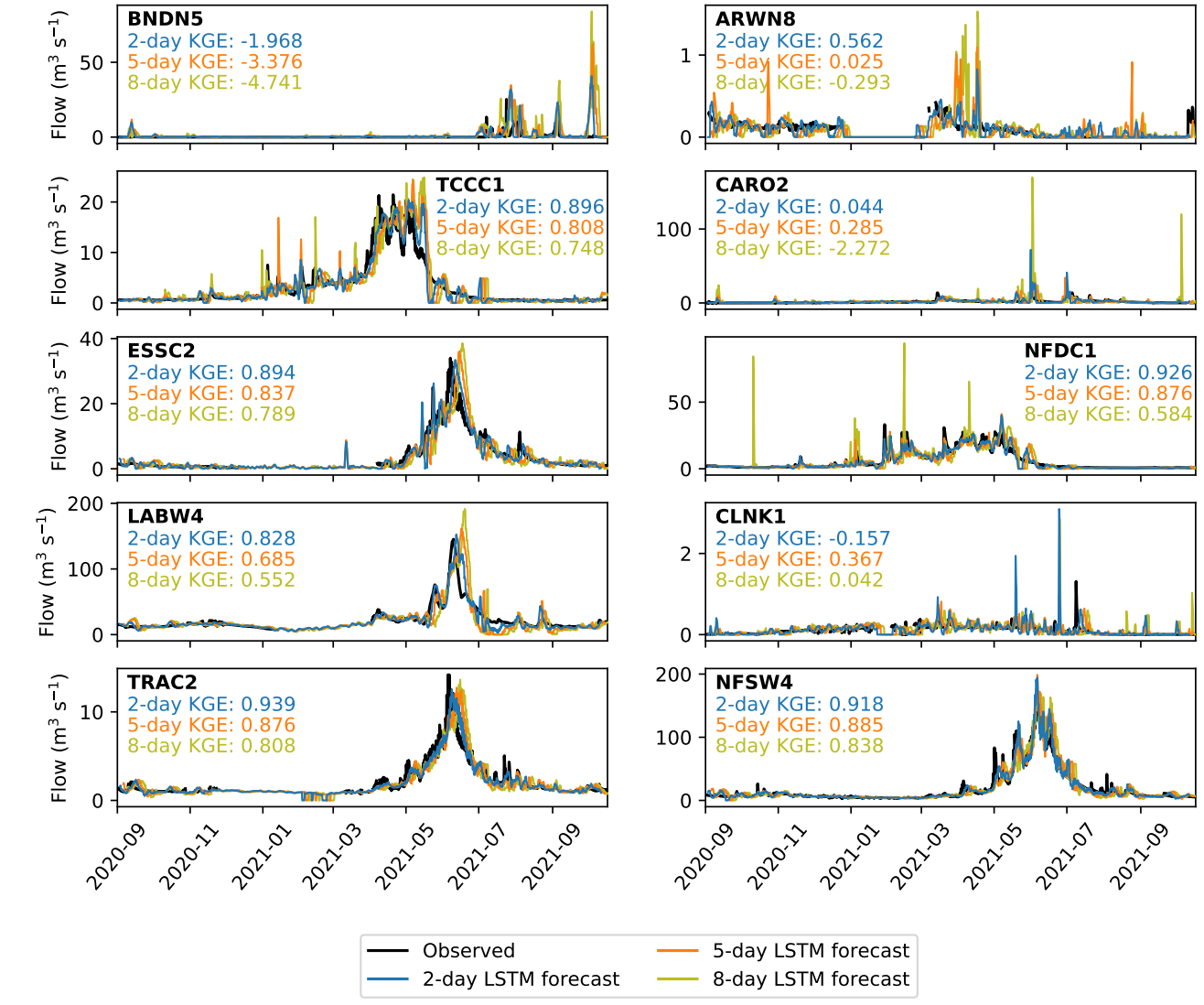


Figure 8. Verification of the 2-day (blue), 5-day (orange) and 8-day (yellow) LSTM forecasts against observations (black) over the operational period (September 2020 to October 2021). Observations are plotted at six-hourly resolution to match the forecasts and are not plotted when the river is frozen. Kling-Gupta efficiencies for each lead time at each station are also given.

Forecasts at each station are evaluated at 2-, 5-, and 8-day lead times against the official observations. Results are given over the entire operational period for the raw GloFAS forecasts in Fig. 6, the bias-corrected GloFAS forecasts in Fig. 7, and for the LSTM forecasts in Fig. 8. As expected, even at short lead times, the forecasts generally perform more poorly than their reanalysis-based counterparts. There are a handful of notable exceptions to this, particularly for the raw GloFAS forecasts at BNDN5, TCCC1, and CLNK1. Although the GloFAS forecasts derive from a marginally more recent version (CHECK!) of the IFS than does ERA5, the gains in this context are likely to be insignificant. At BNDN5 and CLNK1, the difference appears

https://doi.org/10.5194/hess-2022-53
Preprint. Discussion started: 8 February 2022
© Author(s) 2022. CC BY 4.0 License.

Hydrology and Earth System Sciences
Discussions

to be due to the hydrographs of the testing and operational periods having different characteristics. During the testing period, there was no flow at BNDN5 except for several very short (∼daily) pulses of nonzero discharge in autumn 2019; during the operational period, however, autumn 2021 was parked by a period of low but continuous flow. The lagged autocorrelation of the latter situation almost invariably makes it easier to model. At CLNK1, constant very low flow is punctuated by occasional, short-lived peaks. There was only one significant peak in the operational period compared with four in the testing period, making it easier to model correctly. This is arguable also the case at TCCC1, where increased subseasonal variability in the operational period – notably two big storms in December 2019 and January 2021 – made it more challenging to simulate than the testing period.

Evaluating overall performance by computing the KGE of the 5-day forecasts, we find that the bias-corrected GloFAS forecast beats the raw GloFAS forecast at six stations, the LSTM forecast beats the bias-corrected GloFAS forecast at all ten stations and the raw GloFAS forecast at nine. The raw GloFAS 5-day forecasts were skilful at seven stations and highly skilful at one of these; the bias-corrected GloFAS 5-day forecasts were also skilful at seven stations and highly skilful at one of these – though with significant improvement over the raw GloFAS forecasts at six of the seven showing skill. The LSTM 5-day forecasts were skilful at nine stations and highly skilful at five of these.

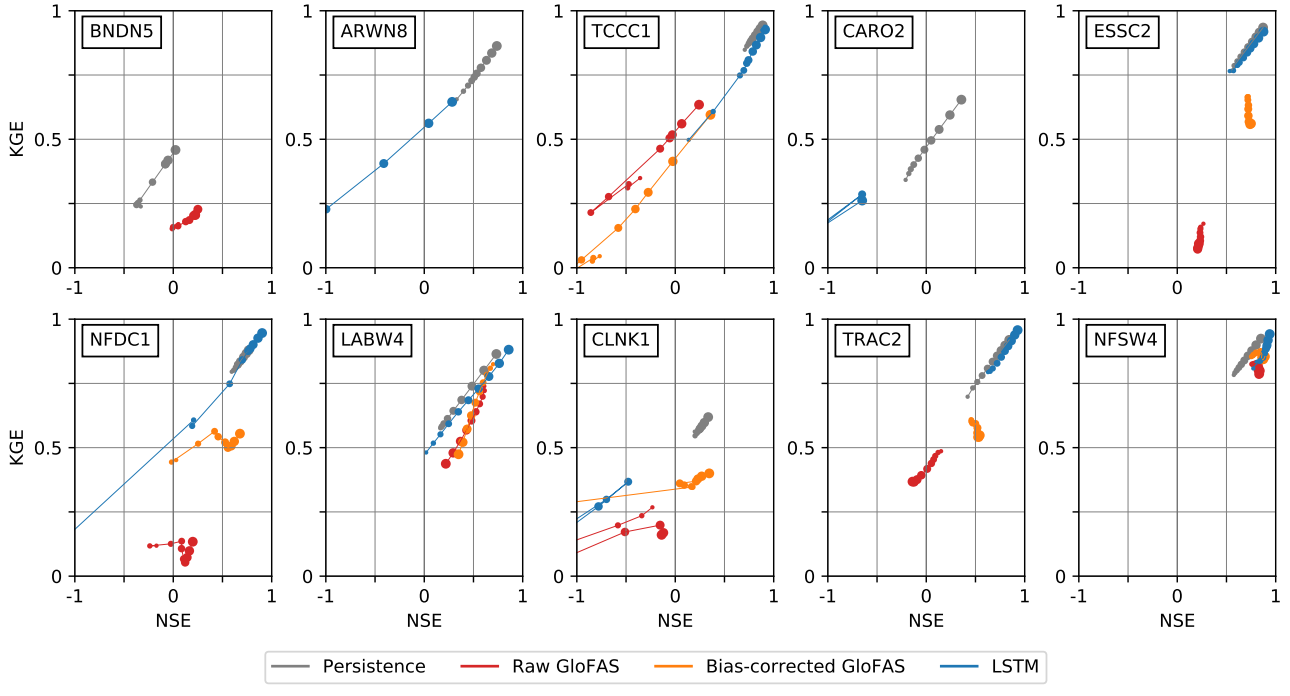


Figure 9. Overall performance of the three models during the operational period (September 2019 to October 2020) compared with a persistence benchmark (black). For each model, at each station, for each lead time (from one to ten days), the Nash-Sutcliffe and Kling-Gupta efficiencies – compared to observations – are plotted. Lead times are connected sequentially by markers of diminishing size, with the ten-day lead time having the smallest marker. Values of KGE below zero or NSE below -1 are not plotted.

https://doi.org/10.5194/hess-2022-53
Preprint. Discussion started: 8 February 2022
© Author(s) 2022. CC BY 4.0 License.

Hydrology and Earth System Sciences
Discussions

To complete this evaluation, we now extend our analysis to all lead times and include a simple persistence model as a benchmark (Sec. 4). For sufficiently long verification periods, the mean and variance of persistence forecasts by definition asymptotically approaches those of the observations. This has two key implications: firstly, since KGE is sensitive to the errors in forecast mean and variance, the persistence model receives a large advantage in this score compared to the other models; secondly, following on from the first point, KGE is no longer a fair descriptor of model performance, so we must also consider NSE.

Fig. 9 shows how the KGE and NSE vary for each model at each station as a function of lead time. Here, we are more interested in the higher values of these metrics (i.e. if a model is skilful, how skilful is it?) than the orders of magnitude of negative KGE and NSE values produced by useless forecasts, and so we truncate the plots at NSE=-1 and KGE=0. For each model at each station, scores are plotted from lead times of one to ten days inclusive, denoted by markers of diminishing radius, and connected in order by thin lines of the same colour, creating 'phase space caterpillars'. Presenting the metrics in this way allows a quick intercomparison between models, for example we can see clearly where bias correction has worked well (ESSC2, NFDC1, TRAC2) by identifying where the bias-corrected points have moved significantly upward and to the right compared to their raw counterparts. With the exception of BNDN5, CLNK1 and CARO2, we see that the LSTM forecasts tend to have comparable KGE to the persistence forecasts, but higher NSEs. Both the persistence and LSTM forecasts tend to outperform the raw and bias-corrected GloFAS forecasts. We also note that stations with a high lagged autocorrelation (i.e. those stations where the persistence model does well) tend also to have discharge that is well simulated by all three of the operational models. Exemplified by LABW4 and NFSW4, these are typically high-discharge sites with large catchment areas and slow response times. Indeed, those stations where the GloFAS and LSTM models struggle the most are characterised by flash responses, with their dominant mode of variability on the diurnal, rather than annual or semiannual timescale.

Next, to understand how the models perform as a function of forecast lead time, we decompose the KGE values into their three components: correlation (Fig. 10, bias ratio (Fig. 11), and variability ratio (Fig. 12). As before, we truncate the graphs to remove poor performing metric values where necessary. In terms of the correlation coefficient, the LSTM performs best at five of the ten stations (ARWN8, TCCC1, NFDC1, TRAC2, NFSW4) and most notably is the only forecast to have a positive correlation with the observation at ARWN8. This is mainly due to a fictional peak in the raw GloFAS which is exacerbated by the bias correction, but which is filtered out by the LSTM. In addition to these five stations, the LSTM also has the highest correlation at short lead times at LABW4 but this decreases at longer lead times, dropping below the correlation coefficient of the other two forecasts after four days. The timing of the largest peak at LABW4 during the operational period is better predicted by the raw and bias-corrected GloFAS forecasts at longer lead times, whereas for the LSTM it is shorter lead times that accurately predict this peak. The raw forecast has comparable correlation coefficients to the bias-corrected and LSTM forecasts at the three largest stations (BNDN5, CARO2, LABW4) but is notably lower for smaller catchments supporting the need for caution when using the raw GloFAS forecast for catchments below the recommended 2000 km^2 threshold. Since the bias-correction method is based on quantiles it can impact the correlation of the forecasts. The bias-corrected GloFAS has an improved correlation compared to the raw GloFAS forecast at seven of the ten stations (TCCC1, ESSC2, NFDC1, CLNK1, TRAC2, NFSW4, ARWN8 (not shown)).

https://doi.org/10.5194/hess-2022-53
Preprint. Discussion started: 8 February 2022
© Author(s) 2022. CC BY 4.0 License.

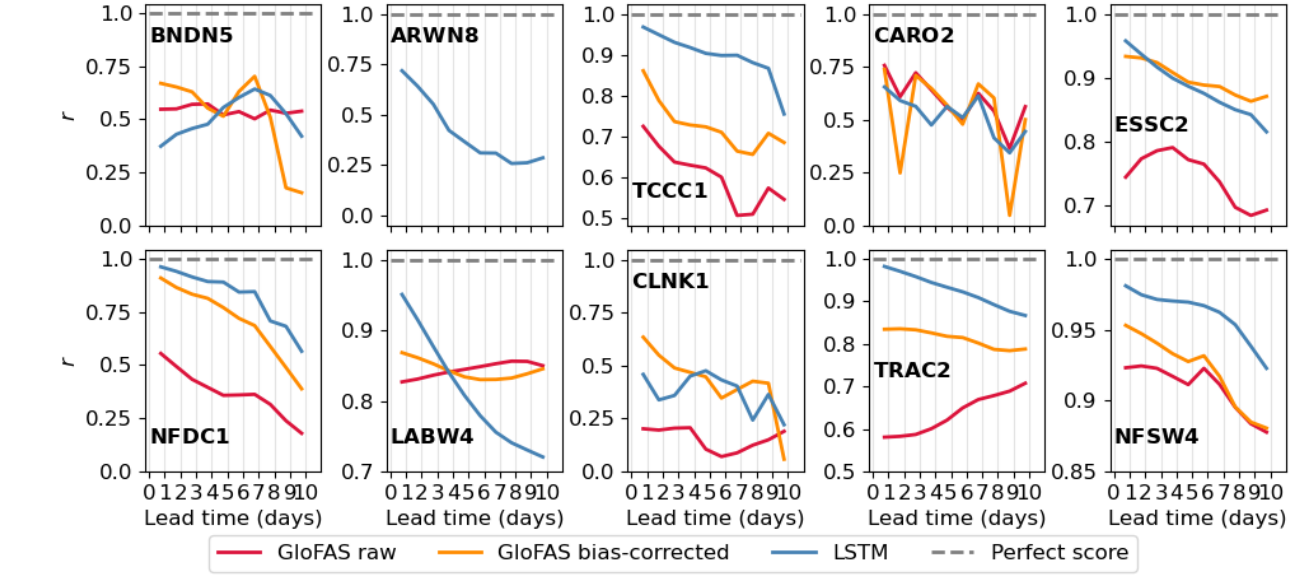


Figure 10. Correlation coefficient, r, calculated over the verification period for all three forecast models at each station for each lead-time (from 1 to 10 days). Negative correlations for ARWN8 have not been plotted.

Surprisngly, the bias ratio of the bias-corrected GloFAS forecasts is worse than than the raw GloFAS forecasts at four of the ten stations (BNDN5, ARWN8, TCCC1, CLNK1). These four stations tend to have low streamflow variability except for some short-duration large peaks. As these peaks are largely seasonal this vulnerability in the bias correction technique is likely due to the change in discharge distribution throughout the year, a problem that could be rectified by applying a season-based quantile mapping. The large bias ratios at BNDN5 and ARWN8, and CLNK1 and CARO2 at longer lead-times, are often due to unprecedentedly large values forecast by the raw GloFAS. The quantile mapping extrapolates these quantities to unphysically large values (see, e.g., 5000 m^3 s^{-1} in September 2021 at CARO2 in Fig. 7). The raw GloFAS forecasts underpredict the streamflow at the small catchments at high elevations (ESSC2, TRAC2). The bias correction does partially correct for this bias, but the LSTM forecast still has the lowest bias at these stations. In fact, the LSTM is the least biased forecast for eight of the ten stations, the exceptions being BNDN5 and CLNK1, where the raw forecast is better. For the largest catchment, BNDN5, the raw GloFAS forecast has a relatively small bias ratio compared to the other two forecasts mainly because it consistently predicted the zero-flow during the first half of the verification period.

The raw GloFAS forecast has the best variability ratio of all three forecasts at the two larger high-elevation catchments (LABW4, NFSW4), although at both stations the bias-corrected forecast has similar variability ratios at some lead times. However, the bias correction only improved the flow variability of raw GloFAS forecasts at half of the stations (BNDN5, ESSC2, NFDC1, CLNK1, TRAC2). However, these stations vary significantly in catchment characteristics (catchment size, elevation, peak duration, meteorological regimes), as do the stations where the bias correction is beneficial, so there is no obviously favourable catchment characteristic. The unphysical streamflow predictions seen at BNDN5, ARWN8, CLNK1 and

https://doi.org/10.5194/hess-2022-53
Preprint. Discussion started: 8 February 2022
© Author(s) 2022. CC BY 4.0 License.

Hydrology and Earth System Sciences
Discussions

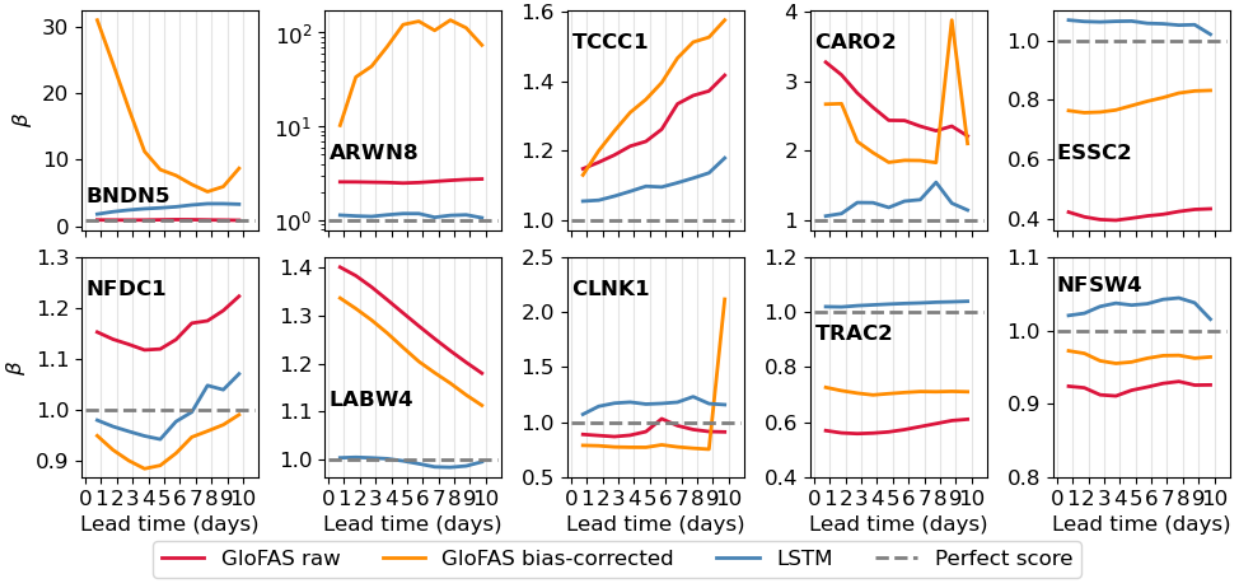


Figure 11. Bias ratio, β, calculated over the verification period for all three forecast models at each station for each lead-time (from 1 to 10 days). Note the logarithmic scale of the y-axis for ARWN8.

CARO2 also impact their variability ratio metric. The LSTM has the best variability ratio at three of the ten stations (BNDN5, ARWN8, TRAC2) and comparable values at a further four (CARO2, ESSC2, LABW4, CLNK1). All forecasts have a higher variability ratio at longer lead-times at NFDC1 with the rate of increase similar for all three forecasts. This suggests that the change with lead-time is due to factor impacting both such as a bias in the IFS forecast which is used to drive the LSTM and the LISFLOOD hydrological model used to create the GloFAS forecasts.

Finally, as this is a study on forecasting, we would be remiss not to analyse a single forecast. Fig. 13 shows the forecasts issued for the ten stations on May 1 2021, along with the persistence forecast, verified against observations. We see more easily here how the bias correction often – but not always – nudges the raw GloFAS forecast in the right direction, with particularly successful results at ARWN8 for this forecast. Of particular interest during this period was a pair of rain-on-snow events that impacted stations in the Rockies (NFSW4, LABW4, ESSC2, and TRAC2). These resulted in large spikes in the streamflow, visible in the observations, centred on May 3 and May 9. Neither the raw nor the bias-corrected GloFAS forecasts captured this, nor did a number of other operational forecasts (Kenneth Nowack; personal communication). While the LSTM mostly underestimated the magnitude of these events, it did predict them. This highlights the ability of LSTMs in general to learn complex non-linear relationships that may be altogether absent from physics-based models.

https://doi.org/10.5194/hess-2022-53
Preprint. Discussion started: 8 February 2022
© Author(s) 2022. CC BY 4.0 License.

Hydrology and Earth System Sciences
Discussions

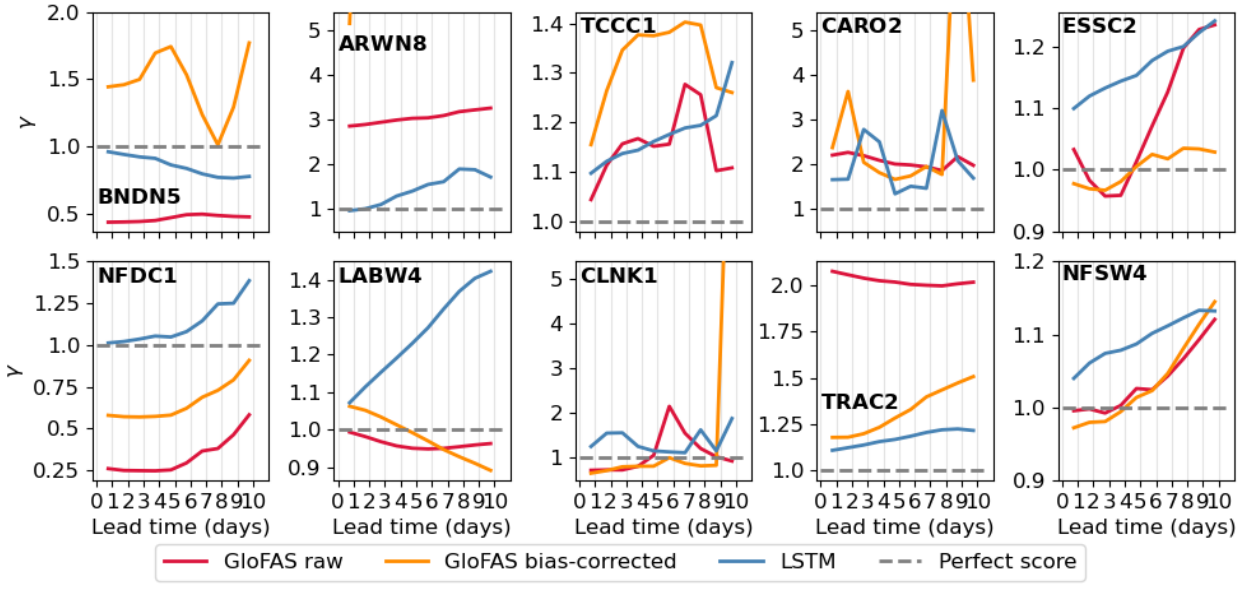


Figure 12. Variability ratio, γ, calculated over the verification period for all three forecast models at each station for each lead-time (from 1 to 10 days). The y-axis has been truncated to γ-values $= 5.2$ for stations ARWN8, CARO2, and CLNK1.

6 Discussion

6.1 Potential improvements

As with any newly-developed forecast models or techniques, we encountered scope for improvement during the operational phase and post-operational analysis. There are two potential improvements to the bias correction, in its current form. The first is to further granularise the process, computing a new quantile mapping and bias-correction matrix for (e.g.) each season. This would have the advantage of further reducing the relative bias caused by seasonally-varying environmental factors (e.g., snow on ground in winter; or intense storms in the summer saturating the soil). Such work must be careful to avoid overfitting the matrices given the increased degrees of freedom.

The bias correction would also benefit from a dependence on forecast lead time. Since biases invariably grow as a function of lead time, the correction required for a 10-day forecast is likely to be larger than the correction required for a 2-day forecast. In using GloFAS-ERA5 to compute our bias correction terms in Sec. 4.2, we effectively limited ourselves to a 0-day lead time correction. The hindcasts that would be required for this now exist from 1997 onwards, and work at ECMWF has used them for bias correction of flood threshold forecasts through CDF mapping (Zsoter et al., 2020).

Similarly, the LSTM was trained on ERA5 data, but then ingested IFS output when run operationally. Although the two products will share some biases, they will inevitably be larger in IFS (the forecast) than ERA5 (the reanalysis), resulting in errors that propagate non-linearly through the LSTM. Originally, we chose to train the model on ERA5 so that our methods

https://doi.org/10.5194/hess-2022-53
Preprint. Discussion started: 8 February 2022
© Author(s) 2022. CC BY 4.0 License.

Hydrology and Earth System Sciences
Discussions

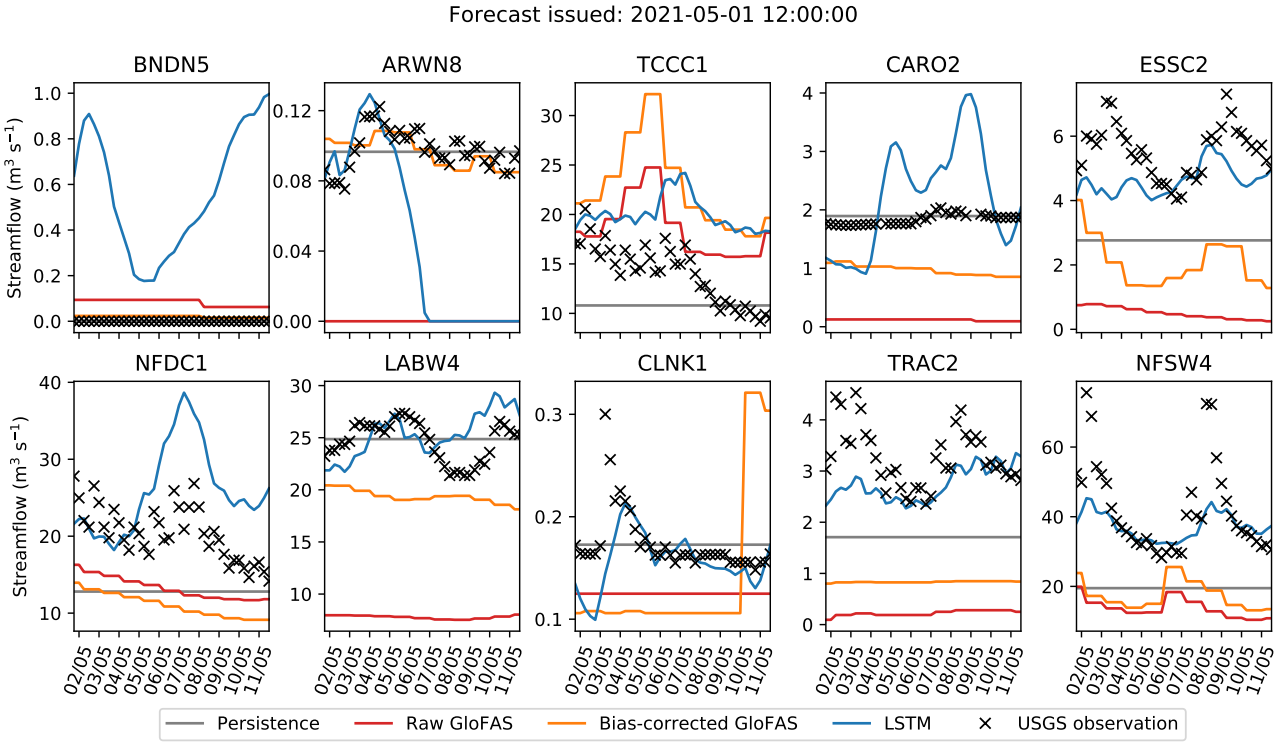


Figure 13. Example forecasts, issued on May 1 2021, for lead times of up to ten days. The four models (persistence: grey, raw GloFAS: red, bias-corrected GloFAS: orange, LSTM: blue) are verified against the official observations, plotted in black crosses.

were reproducible, but there is no reason other forecasters should be bound by this desire. The optimal strategy is to train the LSTM on IFS hindcasts – though this would require careful adjustment of the architecture to account for different lead times. Similarly, such an approach must be careful of changing hindcast model versions.

In our model, the LSTM is trained on, and subsequently ingests, catchment-mean variables (with the exception of streamflow, which is taken at the point of interest). However, as we know, rain falling far away from the station takes longer to reach it than rain falling nearby – a relationship lost when this approximation is made. By replacing the catchment-mean pre-processing step with a convolutional layer in the LSTM, the LSTM would be able to learn such spatiotemporal relationships and likely produce improved forecasts as a result. As we saw in the introduction, work by Le et al. (2019) showed that even a relatively simple LSTM that ingested data from different points upstream could produce excellent results. Despite the potential advantages, there are some caveats to adding a convolutional layer. Firstly, training them is computationally very expensive. Secondly, different basins have different areas, and therefore have to ingest a different number of vectors. This either requires a new architecture for each basin (essentially unfeasible) or intelligent preprocessing, e.g.,grouping the data by distance to the station. Thirdly, related to the second point, it does not allow transferability between products of different spatial resolutions without additional preprocessing.

https://doi.org/10.5194/hess-2022-53
Preprint. Discussion started: 8 February 2022
© Author(s) 2022. CC BY 4.0 License.


Finally, due to data pipeline constraints, we used only the control member of the IFS and GloFAS throughout our operational deployment. Two open research questions remain: what is the best way to combine ensemble members as inputs to an LSTM; and how can ensemble members be leveraged to provide accurate uncertainty estimates in the forecasts?

6.2 Potential applications

We have demonstrated, as have other authors, that LSTMs show a great deal of promise for river streamflow modelling and forecasting. Given this, there are other potential applications that are not immediately obvious from this work.

Perhaps the most exciting is the idea propounded by Kratzert et al. (2018), that LSTMs such as the one in this study are not black boxes, rather they are often 'grey' boxes, often containing many neurons whose output can be physically interpreted. In some cases – Kratzert et al. (2018) highlighted a neuron responsible for calculating snowmelt – representing relationships already captured by physics-based models; in others – e.g., if the rain-on-snow phenomenon described earlier is produced by a single neuron – representing relationships not necessarily captured by existing physics-based models. There is, then, the potential for new hydrological relationships to be discovered through careful investigation of a well-trained LSTM.

Another potential application is to use an LSTM model – in a basin where it has a high KGE and NSE – to infill missing data in the observational record, or to extend it further back in time where reanalysis coverage permits. Continuous long-term streamflow records are useful for both climate and hydrology research, as well as the insurance industry. Similarly, because the LSTM is extremely cheap to run once trained, it could readily be applied to climate model output (either following bias correction of that data, or by using transfer learning to adjust the internal weights of the LSTM) to produce projections of streamflow over selected basins in future climate scenarios.

7 Conclusions

In this study, we explored the efficacy of three models at simulating, and then forecasting up to a 10-day lead time, streamflow at ten different sites across the western United States. The forecasts were then verified against official observations and compared with a benchmark persistence model. The three models were:

1. The control member of the Global Flood Awareness System (GloFAS) ensemble, a physics-based model developed by ECMWF and the Joint Research Centre of the European Commission that provides global forecasts at a resolution of $0.1° \times 0.1°$.

2. A bias-corrected version of the raw GloFAS forecast above. The bias correction technique was newly developed for this study: firstly, each pixel is corrected using a simple quantile-mapping technique, where the mapping is computed using historical observations and the reanalysis version of GloFAS, GloFAS-ERA5. Secondly, the final streamflow is estimated using an optimised linear combination of streamflow from surrounding pixels. The matrix of coefficients for this linear combination is computed by maximising the sum of the Kling-Gupta and Nash-Sutcliffe efficiencies of the output.

https://doi.org/10.5194/hess-2022-53
Preprint. Discussion started: 8 February 2022
© Author(s) 2022. CC BY 4.0 License.

Hydrology and Earth System Sciences Discussions

3. A type of recurrent neural network, known as a long short-term memory network (LSTM), the development of which was a key focus of this study. The LSTM was trained to ingest catchment-mean meteorological and hydrological variables and output streamflow at six-hourly intervals. Trained using historical ERA5 reanalyses and observations, when run operationally, the LSTM ingested forecasts from the ECMWF Integrated Forecasting System (IFS).

Each of the three models were run for a twelve-month testing period (June 2019 to June 2020), for which they used ERA5 as input, to test how well they could simulate streamflow at the ten stations. Defining skilful as having a KGE greater than -0.414 and highly skilful as having a KGE greater than 0.707, the LSTM performed best (skilful at nine stations, highly skilful at six of these, and with a KGE exceeding 0.9 at three of those), followed by the bias-corrected GloFAS (skilful at seven, highly skilful at four), followed by the raw GloFAS (skilful at six, highly skilful at two). The bias correction improved the KGE of simulated streamflow at seven of the ten stations, implying that it is better at improving the skill of already skilful simulations than adding skill to unskilful ones.

The three models were then run operationally for a thirteen-month period (September 2020 to October 2021), using forecast output from the control member of the IFS as input. Forecast efficiencies were calculated for 2-, 5-, and 8-day lead times. Again, the LSTM performed best, with 5-day forecasts being skilful at nine stations, of which five were highly skilful; followed by the bias-corrected GloFAS, with 5-day forecasts being skilful at seven stations, of which one was highly skilful; and then raw GloFAS, which also had skilful 5-day forecasts at seven stations, one of which was highly skilful, but had lower KGE at six of the seven stations showing skill. Seven of the ten stations (all except BNDN5, CARO2, and LABW4) had catchment areas smaller than 2000 km^2, the recommended lower bound for using GloFAS forecasts. Surprisingly, of these seven, the raw and bias-corrected 5-day GloFAS forecasts were skilful at six (highly skilful at one).

Finally, the three models were compared at all lead times and at all stations against a benchmark persistence model. The LSTM had a higher mean NSE than the persistence model at six of the ten stations – NSE is the preferred evaluation metric here given the dependence of KGE on errors in flow mean and variance, which are zero by definition for a long period of persistence forecasts. Overall, stations with a clearly defined annual cycle and low variance about that cycle were the easiest to predict for all models, whereas stations whose variance was dominated by single storms provided the greatest challenge. The results show a promising future for LSTMs in river streamflow forecasting.

Code availability. All code used to generate, train, and test the models, as well as produce the figures and analysis for this paper, are available in a dedicated GitHub repository: https://github.com/kieranmrhunt/us-streamflow/.

Author contributions. Following the CRediT taxonomy: conceptualization – KMRH and FP; methodology – KMRH; software – KMRH and GRM; validation – KMRH and GRM; formal analysis – KMRH and GRM; writing (original draft, review, and editing) – KMRH, GRM, FP, and CP; visualization – KMRH and GRM; project administration and funding acquisition – FP and CP.

https://doi.org/10.5194/hess-2022-53
Preprint. Discussion started: 8 February 2022
© Author(s) 2022. CC BY 4.0 License.

Hydrology and Earth System Sciences
Discussions
Open Access
EGU

485 *Competing interests.* The authors declare no conflict of interest.

Acknowledgements. KMRH would like to thank Andy Wood (UCAR) and Kenneth Nowack (Bureau of Reclamation) for constructive discussions during the early phase of this project, and Ervin Zsoter (ECMWF) for helping to debug file transfer issues.

https://doi.org/10.5194/hess-2022-53
Preprint. Discussion started: 8 February 2022
© Author(s) 2022. CC BY 4.0 License.

References

Amante, C. and Eakins, B. W.: ETOPO1 1 arc-minute global relief model: procedures, data sources and analysis, Technical Memorandum NESDIS NGDC-24, NOAA, https://doi.org/10.7289/V5C8276M, 2009.

Beven, K. J.: Rainfall-runoff modelling: the primer, John Wiley & Sons, 2011.

Chung, J., Gulcehre, C., Cho, K., and Bengio, Y.: Empirical evaluation of gated recurrent neural networks on sequence modeling, arXiv preprint arXiv:1412.3555, 2014.

Daly, C., Halbleib, M., Smith, J. I., Gibson, W. P., Doggett, M. K., Taylor, G. H., Curtis, J., and Pasteris, P. P.: Physiographically sensitive mapping of climatological temperature and precipitation across the conterminous United States, International Journal of Climatology: a Journal of the Royal Meteorological Society, 28, 2031–2064, 2008.

de Melo, G. A., Sugimoto, D. N., Tasinaffo, P. M., Santos, A. H. M., Cunha, A. M., and Dias, L. A. V.: A new approach to river flow forecasting: LSTM and GRU multivariate models, IEEE Latin America Transactions, 17, 1978–1986, 2019.

Ding, Y., Zhu, Y., Wu, Y., Jun, F., and Cheng, Z.: Spatio-temporal attention LSTM model for flood forecasting, in: 2019 International Conference on Internet of Things (IThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData), pp. 458–465, IEEE, 2019.

Freeze, R. A. and Harlan, R. L.: Blueprint for a physically-based, digitally-simulated hydrologic response model, Journal of Hydrology, 9, 237–258, 1969.

Gauch, M., Kratzert, F., Klotz, D., Nearing, G., Lin, J., and Hochreiter, S.: Rainfall–runoff prediction at multiple timescales with a single Long Short-Term Memory network, Hydrology and Earth System Sciences, 25, 2045–2062, 2021.

Gers, F. A., Schmidhuber, J., and Cummins, F.: Learning to forget: Continual prediction with LSTM, Neural computation, 12, 2451–2471, 2000.

Gupta, H. V., Kling, H., Yilmaz, K. K., and Martinez, G. F.: Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling, Journal of hydrology, 377, 80–91, 2009.

Harrigan, S., Zsoter, E., Alfieri, L., Prudhomme, C., Salamon, P., Wetterhall, F., Barnard, C., Cloke, H., and Pappenberger, F.: GloFAS-ERA5 operational global river discharge reanalysis 1979–present, Earth System Science Data, 12, 2043–2060, 2020.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., et al.: The ERA5 global reanalysis, Quarterly Journal of the Royal Meteorological Society, 146, 1999–2049, 2020.

Hochreiter, S. and Schmidhuber, J.: LSTM can solve hard long time lag problems, Advances in neural information processing systems, pp. 473–479, 1997.

Horton, R. E.: The role of infiltration in the hydrologic cycle, Eos, Transactions American Geophysical Union, 14, 446–460, 1933.

Hu, Y., Yan, L., Hang, T., and Feng, J.: Stream-flow forecasting of small rivers based on LSTM, arXiv preprint arXiv:2001.05681, 2020.

Huffman, G. J., Adler, R. F., Rudolf, B., Schneider, U., and Keehn, P. R.: Global precipitation estimates based on a technique for combining satellite-based estimates, rain gauge analysis, and NWP model precipitation information, J. Climate, 8, 1284–1295, https://doi.org/10.1175/1520-0442(1995)008<1284:GPEBOA>2.0.CO;2, 1995.

Imbeaux, E.: Annales des Ponts et Chaussées, Mémoires et Documents, Vve Ch. Dunod, 1892.

Kling, H., Fuchs, M., and Paulin, M.: Runoff conditions in the upper Danube basin under an ensemble of climate change scenarios, Journal of Hydrology, 424, 264–277, 2012.

https://doi.org/10.5194/hess-2022-53
Preprint. Discussion started: 8 February 2022
© Author(s) 2022. CC BY 4.0 License.

Klotz, D., Kratzert, F., Gauch, M., Keefe Sampson, A., Brandstetter, J., Klambauer, G., Hochreiter, S., and Nearing, G.: Uncertainty Estimation with Deep Learning for Rainfall–Runoff Modelling, Hydrology and Earth System Sciences Discussions, pp. 1–32, 2021.

Knoben, W. J., Freer, J. E., and Woods, R. A.: Inherent benchmark or not? Comparing Nash–Sutcliffe and Kling–Gupta efficiency scores, Hydrology and Earth System Sciences, 23, 4323–4331, 2019.

Kollet, S. J., Maxwell, R. M., Woodward, C. S., Smith, S., Vanderborght, J., Vereecken, H., and Simmer, C.: Proof of concept of regional scale hydrologic simulations at hydrologic resolution utilizing massively parallel computer resources, Water Resources Research, 46, 2010.

Kratzert, F., Klotz, D., Brenner, C., Schulz, K., and Herrnegger, M.: Rainfall–runoff modelling using long short-term memory (LSTM) networks, Hydrology and Earth System Sciences, 22, 6005–6022, 2018.

Kratzert, F., Klotz, D., Herrnegger, M., Sampson, A. K., Hochreiter, S., and Nearing, G. S.: Toward improved predictions in ungauged basins: Exploiting the power of machine learning, Water Resources Research, 55, 11 344–11 354, 2019a.

Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., and Nearing, G.: Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets, Hydrology and Earth System Sciences, 23, 5089–5110, 2019b.

Le, X.-H., Ho, H. V., Lee, G., and Jung, S.: Application of long short-term memory (LSTM) neural network for flood forecasting, Water, 11, 1387, 2019.

Linsley, R. K., Kohler, M. A., and Paulhus, J. L. H.: Applied Hydrology, McGraw-Hill Book Company, 1949.

Mulvaney, T. J.: On the use of self-registering rain and flood gauges in making observations of the relations of rainfall and flood discharges in a given catchment, Proceedings of the institution of Civil Engineers of Ireland, 4, 19–31, 1851.

Nash, J. E. and Sutcliffe, J. V.: River flow forecasting through conceptual models part I—A discussion of principles, Journal of hydrology, 10, 282–290, 1970.

Newman, A. J., Clark, M. P., Sampson, K., Wood, A., Hay, L. E., Bock, A., Viger, R. J., Blodgett, D., Brekke, L., Arnold, J. R., et al.: Development of a large-sample watershed-scale hydrometeorological data set for the contiguous USA: data set characteristics and assessment of regional variability in hydrologic model performance, Hydrology and Earth System Sciences, 19, 209–223, 2015.

Nocedal, J. and Wright, S.: Numerical optimization, Springer Science & Business Media, 2006.

Robock, A., Vinnikov, K. Y., Srinivasan, G., Entin, J. K., Hollinger, S. E., Speranskaya, N. A., Liu, S., and Namkhai, A.: The global soil moisture data bank, Bulletin of the American Meteorological Society, 81, 1281–1300, 2000.

Ross, C. N.: The calculation of flood discharge by the use of time contour plan isochrones, Transactions of the Institution of Engineers, Australia, 2, 85–92, 1921.

Sahoo, B. B., Jha, R., Singh, A., and Kumar, D.: Long short-term memory (LSTM) recurrent neural network for low-flow hydrological time series forecasting, Acta Geophysica, 67, 1471–1481, 2019.

Schiemann, R., Vidale, P. L., Shaffrey, L. C., Johnson, S. J., Roberts, M. J., Demory, M.-E., Mizielinski, M. S., and Strachan, J.: Mean and extreme precipitation over European river basins better simulated in a 25 km AGCM, Hydrology and Earth System Sciences, 22, 3933–3950, 2018.

Schmidhuber, J., Hochreiter, S., et al.: Long short-term memory, Neural Comput, 9, 1735–1780, 1997.

Shen, C. and Lawson, K.: Applications of Deep Learning in Hydrology, in: Deep Learning for the Earth Sciences: A Comprehensive Approach to Remote Sensing, Climate Science, and Geosciences, pp. 283–297, Wiley Online Library, 2021.

Silva, D. F. C., Galvão Filho, A. R., Carvalho, R. V., Ribeiro, F. d. S. L., and Coelho, C. J.: Water Flow Forecasting Based on River Tributaries Using Long Short-Term Memory Ensemble Model, Energies, 14, 7707, 2021.

https://doi.org/10.5194/hess-2022-53
Preprint. Discussion started: 8 February 2022
© Author(s) 2022. CC BY 4.0 License.

Slater, L. J., Anderson, B., Buechel, M., Dadson, S., Han, S., Harrigan, S., Kelder, T., Kowal, K., Lees, T., Matthews, T., et al.: Nonstationary weather and water extremes: a review of methods for their detection, attribution, and management, Hydrology and Earth System Sciences, 25, 3897–3935, 2021.

Sudriani, Y., Ridwansyah, I., and Rustini, H. A.: Long short term memory (LSTM) recurrent neural network (RNN) for discharge level prediction and forecast in Cimandiri river, Indonesia, in: IOP Conference Series: Earth and Environmental Science, vol. 299, p. 012037, IOP Publishing, 2019.

Wood, E. F., Roundy, J. K., Troy, T. J., Van Beek, L. P. H., Bierkens, M. F. P., Blyth, E., de Roo, A., Döll, P., Ek, M., Famiglietti, J., et al.: Hyperresolution global land surface modeling: Meeting a grand challenge for monitoring Earth's terrestrial water, Water Resources Research, 47, 2011.

Zhu, S., Luo, X., Yuan, X., and Xu, Z.: An improved long short-term memory network for streamflow forecasting in the upper Yangtze River, Stochastic Environmental Research and Risk Assessment, 34, 1313–1329, 2020.

Zsoter, E., Harrigan, S., Wetterhall, G., Salamon, P., and Prudhomme, C.: River discharge and related forecasted data from the Global Flood Awareness System, v2.1, Copernicus Climate Change Service (C3S) Climate Data Store (CDS), https://doi.org/10.24381/cds.ff1aef77, 2019a.

Zsoter, E., Harrigan, S., Wetterhall, G., Salamon, P., and Prudhomme, C.: River discharge and related forecasted data from the Global Flood Awareness System, v2.1, Copernicus Climate Change Service (C3S) Climate Data Store (CDS), https://doi.org/10.24381/cds.a4fdd6b9, 2019b.

Zsoter, E., Prudhomme, C., Stephens, E., Pappenberger, F., and Cloke, H.: Using ensemble reforecasts to generate flood thresholds for improved global flood forecasting, Journal of Flood Risk Management, 13, e12 658, https://doi.org/https://doi.org/10.1111/jfr3.12658, 2020.

Zsoter, E., Harrigan, S., Barnard, C., Wetterhall, G., Ferrario, I., Mazzetti, C., Alfieri, L., Salamon, P., and Prudhomme, C.: River discharge and related forecasted data from the Global Flood Awareness System, v3.1, Copernicus Climate Change Service (C3S) Climate Data Store (CDS), https://doi.org/10.24381/cds.ff1aef77, 2021.