Comparing machine learning and deep learning models for probabilistic post-processing of satellite precipitation-driven streamflow simulation

Yuhang Zhang[1], Aizhong Ye[1], Bita Analui[2], Phu Nguyen[2], Soroosh Sorooshian[2], Kuolin Hsu[2], Yuxuan Wang[3]

[1]State Key Laboratory of Earth Surface Processes and Resource Ecology, Faculty of Geographical Science, Beijing Normal University, Beijing 100875, China
[2]Center for Hydrometeorology and Remote Sensing, Department of Civil and Environmental Engineering, University of California, Irvine, Irvine, California, CA 92697, USA
[3]College of Arts and Sciences, University of Virginia, Charlottesville, Virginia, 22903, USA

Correspondence to: Aizhong Ye (azye@bnu.edu.cn)

Abstract. Deep learning (DL) and machine learning (ML) are widely used in hydrological post-processing, which plays a critical role in improving the accuracy of hydrological predictions. However, the trade-off between model performance and computational cost has always been a challenge for hydrologists when selecting a suitable model, particularly for probabilistic post-processing with large ensemble members. Moreover, it is unclear whether the performance differences between DL and ML models is significant in hydrological probabilistic post-processing. Therefore, this study aims to systematically compare the quantile regression forest (QRF) model and countable mixtures of asymmetric Laplacians long short-term memory (CMAL-LSTM) model as hydrological probabilistic post-processors. Specifically, we evaluate their ability in dealing with biased streamflow simulation driven by three satellite precipitation products across 522 sub-basins of Yalong River basin in China. Model performance is comprehensively assessed using a series of scoring metrics from both probabilistic and deterministic perspectives. Our results show that the QRF model and the CMAL-LSTM model are comparable in terms of probabilistic prediction, and their performance is closely related to the flow accumulation area (FAA) of the sub-basin. The QRF model outperforms the CMAL-LSTM model in most of the sub-basins with smaller FAA, while the CMAL-LSTM model has an undebatable advantage in sub-basins with FAA larger than 60,000 km^2 in Yalong River basin. In terms of deterministic predictions, the CMAL-LSTM model is preferred, especially when the raw streamflow is poorly simulated and used as an input. However, if we put aside the differences in model performance, the QRF model is more efficient than the CMAL-LSTM model in computation time in all experiments. As a result, this study provides insights into model selection in hydrological post-processing and the trade-offs between model performance and computational efficiency. The findings highlight the importance of considering the specific application scenario, such as the catchment size and the required accuracy level, when selecting a suitable model for hydrological post-processing.

Key words: Bias correction, long-short memory network, quantile regression forest, satellite precipitation, streamflow simulation.

1 Introduction

By generalizing the physical processes, hydrologists or modelers abstract the hydrological mechanism into a series of numerical equations, collectively known as hydrological models (Sittner et al., 1969; Clark et al., 2015; Sivapalan, 2018; Chawanda et al., 2020; Zhou et al., 2021). Hydrological models are widely used for rainfall-runoff simulation, flood forecasting, drought assessment, decision making, and water resource management (Corzo Perez et al., 2011; Tan et al., 2020; Wu et al., 2020; Gou et al., 2020,2021; Miao et al., 2022). Depending on the complexity, hydrological models can be classified as lumped, semi-distributed, and distributed models (Beven, 1989; Jajarmizadeh et al., 2012; Khakbaz et al., 2012; Mai et al., 2022). Although current models simulate the hydrological processes well, they still suffer from multiple uncertainties, including input uncertainty, model structure and parameter uncertainty, and observation uncertainty (Nearing et al., 2016; Herrera et al., 2022). These uncertainties limit the accuracy of hydrological models (Honti et al., 2014; Sordo-Ward et al., 2016; Mai et al., 2022). Among these various sources, input uncertainty is considered to be one of the largest sources of uncertainty. Hence, precipitation, which is the driver of the water cycle, is the most important factor affecting streamflow simulation (Kobold and Sušelj, 2005).

Precipitation information is mainly derived from gauge observations, radar precipitation estimates, satellite precipitation retrievals and reanalysis products (Sun et al., 2018). Gauge stations and radar are limited by the density of the station network and the topography, especially in remote areas such as mountainous regions and high altitudes (Sun et al., 2018; Chen et al., 2020). Reanalysis requires the assimilation of observations from multiple sources and therefore cannot be obtained in real time. Satellite precipitation estimates are available in near-real-time and have shown valuable potentials for applications in regions where ground measurements are scarce. (Jiang and Bauer-Gottwein, 2019; Dembélé et al., 2020). In the past decades, several research institutions have developed various satellite precipitation estimation products with different data sources and algorithms, such as the Integrated Multi-satellitE Retrievals for Global Precipitation Measurement Mission (GPM IMERG) products jointly developed by the National Aeronautics and Space Administration (NASA) and the Japan Aerospace Exploration Agency (JAXA) (Hou et al., 2013; Huffman et al., 2015), the Global Satellite Mapping of Precipitation (GSMaP) products developed by JAXA (Kubota et al., 2007, 2020), and the Precipitation Estimate from Remotely Sensed Information using Artificial Neural Networks-Dynamic Infrared Rain Rate near real-time (PDIR-Now, hereafter, PDIR) product developed by the Centre for Hydrometeorology and Remote Sensing (CHRS) at the University of California, Irvine (UCI) (Nguyen et al., 2020a, 2020b). However, there are still uncertainties in these products due to factors such as data sources and algorithms (Tian et al., 2009; Zhang et al., 2021a). And, the uncertainties are even amplified during the hydrological simulation (Cunha et al., 2012; Falck et al., 2015; Zhang et al., 2021b) and severely limits their capability for meteorological and downstream hydrological applications.

The current study addresses the uncertainty of satellite precipitation as input in hydrological modelling in two ways, namely, pre-processing and post-processing (Wang et al., 2009; Ye et al., 2015; Li et al., 2017; Dong et al., 2020; Shen et al., 2021; Zhang et al., 2022a). Here, we use the terminology of the hydrologic ensemble prediction experiment (HEPEX), where pre- and post-processing are distinguished before and after using the hydrological model (Schaake et al., 2007). That is, precipitation input to the hydrological model and hydrological streamflow output are processed separately (Li et al., 2017). Hydrological pre-processing, also known as precipitation post-processing, is commonly used to obtain bias-corrected precipitation estimates by directly bias-correcting or fusing satellite precipitation estimates and gauge observations (Xu et al., 2020; Zhang et al., 2022a). Pre-processing mainly reduces precipitation input uncertainty. Hydrological post-processing mainly uses the observed streamflow to correct the streamflow simulation or prediction (Ye et al., 2014; Tyralis et al., 2019). Hydrological post-processing not only reduces the effect of input uncertainty, or further reduces input uncertainty after hydrological pre-processing, but also reduces uncertainty caused by hydrological model structure and model parameters (Parrish et al., 2012; Kaune et al., 2020). Both hydrological pre-processing and post-processing can be used to generate deterministic and probabilistic predictions.. Our objective in this study is to compare and evaluate learning algorithms in probabilistic hydrological post-processing.

In addition to the skewed distribution and the heteroscedasticity, the streamflow time series have a strong autocorrelation (Herrera et al., 2022). According to this feature, there are two main types of methods used to perform hydrological post-processing. One is the autoregressive model based on residuals. Its main idea is to use the simulation residuals as forecast factors for the error update. Typical methods are error reduction models based on autoregression (Li et al., 2015, 2016; Zhang et al., 2018). Another way is to use the idea of model output statistics (MOS) (Wang et al., 2009; Bogner and Pappenberger, 2011; Bellier et al., 2018). That is, the simulated streamflow is used directly as a forecast factor to establish statistical relationships between simulations and observations. A representative approach of this type is the general linear model post-processor (GLMPP) (Zhao et al., 2011).

In recent years, machine learning (ML) and deep learning (DL) algorithms have become powerful tools for hydrological modelling (Sit et al., 2020; Zounemat-Kermani et al., 2021; Shen and Lawson, 2021; Fang et al., 2022). For example, long short-term memory (LSTM) models have been used to simulate streamflow in a number of gauged and ungauged basins in North America (Kratzert et al., 2018, 2019), the United Kingdom (Lees et al., 2021), and Europe (Nasreen et al., 2022). In addition to direct streamflow modelling, ML and DL algorithms can also be used as powerful hydrological post-processors for bias correction of streamflow simulation. For example, Frame et al. (2021) used LSTM to build a post-processor to correct the U.S. National Hydrologic Model and validated it on the CAMELS (Catchment Attributes and Meteorology for Large-sample Studies) dataset containing 531 North American watersheds. The results showed that the LSTM post-processing significantly enhanced the output of the raw national hydrological model. Shen et al. (2022) used the random forest as a hydrological post-processor to enhance the simulation performance of the large-scale hydrological model PCR-GLOBAL (PCRaster Global Water Balance) model at three hydrological stations in the Rhine basin. Compared to deterministic forecasts, probabilistic forecasts can provide more insights regarding the uncertainties and improve the risk management strategies. In terms of probabilistic modelling, Tyralis et al. (2019) compared the usability of the statistical model (e.g., quantile regression) and the machine learning algorithm (e.g., quantile regression forests) as hydrological post-processors on the CAMELS dataset. And the results showed that the quantile regression forests model outperformed the quantile regression. Zhu et al. (2020) investigated the applicability of LSTM for probabilistic hydrological forecasting coupled with a Gaussian process model. Similarly, Althoff et al. (2021) quantified the uncertainty of LSTM for hydrological modelling using stochastic deactivation of neurons. Li et al. (2021, 2022) quantified the uncertainty of LSTM for hydrological modelling using variational inference from a Bayesian perspective. All these individual models can quantify the uncertainty. More recently, Klotz et al. (2022) compared the use of dropout and three Gaussian mixture distribution models for uncertainty estimation in LSTM rainfall-runoff modelling. They found that the mixture density model outperformed the random dropout model and provided more reliable probabilistic information. Both ML models and DL models have been successfully practiced in hydrological probabilistic post-processing. In addition, there has been some scholarly work in which DL models have been compared and analysed (e.g., Klotz et al., 2022). However, to our knowledge there has not been a comparison between ML and DL models for hydrological probabilistic post-processing in the literature. DL models, despite their powerful predictive capabilities, are often criticized for their higher computational complexities and costs. ML models, on the other hand, are more efficient and easier to implement but may perform poorly in comparison. Notwithstanding these evidences, their differences in the field of hydrological probabilistic post-processing, such as the scope of application, model performance and computational efficiency is not well studied.

Therefore, in this study, we attempt to comprehensively compare the performance of the two most widely used ML and DL models for streamflow probabilistic post-processing: quantile regression forests (QRF) and countable mixtures of asymmetric Laplacians LSTM (CMAL-LSTM), at a sub-basin scale daily streamflow, respectively. In particular, a full model comparison is performed in a complex basin with 522 nested sub-basins in southwest China. Three sets of global satellite precipitation products are applied to generate uncorrected streamflow simulations. The three precipitation products represent different algorithms. Also, they have been proven to have relatively good accuracy in our previous study (Zhang et al., 2021b). They are also used for single-feature and multi-feature input analysis. A variety of evaluation metrics are used to assess the performance of the proposed models, including probabilistic metrics for multi-point prediction and deterministic metrics for single-point prediction. The relationship between model performance and basin size is also analysed according to the difference in the flow accumulation area of the sub-basin. This study can deepen our understanding of ML and DL models, and enable targeted model selection in practice. The paper is organized as follows: In Sect.2, we introduce the study area and data. In Sect.3, we present the post-processing models, experimental design and evaluation metrics. Sect. 4 presents the streamflow results before and after post-processing with different experiments. In Sect. 5, we discuss the interpretation of post-processing model differences, as well as their limitations. Finally, the conclusions are summarized at the end of this article.

2 Study area and Data

2.1 Study area

The Yalong River (Fig. 1a) is a major tributary of the Jinsha River, which belongs to the upper reaches of the Yangtze River. The Yalong River basin is located between the Qinghai-Tibet Plateau and the Sichuan Basin. The Yalong River basin has a long and narrow shape (96° 52'–102° 48' E, 26° 32'–33° 58' N), with snow-capped mountains scattered in the upper reaches, surrounded by high mountain valleys in the middle reaches, and flowing into the Jinsha River in the lower reaches. It spans seven dimensional zones with complex climate types. The total length of the basin is about 1,570 km, and the total area is about 130,000 km^2. The mean annual precipitation of the basin is about 800 mm.

Following the watershed division method of Du et al. (2017), Yalong River basin is divided into 522 sub-basins with catchment areas ranging from 100 km^2 to 127,164 km^2 (Fig. 1b). The key to sub-basin delineation is the minimum catchment area threshold (100 km^2 in this study), which is related to the total area of the basin, the model architecture complexity, the step size and the spatial resolution of the input data. Location, elevation, area, flow accumulation area and flow direction of each sub-basin can be found in Table S1.

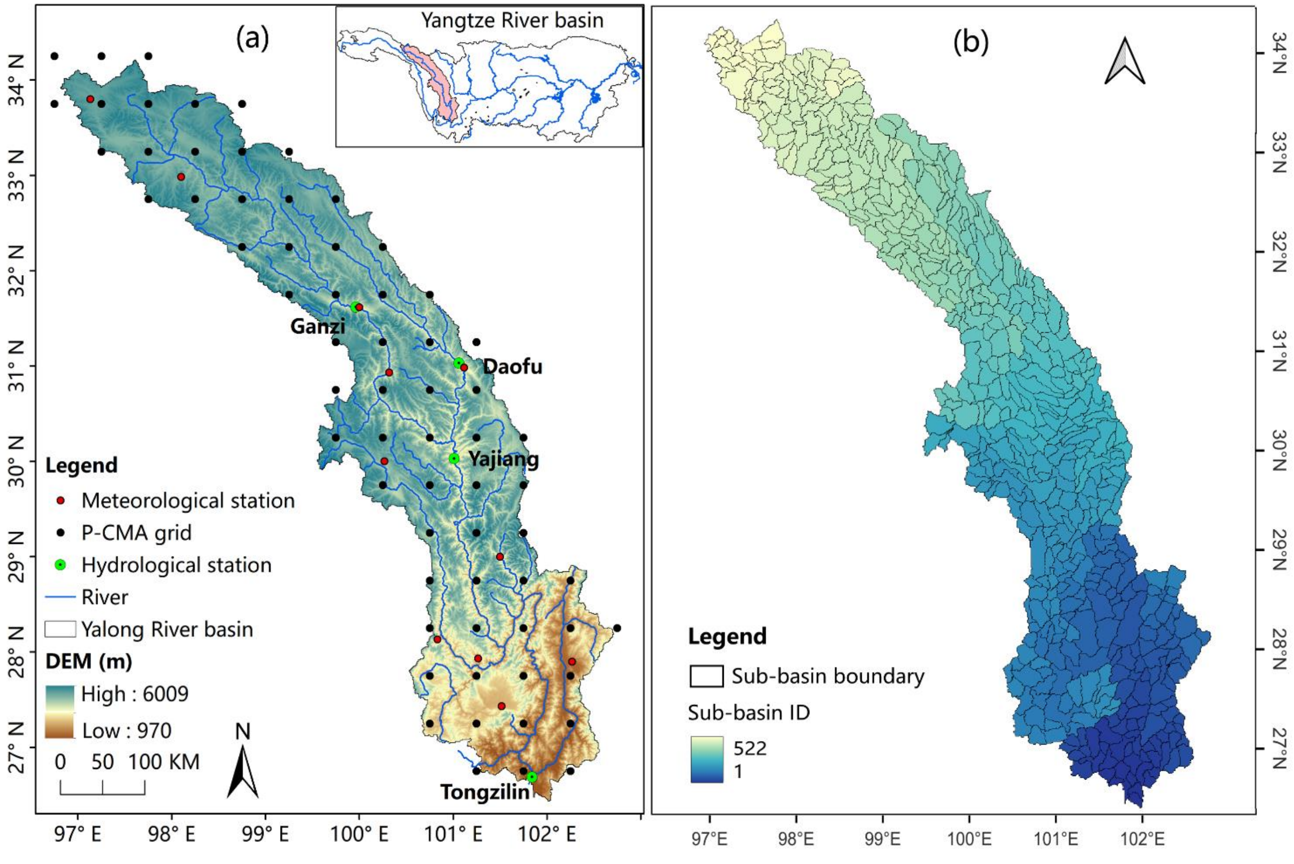

Figure 1. (a) Study area and (b) 522 sub-basins (Zhang et al., 2022a).

2.2 Data

2.2.1 Gauge precipitation observations

The 0.5-degree, daily precipitation observation data were obtained from the National Meteorological Information Centre of the China Meteorological Administration (CMA-NMIC). The product was produced by interpolating gauge data from more than 2000 stations across China. This product has been proven to be highly accurate and has been widely applied to a variety of studies such as streamflow simulation, drought assessment, and water resource management (Gou et al., 2020, 2021; Zhang and Ye, 2021; Miao et al., 2022). In this study, the gridded precipitation observations are used as a reference for the satellite-based precipitation products. Using the inverse distance weighting (IDW) method, they are resampled to each sub-basin. And due to limited hydrological observatories, the streamflow of each sub-basin obtained from the calibrated hydrological model driven by this product is also used as a reference for the satellite precipitation-driven streamflow simulation. Errors caused by factors such as interpolation are ignored. The selected study period is from January 1, 2003 to December 31, 2018.

2.2.2 Global satellite precipitation estimates

Three sets of the latest quasi-global satellite precipitation estimation products are selected. The first one is PDIR product, which solely relies on infrared data. It has a very high spatiotemporal resolution (0.04 degree and 1 hour) and a very short delay time (1 hour). The other two products are bias-adjusted products, IMERG Final Run version 6 (hereafter, IMERG-F) (Huffman et al., 2015, 2019) and Gauge-calibrated GSMaP product (GSMaP_Gauge_NRT_v6, hereafter, GSMaP) (Kubota et al., 2007, 2020), with a spatial resolution of 0.1 degree. The selected study period is also from January 1, 2003 to December 31, 2018. All these products are aggregated to the daily scale and resampled to each sub-basin using IDW. It should be noted that these products are selected as examples only and any other precipitation product can be used as an alternative.

2.2.3 Other data

In addition to precipitation gauge observations and satellite precipitation products, other meteorological data such as: temperature, wind speed and evaporation basin attributes, and streamflow observations are needed for hydrological modelling. The meteorological data were also obtained from the CMA-NMIC, and were used to drive the hydrological model together with precipitation. The streamflow observations (January 1, 2006 to December 31, 2015) were collected from four gauged hydrological stations in the Yalong River basin from the upstream to the downstream, namely Ganzi (GZ), Daofu (DF), Yajiang (YJ), and Tongzilin (TZL) (cf. Figure 1(a)). These data were obtained from the Hydrological Yearbook of the Bureau of Hydrology. The National Aeronautics and Space Administration Shuttle Radar Topographic Mission (NASA SRTM) digital elevation model (DEM) data with a spatial resolution of 90m was obtained from the Geospatial Data Cloud of China. The 1 km soil data was clipped from the China Soil Database issued by the Tibetan Plateau Data Centre of China. The 1km land use data was obtained from the Resource and Environment Science and Data Centre provided by the Institute of Geographical Sciences and Resources, Chinese Academy of Sciences.

3 Methodology

The framework of this study is shown in Fig. 2. We adopt a two-stage streamflow post-processing approach. In the first stage (Sect. 3.1), the hydrological model is calibrated and validated by hydrological station observations. Then, we use the observed precipitation to drive the calibrated hydrological model to generate streamflow references for each sub-basin. And we use satellite precipitation to drive the model to generate uncorrected (raw) streamflow simulations. In the second stage (Sect. 3.2), we perform probabilistic post-processing of the streamflow using the QRF and the CMAL-LSTM models. In the last subsection (Sect. 3.3), we describe the evaluation metrics used in this study.

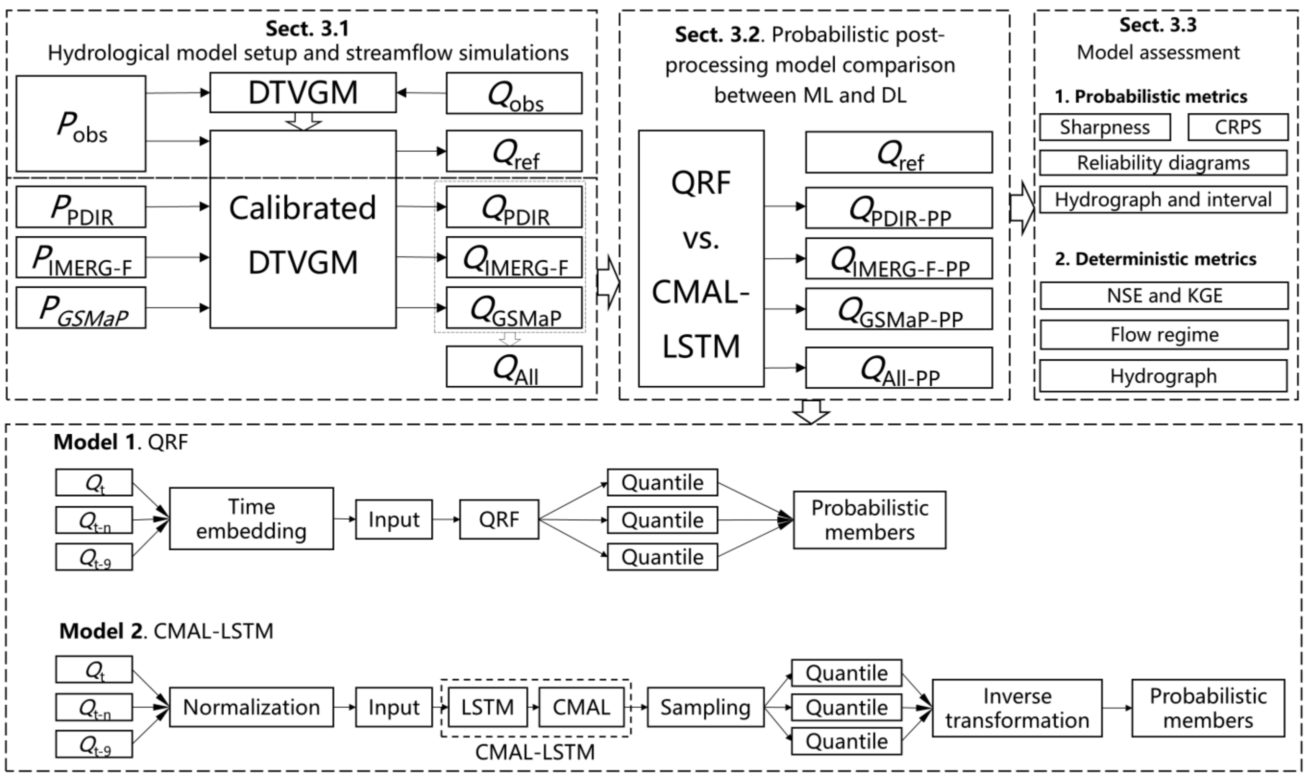

Figure 2. Framework of this study.

3.1 Streamflow reference and uncorrected streamflow simulations

The purpose of this study is to post-process the streamflow simulations for all sub-basin outlets, and therefore corresponding references are needed. Due to the limited streamflow observations, we use streamflow simulations from the hydrological model driven by observed precipitation as a reference. To ensure that the results are reliable, we first use the collected streamflow observations from four hydrological stations to setup, calibrate and validate the hydrological model.

We choose distributed time-variant gain model (DTVGM), a process-based model that uses the rainfall-runoff nonlinear relationship (Xia, 1991; Xia et al., 2005) for simulation. In each sub-basin, runoff is calculated according to Eq. (1). The kinematic wave equation is used for river routing (Ye et al., 2013). The snowmelt process in the high-altitude regions of the basin is simulated by the degree-day method (Bormann et al., 2014). A detailed description of the model can be found in (Xia et al., 2005; Ye et al., 2010).

$$P_t + AW_t = AW_{t+1} + g_1(\frac{AW_{u,t}}{C \cdot WM_u})^{g_2} \cdot P_t + K_r \cdot AW_{u,t} + K_e \cdot EP_t + K_g \cdot AW_{g,t} \quad (1)$$

where t is the time step; P and EP are precipitation and potential evapotranspiration, respectively; AW and WM are soil moisture (mm) and field soil moisture (mm), respectively; u and g are the upper and lower soil layers, respectively; K_e, K_r and K_g are evapotranspiration, interflow and groundwater runoff coefficients, respectively; g_1 and g_2 are factors describing the non-linear rainfall-runoff relationship; and C is the land cover parameter.

Based on the length of the streamflow observation collected from hydrological stations (2006-2014), we divide the streamflow time series into three periods: a one-year spin-up period (2006), a four-year calibration period (2007-2010), and a four-year validation period (2011-2014). We use Nash-Sutcliffe efficiency (NSE) as the objective and regionalize the parameters from upstream to downstream using manual tuning, while ensuring that the water balance coefficient converges to 1. The model calibration and validation are shown in Fig. S1 in the supplement. The NSE for the four gauged hydrological stations (GZ, DF, YJ, and TZL) are 0.89, 0.91, 0.93, 0.79, and 0.79, 0.86, 0.87, and 0.59 for calibration and validation periods, respectively. In the remaining part of this study, the hydrological model is fixed and we mainly post-process the streamflow bias introduced by satellite precipitation, disregarding other sources of uncertainty such as model structure, DEM and other forcing data.

After model calibration and validation, to ensure the number of data samples for data-driven post-processing methods, we use the observed precipitation from 2003 to 2018 to drive the hydrological model. A 16-year streamflow simulation reference data for 522 sub-basin outlets is obtained. Streamflow from different sub-basins can also reflect hydrological processes of diverse climate types and scales.

Finally, we use three different satellite precipitation products PDIR, IMERG-F and GSMaP for period 2003-2018 to separately drive the hydrological model and obtain three different streamflow simulations accordingly. In addition, the equally weighted average of the three outputs can be viewed as a multi-product driven simulation (All). The reason for considering the multi-product simulation (All) for reference is two-fold. First, the model performance of two different post-processing models can be adequately compared in three different contexts. Secondly, the multiple inputs can be used to compare the effects of model averaging and multi-dimensional features on the post-processing models. The experimental design is described in the following Sect. 3.2.

3.2 Post-processing model and experimental design

The two post-processing models selected are the QRF model (Meinshausen and Ridgeway, 2006) and the CMAL-LSTM model (Klotz et al., 2022). The QRF model was chosen because it enables us to analyse the distribution of the entire data based on different quantiles, and it has been previously used in several studies (Taillardat et al., 2016; Evin et al., 2021; Kasraei et al., 2021; Tyralis et al., 2019; Tyralis and Papacharalampous, 2021). The CMAL-LSTM model is a combination of an LSTM model and a CMAL mixture density function, which allows it to estimate prediction uncertainty. To the best of our knowledge, these two models currently considered state-of-the-art in ML and DL for hydrological probabilistic modelling. Avid readers are highly encouraged to read the original papers for more detailed information about each model.

Both post-processing models require input features. Here, in order to keep the complexity of the models low, only the uncorrected streamflow simulations are chosen as input features. Considering the autocorrelation skill of the streamflow (see Fig. S2 in the supplement), for the post-processing (Q_t) on day t, we select the simulated streamflow for the first 9 days (Q_{t-9}^{sim}, Q_{t-8}^{sim},…, Q_{t-1}^{sim}) and the simulated streamflow of that day (Q_t^{sim}) as the inputs. In the QRF model, the input features are fed by temporal embedding. And in the CMAL-LSTM model, the seq-length is set to 9. For both models, we select the streamflow reference (Q_t^{ref}) on day t as the target. In addition, since we used three different satellite precipitation products, the experiments are divided into a single-product experiment and a multi-product experiment (All). The information for each experiment is summarized in Table 1. The training period is from 1 January 2003 to 31 December 2010. The validation period is from 1 January 2011 to 31 December 2014. And the test period is from 1 January 2015 to 31 December 2018.

Table 1. Experimental design.

Streamflow simulation	Model	Input feature		Target
PDIR	QRF	$Q_{t-9}^{sim}, Q_{t-8}^{sim},\ldots, Q_t^{sim}$	10	Q_t^{ref}
	CMAL-LSTM		1	
IMERG-F	QRF		10	
	CMAL-LSTM		1	
GSMaP	QRF		10	
	CMAL-LSTM		1	
All (PDIR, IMERG-F, GSMaP)	QRF		30	
	CMAL-LSTM		3	

We implemented the QRF model using *pyquantrf* package (Jnelson18, 2022). We tuned three sensitive hyperparameters in the QRF model by grid search, finally setting the number of trees (K) to 70, the number of non-leaf node splitting features to 10, and the number of samples used for leaf node predictions (N_{leaf}) to 10. All other hyperparameters were set to default values.

We implemented the CMAL-LSTM model using *NeuralHydrology* package (Kratzert et al., 2022a). We followed the model architecture of Klotz et al. (2022), which contains an LSTM layer and a CMAL layer. In contrast to the QRF model, the input data of the CMAL-LSTM model needs to be normalized. Here, by several comparisons, we used the normal quantile transform method (Fig. S3 in the supplement). The hyperparameters of the model include the number of neurons in the LSTM layer (N_{LSTM}), the number of components of the mixture density function (N_{MDN}), the dropout rate, the learning rate, the epoch, and the batch size. N_{MDN}, is set to 3, which follows Klotz et al. (2022). The other hyperparameters are also fine-tuned such that the final learning rate is set to 0.0001, the dropout to 0.4, the epoch to 100, the batch size to 256, and the N_{LSTM} to 256.

For the QRF model, 100 percentiles (0.005 to 0.995) were equally sampled for each basin and time step and fed directly into the model to obtain the final probabilistic (100) members. For the CMAL-LSTM model, first 10,000 sample points for each basin and time step by sampling from the mixture distribution were generated and the same 100 percentiles (0.005 to 0.995) from these sample points were extracted and remapped to the original streamflow space using inverse quantile normal transformation, where finally the probabilistic members were produced.

Our computing platform is a workstation configured with an Intel(R) Xeon(R) Gold 6226R CPU @ 2.9GHz and an RTX3090 GPU with 24G video memory. It is worth noting that we modelled each sub-basin separately due to the random sampling process of the CMAL-LSTM model exceeding the GPU's video memory. For consistency, the QRF model was also modelled locally. The computational time was approximately 12 hours to complete all CMAL-LSTM and 6 hours to complete all QRF experiments.

3.3 Performance evaluation

In this section two post-processing models are evaluated from both probabilistic and deterministic perspectives. The evaluation metrics are presented in Sect. 3.3.1 and Sect. 3.3.2, respectively.

3.3.1 Probabilistic (multi-point) metrics

We followed the criterion for probabilistic predictions proposed by Gneiting et al. (2007): to maximize the sharpness of the prediction distributions subject to reliability. We both use scoring rules and diagnostic graphs to assess reliability and sharpness holistically.

The continuous rank probability score (CRPS) is a widely used proper scoring rule that assesses reliability and sharpness simultaneously (Gneiting et al., 2007). For given probabilistic members, the CRPS calculates the difference between the cumulative distribution function (CDF) of the probabilistic members and the observations. We also used a weighted version of CRPS (threshold weighted CRPS, twCPRS), which is commonly used to give more weight to extreme cases (Gneiting and Ranjan, 2011). These two metrics can be expressed as follows:

$$CRPS(F,x) = \int_{-\infty}^{\infty} \{F(y) - \mathbf{1}(y \geq x)\}^2 dy \tag{2}$$

$$twCRPS(F,x) = \int_{-\infty}^{\infty} \{F(y) - \mathbf{1}(y \geq x)\}^2 \omega(y) dy \tag{3}$$

where $\omega(y)$ is a threshold weighted function and is calculated based on the threshold q (80%, 90% and 95% percentiles of observations in this study). When $y \geq q$ ($y < q$), $\omega(y)$ equals 1 (0). $F(y)$ is the CDF obtained from the probabilistic members for the corrected streamflow; $\mathbf{1}(y \geq x)$ is the Heaviside function. The better performing model has both metrics ($CRPS$ and $twCRPS$) closer to 0.

The CRPS skill score ($CRPSS$) is also used to define the relative differences between the two post-processing models. For QRF and CMAL-LSTM, the $CRPSS$ can be calculated as:

$$CRPSS_{QRF/PLSTM} = \left(1 - \frac{CRPS_{QRF}}{CRPS_{PLSTM}}\right) \times 100\% \tag{4}$$

A $CRPSS$ greater than 0 indicates that the QRF model is better than the CMAL-LSTM model, and vice versa.

The reliability diagram is used as diagnostic graph to assess the agreement between the predicted probability and the observed frequency (Jolliffe and Stephenson, 2003). Namely, if the predicted probability of a particular event is 30%, then the observed relative frequency should also be 30%. Ultimately, perfectly reliable predictions at multiple levels of probability result in a distribution along the diagonal line corresponding to the same levels of observed frequency. A point above (below) the diagonal line in the reliability diagram indicates that the observed relative frequency is higher (lower) than the predicted probability and that there is an underprediction (overprediction) phenomenon. Here again, three thresholds (80%, 90% and 95%) are chosen to better evaluate the reliability of extreme cases (Yang et al., 2021).

Sharpness is a fundamental characteristic of predictive distributions. A sharp probabilistic output corresponds to a low degree of variability in the predictive distribution. To evaluate the sharpness of probabilistic predictions, prediction intervals are commonly employed (Gneiting et al., 2007). For this study, the 50% and 90% percentile intervals were chosen. Furthermore, to establish the relationships between predictive distributions and observations, we assessed the coverage of the prediction intervals over the observations. The average Euclidean distance of the 25% and 75% probabilistic members is adopted as the sharpness metric ($DIS_{25\text{-}75}$) for the 50% prediction interval, and the 5% and 95% probabilistic members were used to compute the sharpness metric ($DIS_{5\text{-}95}$) for the 90% prediction intervals. The ratio of the number of observations in the prediction intervals to the total number of observations was used as the coverage of observations ($CO_{25\text{-}75}$ and $CO_{5\text{-}95}$). In addition, three additional metrics used in a previous study (Klotz et al., 2022) are also employed to calculate the sharpness metric for the full probabilistic members, including mean absolute deviation (MAD), standard deviation (STD) and variance (VAR).

3.4.2 Deterministic (single-point) metrics

The widely used Nash-Sutcliffe efficiency (NSE) (Nash and Sutcliffe, 1970) and Kling-Gupta efficiency (KGE) (Gupta et al., 2009) are applied for assessing the deterministic model performance. In addition, two components of NSE, namely Pearson correlation coefficient (PCC) and relative bias (RB) are calculated to assess the temporal consistency and systematic bias of the difference between simulations and observations, respectively. Furthermore, to account for the seasonality of the flow regime, four metrics are selected to characterize the different aspects of flow regimes, including the peak flow bias (FHV, Eq. (A3) in Yilmaz et al., 2008), low-flow bias (FLV, Eq. (A4) in Yilmaz et al., 2008), flow duration curve bias (FMS, Eq. (A2) in Yilmaz et al., 2008), and mean peak time lag bias (in days) (PT, Appendix D in Kratzert et al., 2021). These metrics provide a comprehensive assessment of model performance across different flow conditions and facilitate a more accurate evaluation of model ability to reproduce the hydrological processes.

4 Results

4.1 Uncorrected streamflow simulations

Figure 3 shows the spatial distribution of NSE for streamflow simulations in 522 sub-basins, driven by three different satellite precipitation products and multi-product outputs using equal-weighting averaging (All). Among the three satellite precipitation products, the IMERG-F achieves the best model performance, followed by PDIR and GSMaP. PDIR performs poorly in the upstream and outlet regions of the basin. GSMaP exhibits significant deviations from the streamflow reference in almost all sub-basins. The precipitation product quality plays a crucial role in streamflow performance with the same hydrological model configuration. The high precipitation bias in GSMaP (Fig. S4f in the supplement) leads to high biases in streamflow simulations (Fig. 8b), resulting in the lowest NSE values (Fig. 3c and Fig.8c) among the three products. The performance of PDIR-driven streamflow is mainly influenced by the poor temporal variability (PCC) against observations (Fig. S4a in the supplement and Fig. 8a). Equal-weighting averaging (All) that incorporates biased information from PDIR and GSMaP has little impact on improving model performance.

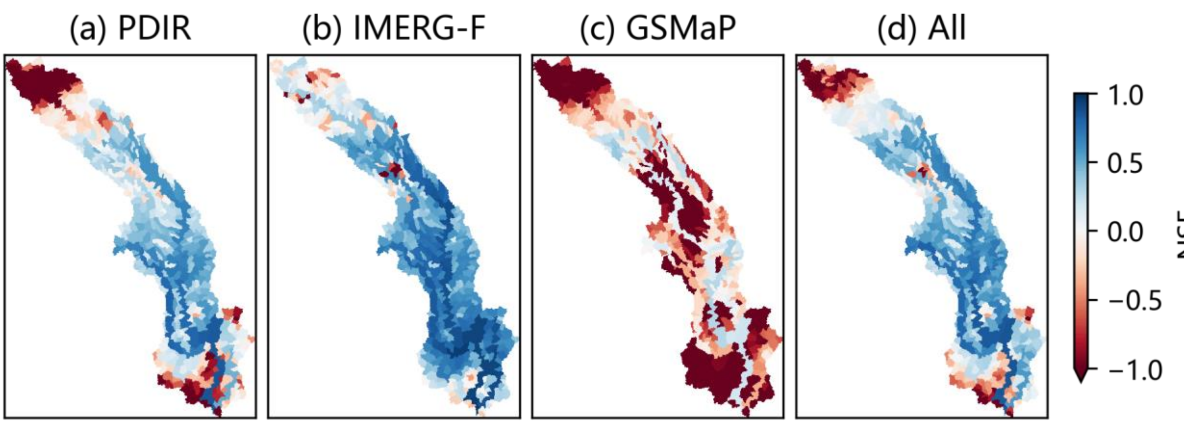

Figure 3. The NSE of uncorrected streamflow simulation for the 522 sub-basins.

4.2 Probabilistic (multi-point) assessment

The flow magnitudes in different sub-basins vary widely. Therefore, in the presented results for each sub-basin the results are normalized separately according to the probabilistic membership of all experiments. By doing so, the probabilistic members of all sub-basins are mapped to the range between 0 and 1.

4.2.1 CRPS overall performance

Overall, the QRF and CMAL-LSTM models demonstrate similar performance in terms of CRPS and twCRPS across all threshold conditions (as shown in Fig. 4 and Fig. S5). However, it is noteworthy that the QRF model exhibits more outliers compared with the CMAL-LSTM model, indicating that the latter is more stable across sub-basins. When it comes to different precipitation-driven streamflow inputs, the IMERG-F-QRF and IMERG-F-CMAL-LSTM experiments have median CRPS values of 0.0197 and 0.0199, respectively, for 522 sub-basins; the GSMaP-QRF and GSMaP-CMAL-LSTM experiments have median CRPS values of 0.024 and 0.0241, respectively; the PDIR-QRF and PDIR-CMAL-LSTM experiments have median CRPS values of 0.0287 and 0.0292, respectively. The results show that IMERG-F performs better than GSMaP, and both bias-corrected products outperform the near real-time product PDIR in post-processing performance. The results of the multi-product approach (All) are close to those of IMERG-F, but better than those of PDIR and GSMaP. As the threshold conditions increase, the performance of the multi-product approach is slightly worse than that of IMERG-F (Fig. S5). This suggests that introducing features that perform well in a model, such as IMERG-F driven raw streamflow, can improve the performance of post-processing models, but introducing features that perform poorly, such as GSMaP and PDIR driven raw streamflow, can worsen the performance of post-processing model. The results indicate that the QRF and CMAL-LSTM models can automatically perform feature filtration, but cannot completely avoid learning from disruptive information. Using IMERG-F driven raw streamflow as input, the post-processing models perform better than when driven by the other two products as input features, which is related to the quality of IMERG-F features. In terms of temporal correlation and bias, IMERG-F is the optimal product. The raw streamflow simulation of GSMaP performs worse than PDIR, but the post-processing model performs better than PDIR, because the raw streamflow of GSMaP has higher temporal correlation and better autocorrelation skill as input features compared to PDIR. This leads to PDIR being the worst-performing post-processing experiment among the selected datasets.

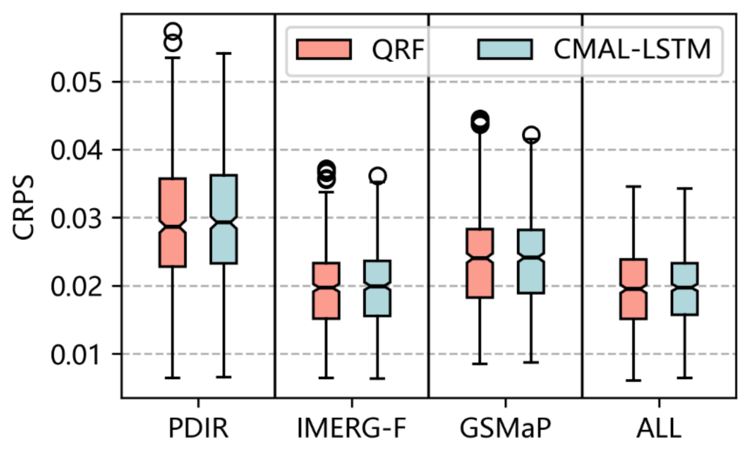

Figure 4. The boxplot of CPRS for different post-processing experiments.

In addition to their overall performance (Fig. 4), the QRF and CMAL-LSTM models exhibit similar spatial performance as it is reported in Fig. 5. Compared to PDIR and GSMaP, IMERG-F and multi-product results achieve relatively good performance in most of the 522 sub-basins. PDIR performs the worst, which inherently is attributed to its poorer input features, such as low autocorrelation skill of streamflow. The third row in the Fig.5 (i.e., Fig. 5i–l) shows that the differences between QRF and CMAL-LSTM are mostly within 10%. However, the introduction of multi-product features increased the gap between them, indicating that CMAL-LSTM has an advantage over the QRF model in processing multi-dimensional features. In the PDIR experiment, the QRF model demonstrates superior performance in 68.2% of the sub-basins (356 out of 522), while the CMAL-LSTM model performs better in the remaining 31.8% of sub-basins. Regarding the experiments conducted on IMERG-F, GSMaP, and multi-product (All), the proportions of QRF and CMAL-LSTM models are 65.5% and 34.5%, 54.2% and 45.8%, and 64.6% and 35.4% respectively.

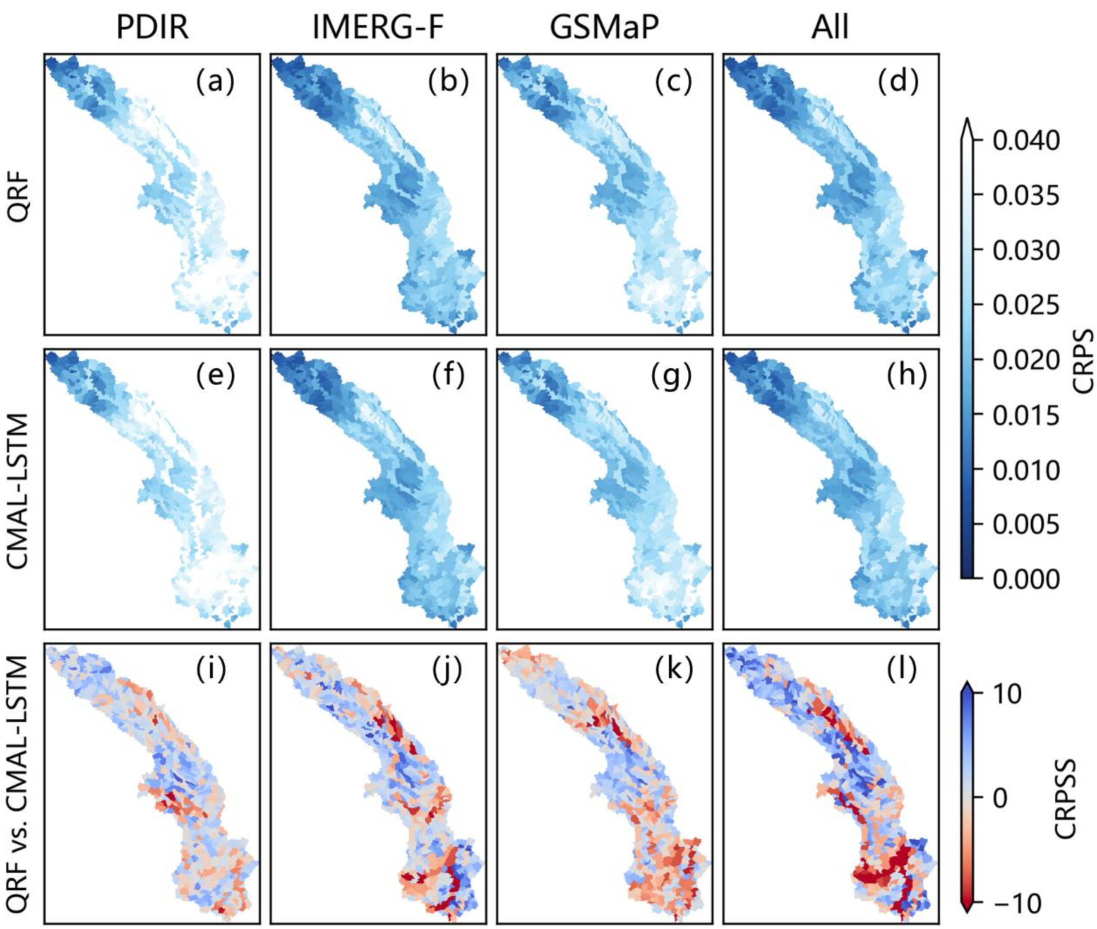

Figure 5. The spatial distribution of CRPS and CRPSS for different post-processing experiments.

4.2.2 The relationship between model performance and flow accumulation area (FAA)

To further investigate the differences between the two post-processing models, the relationship between the CRPS/CPRSS metrics and the FAA of sub-basins are presented in Fig. 6. Overall, the CRPS values of both post-processing models increases with increasing FAA, which is related to the streamflow amplitude of different sub-basins. Therefore, the relationship between the CRPSS score and the FAA as reported in Fig. 6e–h is of interest in order to compare the differences between the two post-processing models. It is observed that when the FAA is small, the QRF model performance is superior to the CMAL-LSTM model. However, as the FAA increases, the post-processing skill of the CMAL-LSTM model surpasses that of the QRF model. Additionally, the sub-basins are categorised, based on their size, into five intervals: less than 20,000 km^2, 20,000–40,000 km^2, 40,000–60,000 km^2, 60,000–100,000 km^2, and greater than 100,000 km^2. The corresponding number of sub-basins for each of the five intervals are 476, 15, 4, 13 and 14, respectively. The statistics of model performance in different FAA intervals are summarized in Table 2. In sub-basins with FAA less than 20,000 km^2, the QRF model shows a better performance. In the PDIR experiment, the QRF model has a higher CRPS value in 69.5% of sub-basins. In the IMERG-F, GSMaP, and multi-product experiments, the percentage of sub-basins where the QRF model outperforms the CMAL-LSTM model are 69.7%, 57.4%, and 67.2%, respectively. In sub-basins with FAA greater than 60,000 km^2, the CMAL-LSTM model shows an absolute advantage. In the PIDR experiment, the CMAL-LSTM model has a higher CRPS value in 16 sub-basins. In the IMERG-F, GSMaP, and multi-product experiments, the number of sub-basins where the CMAL-LSTM model has a higher CRPS value are 24, 27, and 25, respectively.

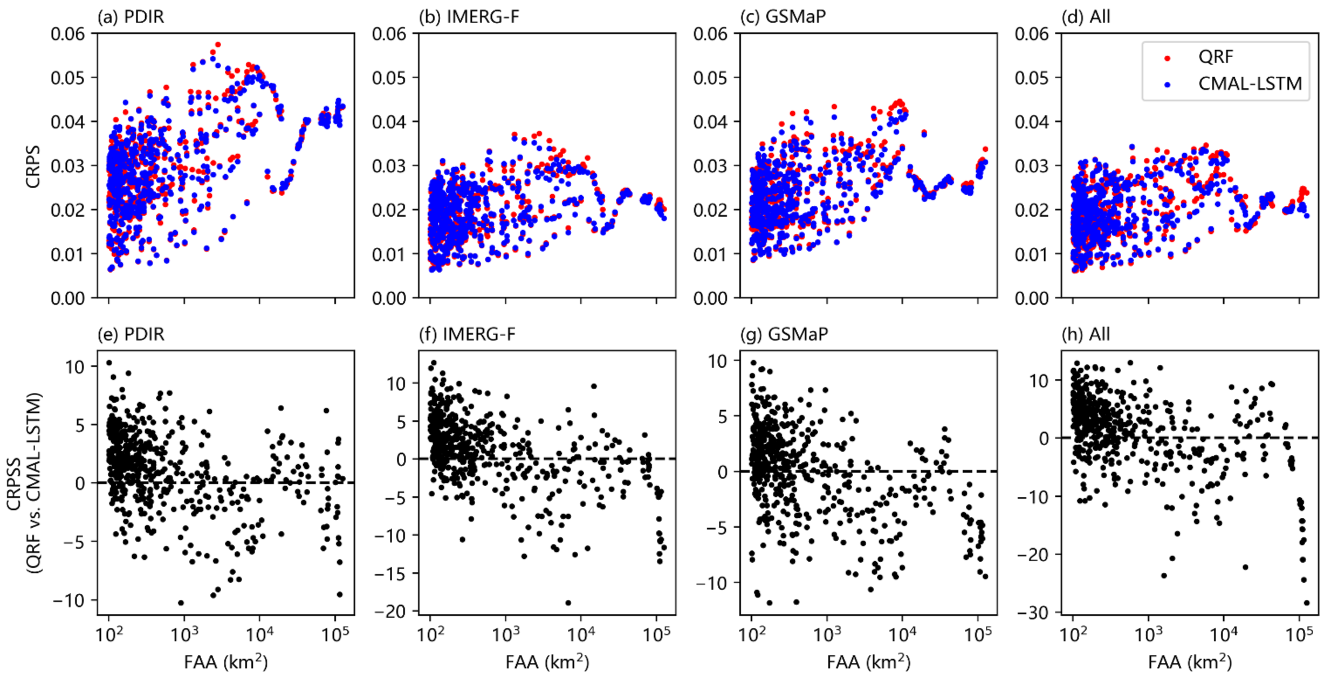

Figure 6. The relationships between (a–d) CRPS, (e–h) CRPSS and FAA.

Table 2. The probabilistic performance of two post-processing models for different FAA intervals. The bold numbers indicate better performance in each group.

FAA (10^4 km)	Number of sub-basins	PDIR		IMERG-F		GSMaP		ALL	
		QRF	CMAL-LSTM	QRF	CMAL-LSTM	QRF	CMAL-LSTM	QRF	CMAL-LSTM
< 2	476	**331**	145	**332**	144	**273**	203	**320**	156
2–4	15	**11**	4	6	**9**	**9**	6	**11**	4

4–6	4	**3**	1	1	**3**	1	**3**	**4**	0
6–10	13	4	**9**	3	**10**	0	**13**	2	**11**
> 10	14	7	7	0	**14**	0	**14**	0	**14**

4.2.3 Reliability and sharpness

The reliability diagram is further used to diagnose the difference in post-processing model performance in terms of reliability. To distinguish the differences in model performance of the CMAL-LSTM and QRF models with the change of FAA, the calculation of the reliability diagram is divided into two parts. On part is for FAA less than 60,000 km^2 as shown in Fig. 7a–c, which is obtained by combing all the streamflow prediction of 495 sub-basins. The second part is for FAA greater than 60,000 km^2 as shown in Fig. 7d–f, which is obtained by combing all the streamflow prediction of 27 sub-basins. Overall, when the FAA is less than 60,000 km^2, the performance of the two post-processing models is similar. The QRF model is slightly better than the CMAL-LSTM model. Except for the PDIR experiments, all experiments have a high reliability. As the threshold increases, all experiments show an increasing deviation from the diagonal line and a decrease in reliability. Moreover, when the FAA of sub-basin exceeds 60,000 km^2, the reliability of the post-processing experiments declines and the CMAL-LSTM model performs slightly better than the QRF model, with more points distributed along the diagonal line. As the threshold increases, the curve becomes more oscillatory, resulting in a significant decrease in reliability. Especially under extreme conditions and as is shown in Fig. 7f, the difference between the two post-processing models is large, with the CMAL-LSTM performing relatively better.

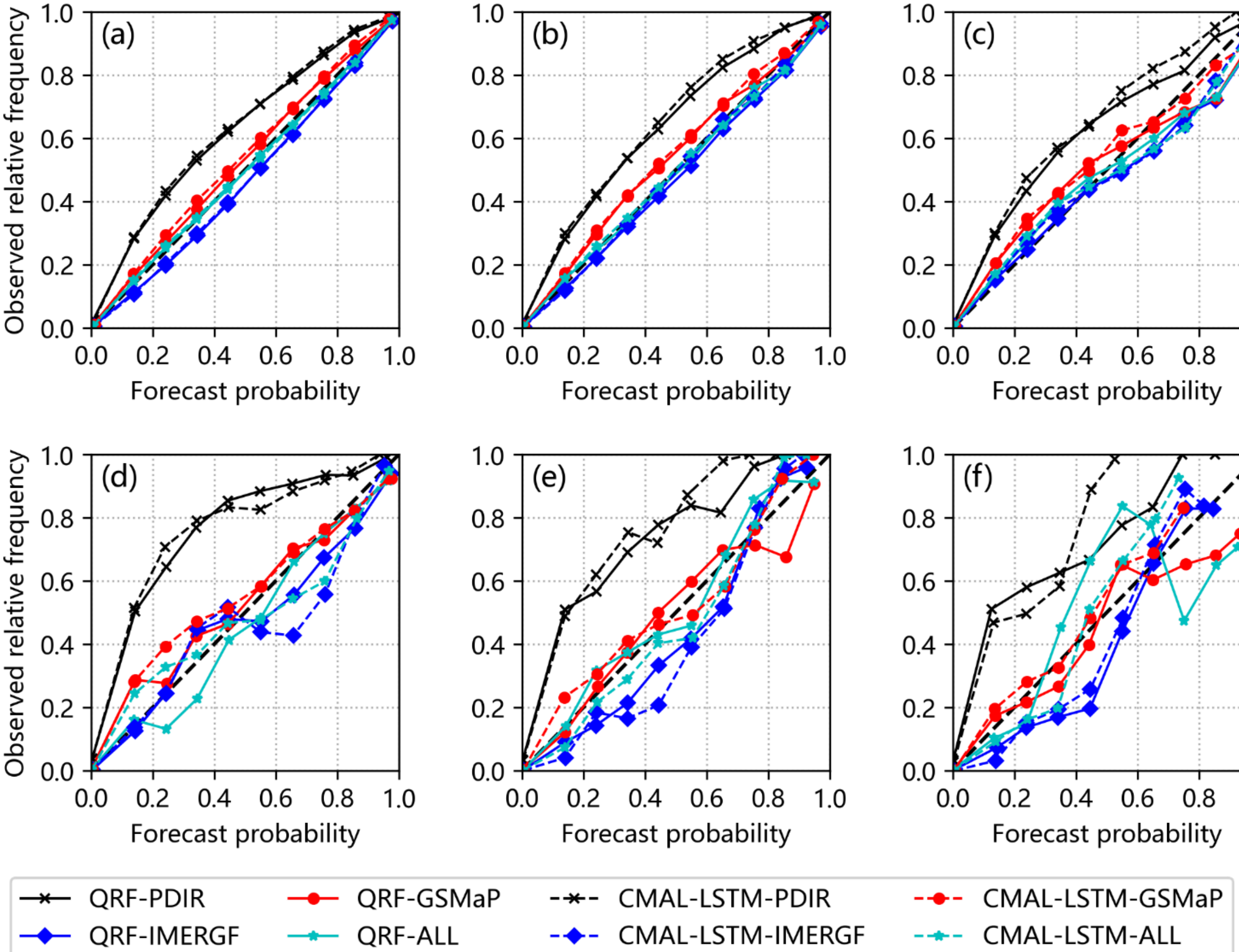

Figure 7. Reliability diagrams. (a) 80%, (b) 90% and (c) 95% percentiles of observations for the sub-basins with FAA less than 60,000 km^2 and (d) 80%, (e) 90% and (f) 95% percentiles of observations for the sub-basins with FAA greater than 60,000 km^2.

Sharpness describes the variability properties of predictive distribution and can be used to assess the differences between post-processing models from the uncertainty estimation perspective. To eliminate the influence of different flow regimes, all data are divided into high-flow seasons (May to October) and low-flow seasons (November to April). Sharpness metrics are calculated separately for each sub-basin. The average values of the metrics for all 522 sub-basins are listed in Table 3. The results show that, on average across all 522 sub-basins, the QRF model produces narrower prediction intervals than the CMAL-LSTM model during both high and low-flow seasons, indicating higher sharpness of the QRF model compared to CMAL-LSTM. This partially explains why the QRF model has higher CRPS values in most sub-basins. It is worth noting that the QRF model shows high coverage of the observations as well as narrower prediction intervals during high flow seasons. The average coverage of observations for the 25th to 75th quantiles ($CO_{25\text{-}75}$) is 1.5% higher for the QRF model than for the CMAL-LSTM model. However, the wider prediction interval of the CMAL-LSTM model results in higher coverage of observations during low flow seasons. The average coverage of observations for the 25th to 75th quantiles (CO_{25-75}) is 2% higher for the CMAL-LSTM model than for the QRF model. Interestingly, the 90% prediction intervals obtained by both post-processing methods contain 100% of the observations, based on the average values across 522 sub-basins during both high and low-flow seasons.

Table 3. Sharpness metrics. The bold numbers indicate better performance in each group.

Flow seasons	Metric	PDIR		IMERG-F		GSMaP		All	
		QRF	CMAL-LSTM	QRF	CMAL-LSTM	QRF	CMAL-LSTM	QRF	CMAL-LSTM
High-flow (May–Oct.)	MAD	**0.046**	0.048	**0.047**	0.052	**0.050**	0.054	**0.045**	0.047
	STD	**0.109**	0.112	**0.133**	0.139	**0.129**	0.133	**0.129**	0.134
	VAR	**0.013**	0.014	**0.020**	0.021	**0.018**	0.019	**0.018**	0.020
	DIS_{25-75}	0.0714	**0.0703**	**0.0753**	0.0757	**0.0781**	0.0785	0.0710	**0.0687**
	DIS_{5-95}	**0.184**	0.194	**0.192**	0.215	**0.206**	0.223	**0.184**	0.195
	CO_{25-75} (%)	**51.5**	50.1	**76.9**	76.0	**64.2**	62.8	**73.3**	71.4
	CO_{5-95} (%)	100	100	100	100	100	100	100	100
Low-flow (Nov.–Apr.)	MAD	**0.0085**	0.0100	**0.0073**	0.0094	**0.0088**	0.0104	**0.0064**	0.0069
	STD	**0.0264**	0.0284	**0.0280**	0.0301	**0.0305**	0.0323	**0.0258**	0.0262
	VAR	**8.32**	9.48	**9.10**	10.47	**10.40**	11.52	**7.71**	7.86
	DIS_{25-75}	**0.0121**	0.0124	**0.0099**	0.0112	**0.0121**	0.0122	0.0086	**0.0086**
	DIS_{5-95}	**0.033**	0.039	**0.029**	0.037	**0.036**	0.042	**0.026**	0.027
	CO_{25-75} (%)	72.2	**75.1**	88.8	**90.2**	69.1	**73.9**	**79.6**	79.2
	CO_{5-95} (%)	100	100	100	100	100	100	100	100

4.3 Deterministic (single-point) assessment

Although the post-processing model proposed in this study is probabilistic, decision-makers tend to prefer deterministic (single-point) prediction. Therefore, the average of the probability members is utilized as deterministic predictions to further compare the prediction accuracy of the models. Also, it can be viewed as a post hoc model examination.

4.3.1 Overall model performance

Figure 8 shows the performance evaluation of the streamflow simulations before (RAW) and after post-processing using the QRF and CMAL-LSTM models for 522 sub-basins. PCC, RB and NSE are used as performance metrics, with each sub-basin being evaluated separately. The median and mean of each metric across all 522 sub-basins are computed and reported in the first three columns of Table 4. The results indicate that both post-processing models significantly improved the simulation performance over the uncorrected streamflow. However, the CMAL-LSTM model consistently outperforms the QRF model across the precipitation products and the sub-basins.

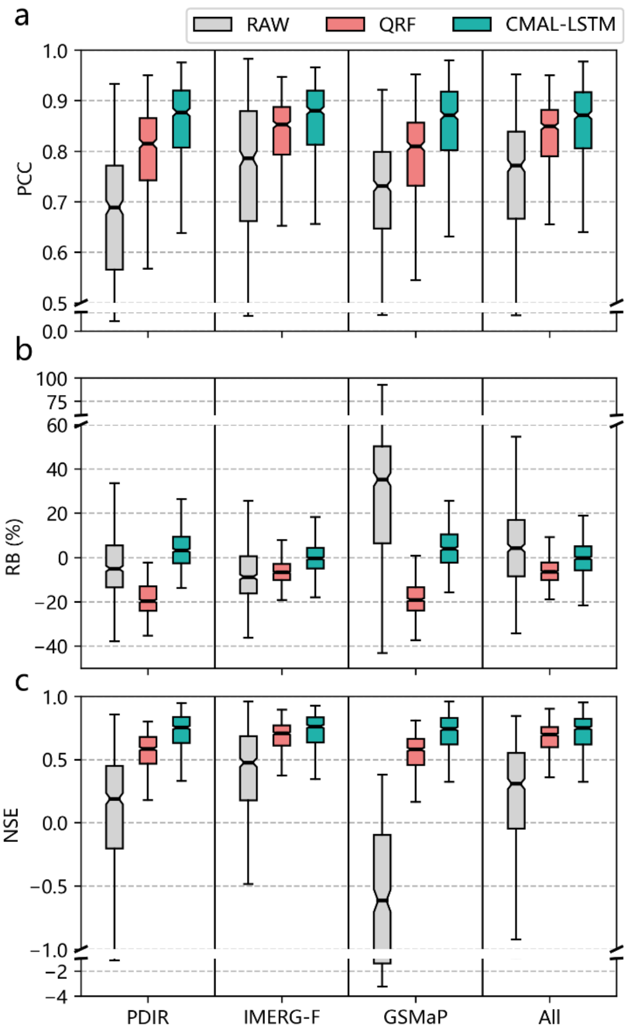

Figure 8. Boxplots of different model performance in 522 sub-basins. (a) PCC; (b) RB; and (c) NSE.

Figure 9 illustrates the spatial characteristics of the NSE improvement in streamflow simulations obtained through model comparison. Compared to the raw simulations (RAW), both QRF and CMAL-LSTM models exhibit significant improvements in almost all sub-basins. Among all post-processing experiments, GSMaP-CMAL-LSTM and GSMaP-QRF provide the most significant improvement in accuracy due to the poorer performance of the raw GSMaP-driven streamflow simulations. Conversely, the absolute NSE improvement brought by post-processing models are relatively small for the IMERG-F-driven streamflow simulations, and even a slight performance decline in 14.8% of sub-basins is observed in the IMERG-F-QRF experiment (Fig 9b). Compared to CMAL-LSTM, the QRF model does not show its advantage of deterministic (single-point) estimation in almost all sub-basins. The maximum difference in model performance appears in GSMaP experiments, followed by PDIR, IMERG-F and multi-product (All) experiments. This indicates that the deterministic (single-point) estimation ability of the QRF model differs significantly from the CMAL-LSTM model for streamflow with poor raw simulation.

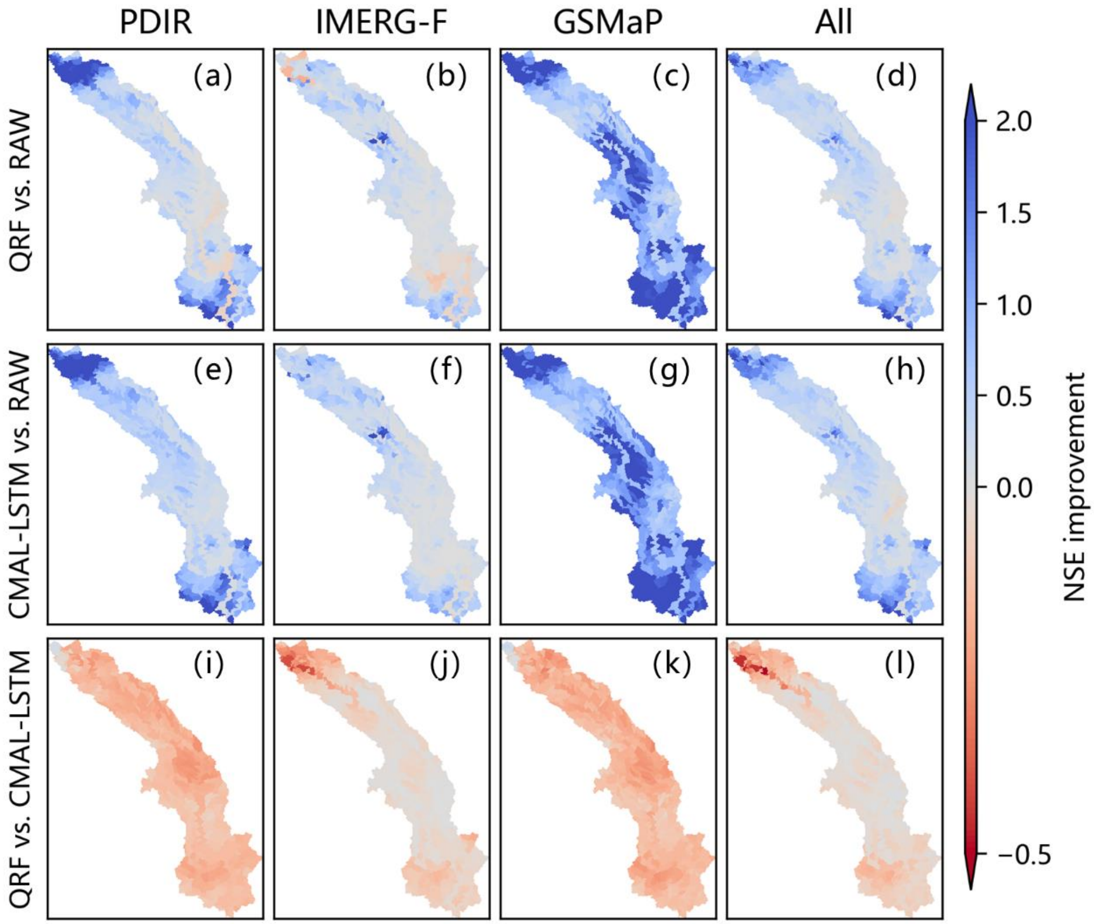

Figure 9. The spatial distribution of NSE improvement ($NSE_{PP} - NSE_{raw}$) between (a–d) QRF and RAW, (e–h) CMAL-LSTM and RAW and (i–l) QRF and CMAL-LSTM in 522 sub-basins.

4.3.2 The relationship between model performance and flow accumulation area

Based on the spatial distribution shown in Fig.9, the relationship between model performance and the flow accumulation area (FAA) of the sub-basin is further investigated, following a similar analysis approach as in Sect. 4.2.2 and Fig. 6. The findings, presented in Fig. 10, show that the performance of the model improves as the FAA of sub-basin increases. Moreover, the CMAL-LSTM model outperforms the QRF model in all experiments (see statistics in Table S2). However, as the FAA of sub-basin increase, the gap between the CMAL-LSTM model and QRF model narrows to some extent. This trend is particularly evident the IMERG-F driven experiment. Nonetheless, in experiments such as PDIR, GSMaP and multi-product (All), and the increase in FAA has little effect on the difference between the CMAL-LSTM and QRF models. This suggests that highly biased information from raw streamflow simulation has a greater impact on the QRF than on the CMAL-LSTM model.

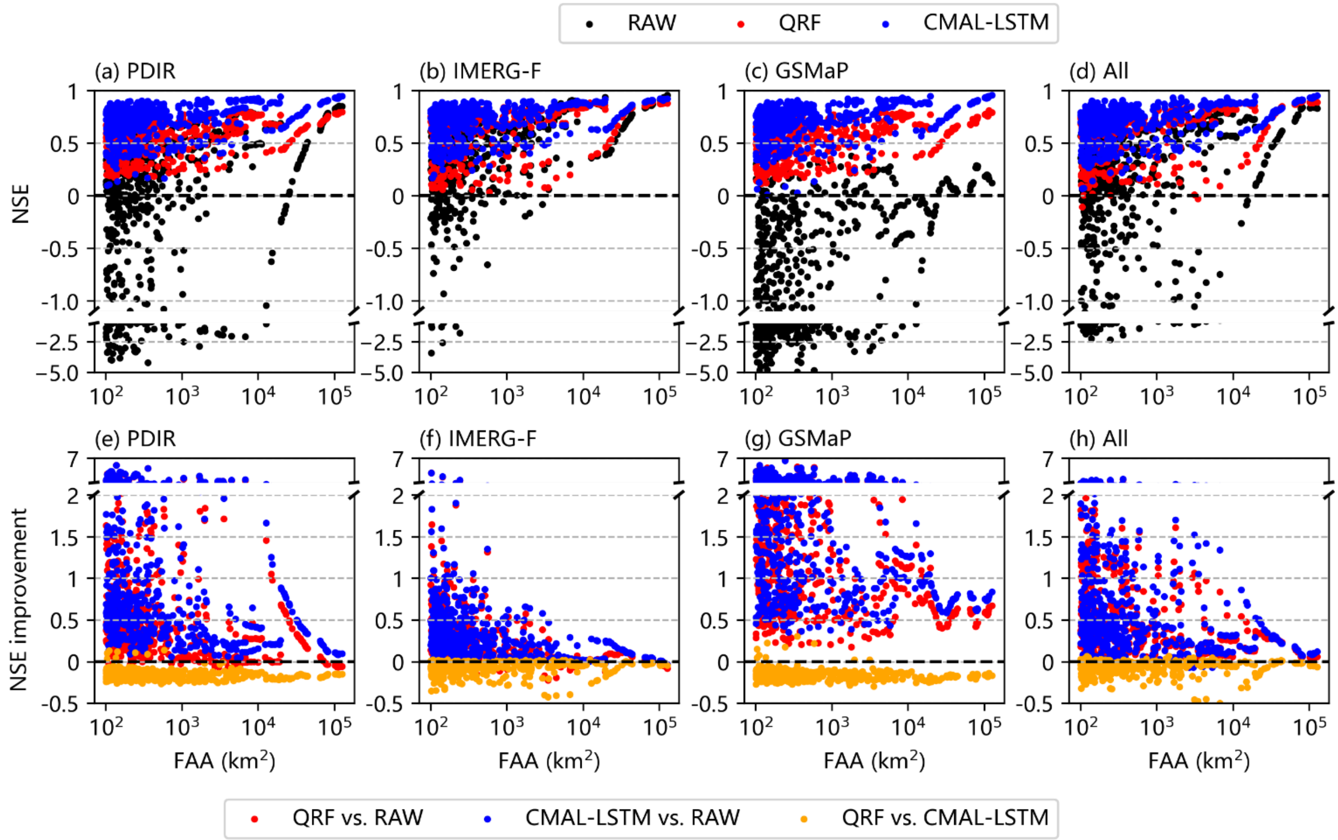

Figure 10. The relationships between (a–d) NSE, (b–h) NSE improvement ($NSE_{PP} - NSE_{raw}$) and FAA.

4.3.3 High-flow, low-flow, and peak timing

Table 4 summarizes the means and medians of integrated metrics and flow regime indicators for the 522 sub-basins in different experiments. The first three columns of the table are the same as the metrics used in Fig. 8. PCC and RB are the components of Nash-Sutcliffe efficiency (NSE). In order to guarantee the consensus of the results, another integrated indicator KGE is also calculated. The KGE performs identical to NSE, confirming the superiority of the CMAL-LSTM model. The last four columns of the table are flow-related indicators. Overall, the CMAL-LSTM model remains the best, except for the low-flow bias (FLV), where the QRF model is more effective. However, as indicated by the high-flow bias (FHV), both post-processing models have limitations in handling flood peaks. Regardless of the precipitation product used to drive the streamflow simulations, the bias of the flood peak changes from an overestimation (RAW) to an underestimation (post-processing). In addition, there is a certain degree of deviation in the simulations of peak time. Flood peaks have always posed a challenging problem in hydrological simulation, which highlights the necessity of probabilistic post-processing.

Table 4. Summary of integrated metrics and flow regime indicators of different models in 522 sub-basins. The bold numbers indicate better performance in each group.

Input	Aggregation	Model	PCC	RB	NSE	KGE	FHV	FMS	FLV	PT
			Metric							
PDIR	Mean	RAW	0.656	**-0.02**	-0.1	0.521	33.11	-5.3	-17.3	1.68
		QRF	0.785	-0.19	0.558	0.621	-43.4	-9.85	**3.143**	1.441
		CMAL-LSTM	**0.851**	0.032	**0.712**	**0.755**	**-28.8**	**1.201**	15.24	**1.328**
	Median	RAW	0.689	-0.05	0.19	0.572	**24.77**	-7.63	-12.5	1.692
		QRF	0.815	-0.2	0.584	0.645	-44.6	-10.5	**9.833**	1.417
		CMAL-LSTM	**0.877**	**0.032**	**0.752**	**0.778**	-29.6	**0.978**	19.13	**1.273**
IMERG-F	Mean	RAW	0.759	-0.06	0.389	0.664	**10.92**	-4.04	-14.3	1.459
		QRF	0.808	-0.06	0.648	0.718	-35.3	4.268	**-4.29**	1.394
		CMAL-LSTM	**0.852**	**-0.01**	**0.715**	**0.765**	-30.4	**2.409**	-5.05	**1.282**
	Median	RAW	0.785	-0.09	0.475	0.672	**9.555**	-6.35	-4.14	1.417
		QRF	0.852	-0.07	0.706	0.739	-37.6	**2.068**	5.878	1.333
		CMAL-LSTM	**0.88**	**-0.01**	**0.761**	**0.788**	-32.1	2.159	**2.467**	**1.231**
GSMaP	Mean	RAW	0.687	0.286	-0.92	0.308	88.82	8.465	-45.1	1.519
		QRF	0.778	-0.19	0.545	0.61	-45.4	-11.2	**15.94**	1.703
		CMAL-LSTM	**0.848**	**0.043**	**0.703**	**0.741**	**-31.2**	**0.708**	23.71	**1.44**
	Median	RAW	0.731	0.352	-0.62	0.393	82.86	12.08	-34.1	1.5
		QRF	0.809	-0.19	0.579	0.633	-48	-11.1	**23.73**	1.696
		CMAL-LSTM	**0.871**	**0.04**	**0.742**	**0.762**	**-32.3**	**1.037**	26.36	**1.417**
All	Mean	RAW	0.733	0.059	0.154	0.603	34.38	**2.332**	-15.5	1.456
		QRF	0.803	-0.06	0.637	0.704	-38.8	3.494	**8.635**	1.532
		CMAL-LSTM	**0.846**	**-0.01**	**0.703**	**0.76**	**-32.3**	4.855	10.27	**1.44**
	Median	RAW	0.771	0.042	0.306	0.664	**30.53**	2.228	**-4.74**	1.417
		QRF	0.849	-0.07	0.695	0.727	-42.3	**1.317**	14.96	1.542
		CMAL-LSTM	**0.871**	**-0.003**	**0.749**	**0.781**	-33.8	4.436	13.83	**1.417**

5 Discussion

5.1 Model comparison

Previous studies have demonstrated that the quantile regression forests (QRF) approach outperforms other quantile-based models, such as quantile regression and quantile neural networks (Taillardat et al., 2016; Tyralis et al., 2019; Tyralis and Papacharalampous, 2021). Additionally, recent research has indicated the effectiveness of mixture density networks based on the countable mixtures of asymmetric Laplacians models and long short-term memory networks (CMAL-LSTM) for hydrological probabilistic modelling (Klotz et al., 2022). In terms of reliability and sharpness evaluation for probabilistic prediction, CMAL-LSTM has been proven to achieve the best results compared to other models such as LSTM coupled with Gaussian mixture models, uncountable mixtures of asymmetric Laplacians models, and Monte Carlo dropout. These findings suggest that currently, QRF and CMAL-LSTM may be the most effective machine learning and deep learning model for hydrological probabilistic modelling. In this study, we conducted a comprehensive evaluation of the performance of these two advanced data-driven models in the context of streamflow probabilistic post-processing.

Our findings suggest that the QRF model outperformed the CMAL-LSTM model in terms of probability prediction in most sub-basins. And the performance difference between the two models was found to be associated with the catchment area of the sub-basins. The QRF model was superior in sub-basins with smaller catchment area, while the CMAL-LSTM model demonstrated better performance in larger sub-basins. However, when evaluated from a deterministic standpoint, the CMAL-LSTM model achieved higher NSE scores than the QRF model across nearly all sub-basins. The authors believe that the primary reason for the disparity in model performance is due to the differences in their respective model structure. As illustrated in Fig 2, the QRF model and the CMAL-LSTM model have dissimilar probabilistic procedure.

First, the QRF model and the CMAL-LSTM model differ in their treatment of input features. Specifically, the QRF model utilizes time embedding to flatten time-series features as input for the model. In contrast, the CMAL-LSTM model is capable of better learning the temporal autocorrelation of input features due to the inherent time-series learning capabilities of LSTM. As a result, the CMAL-LSTM model is more responsive to the autocorrelation of uncorrected streamflow features compared to the QRF model. The results depicted in Fig. S6 in the supplement provide evidence to support the interpretation that the performance difference between the QRF model and the CMAL-LSTM model is related to the autocorrelation of input features. The CMAL-LSTM model performs better in the sub-basin No. 250, where streamflow feature autocorrelations are more skillful, than in sub-basin No. 10, where streamflow feature autocorrelation skills are lacking.

Second, the QRF model and CMAL-LSTM differ in how they generate probabilistic members. The QRF model calculates the final probabilistic members by grouping them based on a predetermined number of quantiles (100 in this study). In contrast, the CMAL-LSTM model first specifies the form of the probabilistic distribution, then learns the parameters of the distribution using neural networks, and finally obtains the final probabilistic members by sampling. The QRF model produces an approximate and implicit probabilistic distribution, while the CMAL-LSTM model produces an accurate and explicit probabilistic distribution. Moreover, the predicted distribution from the CMAL-LSTM model using the mixture density function is more flexible. As a result, the QRF model produces narrower prediction intervals compared to the CMAL-LSTM model as is reported in Table 3. This is especially true when the sub-basin catchment area is smaller, and the streamflow amplitude is lower. This also explains the reason that the QRF model has higher sharpness in these cases compared to the CMAL-LSTM model. Figure. S7 presents the hydrograph and prediction intervals in two randomly selected sub-basins as an example. In sub-basin No.10, the CMAL-LSTM model achieves a balance between the width of the prediction interval and the observation coverage, which is more important for high-flow predictions and also explains why the CMAL-LSTM model has a higher CRPS value in the sub-basin with larger catchment area. In contrast, although the prediction interval of the QRF model is narrower, it is affected by systematic bias. For example, IMERG-F-QRF underestimates the peak flow in the high-flow season, leading to its smaller CRPS value compared to the CMAL-LSTM model. For sub-basin No.250 with a smaller catchment area, its rainfall-runoff response is faster, and the fluctuation of streamflow is greater. Localized precipitation events can also cause large pulse flow, which is the main feature of flash floods. Therefore, there are relatively more extreme samples. In this case, the QRF model learns and captures more observations with narrower prediction intervals, resulting in a better CRPS value.

Third, the QRF model and CMAL-LSTM model differ in their inference process. The QRF model utilizes a decision tree model as its base learner, which is a classification algorithm based on historical searches. Whereas, the CMAL-LSTM model uses a neural network with LSTM layer as its base learner, which is a more powerful fitting model. Due to the differences in model structure, the two models have different abilities to handle extreme events. When extreme event samples are limited, the QRF model tends to underestimate predictions due to its historical search-based approach. On the other hand, the CMAL-LSTM uses the mixture density function for extrapolation. However, both post-processing models still underestimates streamflow extreme events. The QRF model exhibits a higher degree of underestimation in sub-basins with larger catchment areas, resulting in unsatisfactory performance compared to the CMAL-LSTM model in these regions. These discrepancies also lead to lower NSE scores for the QRF model across all sub-basins, as the squared term in the NSE metric increases the sensitivity to high-flow processes which is reported in Fig. S8 in the supplement.

Furthermore, besides examining the differences in model performance, we investigated the effects of different input features on the post-processing model by using three different satellite precipitation products in this study. We observed a cascading impact on model performance in the rainfall-runoff and post-processing process. Given a fixed hydrological model, in areas with a small catchment area, the response of streamflow to precipitation is quicker, and the quality of satellite precipitation products directly influences the quality of streamflow prediction through the rainfall-runoff process. The temporal correlation of satellite precipitation determines the temporal correlation of streamflow prediction. Deviations in satellite precipitation led to the biased streamflow prediction, which have a more significant effect on the NSE score of streamflow prediction. This explains the reason that IMERG-F is optimal and PDIR is superior to GSMaP. During the transfer process from raw streamflow to post-processed streamflow, the autocorrelation skill of the raw runoff dictates the performance of the streamflow post-processing model. This clarifies why IMERG-F is still optimal, but GSMaP is superior to PDIR. Based on the results of the multi-product experiment, we observed that the post-processing model can learn better features to a larger extent, however, it cannot completely filter out the information that affects the model accuracy. Regrading information filtering, the CMAL-LSTM model surpasses the QRF model. These findings suggest that although streamflow post-processing can enhance model performance, opting for the best quality product is still a prudent decision when multiple precipitation products are available, and it can also save more computing resources. Another strategy is to execute precipitation post-processing before the hydrological model, which can assist the model to better learn the features and ultimately improve model performance.

5.2 Limitations and future work

This study provides a systematic evaluation of QRF and CMAL-LSTM models in probabilistic streamflow post-processing, yielding valuable insights and practical experience on model selection. However, there are still some deficiencies that need to be addressed in future research, which are summarized as follows.

First, we used simulated streamflow driven by observed precipitation as a proxy for true streamflow. This study diverges from previous research by focusing on sub-basin scale streamflow post-processing in a nested basin comprised of 522 sub-basins exhibiting varying flow accumulation areas, ranging from 100 km^2 to 127,164 km^2. To achieve the streamflow post-processing for these 522 sub-basins, corresponding streamflow observations are required, but such data are not readily available. As an alternative, we employed streamflow simulations generated by a calibrated hydrological model driven by observed precipitation. This approach yields a post-processing model performance that closely approximates the given reference; however, it is not an exact representation of actual streamflow post-processing. Despite this limitation, the reference generated was used to evaluate the performance of various post-processing models. Future studies could conduct a more in-depth comparison of different post-processing models in basins with more streamflow records. Nonetheless, our dataset remains scarce in the current community, and we have made it available along with this study to enable other researchers to evaluate and compare different methods against the benchmark presented in this study (Zhang et al., 2022b).

Second, there exists data imbalance among the studied sub-basins. Among the selected 522 sub-basins, it can be observed that model performance is related to the catchment size. However, the number of sub-basins corresponding to each of the five intervals (100–20,000 km^2, 20,000–40,000 km^2, 40,000–60,000 km^2, 60,000–100,000 km^2, and greater than 100,000 km^2) are 476, 15, 4, 13 and 14, respectively. Only 5.2% of the sub-basins have a catchment area larger than 60,000 km^2. This could potentially affect the generality of conclusions drawn. To address this limitation, more extensive and balanced datasets (such as Caravan, Kratzert et al., 2022b) can be utilized to achieve further validation of the research findings and a better understanding of different post-processing models.

Third, the selection of input features and hydrological models could be extended. In order to maintain model complexity and keep computational costs low, this study only used one variable, uncorrected streamflow, as the predictor. However, there are more variables that can be used as predictors, including other meteorological variables such as temperature and wind speed (Frame et al., 2021). In addition, basin-related attributes can provide us with local information, which is particularly helpful for the prediction in ungauged areas. In previous studies, all of these variables have been shown to have varying degrees of contributions to the model (Jiang et al., 2022). For post-processing, there are also studies that use model state variables and other output variables as predictors (Frame et al., 2021), which can provide us with information about the hydrological processes and increase the physical interpretability of the post-processing framework (Razavi, 2021; Tsai et al., 2021). However, state variables and outputs generated by hydrological models tend to be biased due to inherent bias in the satellite precipitation. It is unclear whether this is helpful for streamflow post-processing and requires further exploration. In terms of hydrological model selection, only the distributed time-variant gain model (DTVGM) was used to simulate streamflow from three different satellite precipitation products to increase the diversity of post-processing experiments. By doing so, the other two sources of uncertainty, namely, model structure and parameters, were eliminated, since the focus of this study was on comparing post-processing model with input uncertainty. It is worthing noting that in addition to input uncertainty, hydrological model structure and parameter uncertainty are also significant sources of uncertainty, as highlighted by Herrera et al. (2022) and Mai et al. (2022). For future post-processing model comparisons, we suggest to adopt the approach of using multiple hydrological models to analyse the uncertainty of model structure and parameters (Ghiggi et al., 2021; Troin et al., 2021; Mai et al., 2022).

Fourth, the post-processing models have limitations in handling streamflow extreme events, as observed through comparative analysis and visualization as reported in Table 4 and Fig. S8 in the supplement. The QRF model is based on a historical analogy search, wherein the model finds a group of similar samples and averages them at the leaf nodes to obtain the final prediction (Li and Martin, 2017). As a result, the limited number of samples, particularly for extreme events, hinders its ability to predict such events. However, this limitation can be addressed by introducing additional parameter mixing methods, such as combining QRF and extreme value distribution. Previous attempts, such as combining QRF and extended generalized Pareto distribution, have shown promising results (Taillardat et al., 2019). Nonetheless, these mixing methods add complexity to the model and require additional calibration of hyperparameters. The CMAL-LSTM model is also constrained by the number of extreme event samples, but its performance in these extreme events exceeds that of the QRF model. Additionally, the CMAL-LSTM model chosen in this study is a mixture density network and the corresponding parameters are directly learned through neural network optimization algorithms like gradient descent. The authors believe that collecting more data samples and introducing additional predictors and distribution functions for extreme events can lead to further improvements.

Finally, it is important to constantly enhance and update the model comparison iteratively. The CMAL-LSTM model was selected based on its superior performance as proposed by Klotz et al. (2022). They also evaluated two other hybrid density networks and a probabilistic method using Monte Carlo dropout. Additionally, there are other probabilistic prediction methods such as the variational inference (Li et al., 2021) and generative adversarial networks (Pan et al., 2021). In a rapidly evolving community, new methods can be applied and tested to further improve the performance of streamflow post-processing in future research.

6 Conclusions

In this study, a series of well-designed experiments to compare the performance of two state-of-the-art models for streamflow probabilistic post-processing were conducted: a machine learning model (quantile regression forests) and a deep learning model (countable mixtures of asymmetric Laplacians long short-term memory network). Using observed precipitation and three different satellite precipitation products to drive the calibrated hydrological model, we generated a large-sample dataset of 522 sub-basins with paired streamflow reference and biased streamflow simulations. We evaluated the model performance from both probabilistic and deterministic perspectives, including reliability, sharpness, accuracy, and flow regime, through intuitive case studies. These experiments established a path for understanding the model differences in probabilistic modelling and post-processing, provided practical experience for model selection, and extracted insights for model improvement. It also serves as a reference for establishing benchmark tests for model evaluation, including dataset construction and metrics selection. Furthermore, streamflow post-processing provides dependable data support for a range of downstream tasks, such as flood risk analysis, reservoir scheduling, and water resource management. The empirical findings of this study for the two post-processing models are summarized below.

(1) Based on the probabilistic assessment, the QRF and CMAL-LSTM models exhibit comparable performance. However, their model differences are correlated with the flow accumulation area (FAA) of sub-basins. In cases where the catchment area of a sub-basin is small, the QRF model generates a narrower prediction interval, resulting in better CRPS scores compared to the CMAL-LSTM model in most sub-basins. Conversely, for larger sub-basins (over 60,000 km^2 in this study), the CMAL-LSTM model outperforms the QRF model due to its ability to learn autocorrelation skills of features and capture extreme values.

(2) Based on the deterministic assessment, it can be concluded that the CMAL-LSTM model performs better than the QRF model in capturing high-flow process and flow duration curve. On the other hand, the QRF model tends to underestimate the high-flow process, resulting in worse NSE score across all sub-basins. Both models, however, have the issue of underestimating flood peaks due to sparse samples of extreme events.

(3) For the input uncertainties introduced by the different satellite precipitation products, both models are able to reduce their impact on the streamflow simulation. However, the performance of the post-processing models does not improve further in the multi-product experiments. Instead, the inclusion of heavily biased inputs leads to a deterioration in model performance. Opting for a single precipitation product that is best suited to the task at hand is a more prudent approach to safeguard model performance and minimize computational cost, rather than using multiple precipitation products with varying degrees of quality.

(4) Given the performance of post-processing models, the authors believe they have the potential to be applied to other sources of uncertainty that affect hydrological modelling, such as model structure and parameter uncertainty.

Data and code availability. The GPM IMERG Final Run is free available at GES DISC (https://gpm.nasa.gov/node/3328). The PDIR data can be freely download from CHRS Data Portal (http://chrsdata.eng.uci.edu/). The GSMaP data is publicly available (at https://sharaku.eorc.jaxa.jp/GSMaP/index.htm). The CMA precipitation observation is provided by the National Meteorological Information Centre of China Meteorological Administration. The soil types are free available at http://www.fao.org/soils-portal/soil-survey/soil-maps-and-databases/harmonized-world-soil-database-v12/en/. The land use data is free available from Chinese National Tibetan Plateau Third Pole Environment Data Centre at http://data.tpdc.ac.cn/en/data/a75843b4-6591-4a69-a5e4-6f94099ddc2d/. The DEM data is free available at https://www.gscloud.cn/. The QRF model code is available at Github repository (https://github.com/jnelson18/pyquantrf)

(Jnelson18, 2022). The CMAL-LSTM model code is available at Github repository (https://github.com/neuralhydrology/neuralhydrology) (Kratzert et al., 2022a). The dataset and results of this study are available at Zenodo repository (https://zenodo.org/record/7187505) (Zhang et al., 2022b).

Author contribution. Conceptualization, YZ, AY, PN, BA, SS, KH and YW; methodology, YZ and AY; software, YZ and AY; validation, YZ; data curation, YZ, AY, PN and BA; visualization, YZ; supervision, AY KH, and SS; project administration, AY. and SS; funding acquisition, AY and SS. original draft preparation, YZ; review and editing, YZ, AY, PN, BA, SS, KH and YW; All authors have read and agreed to the published version of the manuscript.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. This research is jointly supported by the Natural Science Foundation of China (No. 42171022, 51879009), the Second Tibetan Plateau Scientific Expedition and Research Program (No. 2019QZKK0405), the National Key Research and Development Program of China (No. 2018YFE0196000), the U.S. Department of Energy (DOE Prime Award DE-IA0000018).

References

Althoff, D., Rodrigues, L. N., and Bazame, H. C.: Uncertainty quantification for hydrological models based on neural networks: the dropout ensemble, Stochastic Environmental Research and Risk Assessment, 35(5), 1051-1067, https://doi.org/10.1007/s00477-021-01980-8, 2021.

Beven, K.: Changing ideas in hydrology—the case of physically-based models, Journal of Hydrology, 105(1-2), 157-172, https://doi.org/10.1016/0022-1694(90)90161-P, 1989.

Bogner, K., and Pappenberger, F.: Multiscale error analysis, correction, and predictive uncertainty estimation in a flood forecasting system, Water Resources Research, 47(7), e2010WR009137, https://doi.org/10.1029/2010WR009137, 2011.

Bormann, K. J., Evans, J. P., and McCabe, M. F.: Constraining snowmelt in a temperature-index model using simulated snow densities, Journal of Hydrology, 517, 652-667, https://doi.org/10.1016/j.jhydrol.2014.05.073, 2014.

Breiman, L.: Random forests, Machine Learning, 45(1), 5-32, https://doi.org/10.1023/a:1010933404324, 2001.

Bröcker, J.: Evaluating raw ensembles with the continuous ranked probability score, Quarterly Journal of the Royal Meteorological Society, 138(667), 1611-1617, https://doi.org/10.1002/qj.1891, 2012.

Chawanda, C. J., George, C., Thiery, W., Griensven, A. V., Tech, J., Arnold, J., and Srinivasan, R.: User-friendly workflows for catchment modelling: Towards reproducible SWAT+ model studies, Environmental Modelling and Software, 134, 104812, https://doi.org/10.1016/j.envsoft.2020.104812, 2020.

Chen, H., Yong, B., Shen, Y., Liu, J., Hong, Y., and Zhang, J.: Comparison analysis of six purely satellite-derived global precipitation estimates, Journal of Hydrology, 581, 124376, https://doi.org/10.1016/j.jhydrol.2019.124376, 2020.

Clark, M. P., Nijssen, B., Lundquist, J. D., Kavetski, D., Rupp, D. E., Woods, R. A., Freer, J. E., Gutmann, E. D., Wood, A. W., Brekke, L. D., Arnold, J. R., Gochis, D. J., and Rasmussen, R. M.: A unified approach for process-based hydrologic modelling: 1. Modelling concept, Water Resources Research, 51(4), 2498-2514, https://doi.org/10.1002/2015WR017198, 2015.

Corzo Perez, G. A., Van Huijgevoort, M., Voß, F., and Van Lanen, H.: On the spatio-temporal analysis of hydrological droughts from global hydrological models, Hydrology and Earth System Sciences, 15(9), 2963-2978, https://doi.org/10.5194/hessd-8-619-2011, 2011.

Cunha, L. K., Mandapaka, P. V., Krajewski, W. F., Mantilla, R., and Bradley, A. A.: Impact of radar-rainfall error structure on estimated flood magnitude across scales: An investigation based on a parsimonious distributed hydrological model, Water Resources Research, 48(10), https://doi.org/10.1029/2012WR012138, 2012.

Dembélé, M., Hrachowitz, M., Savenije, H. H. G., Mariéthoz, G., and Schaefli, B.: Improving the Predictive Skill of a Distributed Hydrological Model by Calibration on Spatial Patterns With Multiple Satellite Data Sets, Water Resources Research, 56(1), https://doi.org/10.1029/2019WR026085, 2020.

Dong, J., Crow, W. T., and Reichle, R.: Improving Rain/No-Rain Detection Skill by Merging Precipitation Estimates from Different Sources, Journal of Hydrometeorology, 21(10), 2419-2429. https://doi.org/10.1175/JHM-D-20-0097.1, 2020.

Evin, G., Lafaysse, M., Taillardat, M., and Zamo, M.: Calibrated ensemble forecasts of the height of new snow using quantile regression forests and ensemble model output statistics, Nonlinear Processes in Geophysics, 28(3), 467-480, https://doi.org/10.5194/npg-28-467-2021, 2021.

Falck, A. S., Maggioni, V., Tomasella, J., Vila, D. A., and Diniz, F. L. R.: Propagation of satellite precipitation uncertainties through a distributed hydrologic model: A case study in the Tocantins–Araguaia basin in Brazil, Journal of Hydrology, 527, 943-957, https://doi.org/10.1016/j.jhydrol.2015.05.042, 2015.

Fang, K., and Shen, C.: Near-Real-Time Forecast of Satellite-Based Soil Moisture Using Long Short-Term Memory with an Adaptive Data Integration Kernel, Journal of Hydrometeorology, 21(3), 399-413, https://doi.org/10.1175/JHM-D-19-0169.1, 2020.

Fang, K., Kifer, D., Lawson, K., Feng, D., and Shen, C.: The data synergy effects of time-series deep learning models in hydrology, Water Resources Research, e2021WR029583. https://doi.org/10.1029/2021WR029583, 2022.

Fang, K., Shen, C., Kifer, D., and Yang, X.: Prolongation of SMAP to spatiotemporally seamless coverage of continental US using a deep learning neural network, Geophysical Research Letters, 44(21), 11-30, https://doi.org/10.1002/2017GL075619, 2017.

Frame, J. M., Kratzert, F., Raney, A., Rahman, M., Salas, F. R., and Nearing, G. S.: Post-Processing the National Water Model with Long Short-Term Memory Networks for Streamflow Predictions and Model Diagnostics, JAWRA Journal of the American Water Resources Association, 57(6), 885-905, https://doi.org/10.1111/1752-1688.12964, 2021.

Ghiggi, G., Humphrey, V., Seneviratne, S. I., and Gudmundsson, L.: G-RUN ENSEMBLE: A Multi-Forcing Observation-Based Global Runoff Reanalysis, Water Resources Research, 57(5), e2020WR028787, https://doi.org/10.1029/2020WR028787, 2021.

Gneiting, T., Balabdaoui, F., and Raftery, A. E.: Probabilistic forecasts, calibration and sharpness, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 69(2), 243-268, https://doi.org/10.1111/j.1467-9868.2007.00587.x, 2007.

Gou, J., Miao, C., Duan, Q., Tang, Q., Di, Z., Liao, W., Wu, J., and Zhou, R.: Sensitivity Analysis-Based Automatic Parameter Calibration of the VIC Model for Streamflow Simulations Over China, Water Resources Research, 56(1), e2019WR025968, https://doi.org/10.1029/2019WR025968, 2020.

Gou, J., Miao, C., Samaniego, L., Xiao, M., Wu, J., and Guo, X.: CNRD v1.0: A High-Quality Natural Runoff Dataset for Hydrological and Climate Studies in China. Bulletin of the American Meteorological Society, 102(5), E929-E947. https://doi.org/10.1175/BAMS-D-20-0094.1, 2021.

Hartmann, H. C., Pagano, T. C., Sorooshian, S., and Bales, R.: Confidence builders: Evaluating seasonal climate forecasts from user perspectives, Bulletin of the American Meteorological Society, 83(5), 683-698, https://doi.org/10.1175/1520-0477(2002)083<0683:CBESCF>2.3.CO;2, 2002.

Herrera, P. A., Marazuela, M. A., and Hofmann, T.: Parameter estimation and uncertainty analysis in hydrological modelling, Wiley Interdisciplinary Reviews-Water, 9(1), e1569, https://doi.org/10.1002/wat2.1569, 2022.

Honti, M., Scheidegger, A., and Stamm, C.: The importance of hydrological uncertainty assessment methods in climate change impact studies, Hydrology and Earth System Sciences, 18(8), 3301-3317, https://doi.org/10.5194/hess-18-3301-2014, 2014.

Hori, T., Cho, J., and Watanabe, S.: End-to-end speech recognition with word-based RNN language models, IEEE Spoken Language Technology Workshop (SLT), 389-396, https://doi.org/10.1109/SLT.2018.8639693, 2018.

Hou, A. Y., Kakar, R. K., Neeck, S., AA, A., Kummerow, C. D., Kojima, M., Oki, R., Nakamura, K., and Iguchi, T.: The Global Precipitation Measurement Mission, Bulletin of the American Meteorological Society, 95(5), 701-722, https://doi.org/10.1175/BAMS-D-13-00164.1, 2013.

Huffman, G. J., Bolvin, D. T., Nelkin, E. J., and Tan, J.: Integrated Multi-satellitE Retrievals for GPM (IMERG) technical documentation, NASA/GSFC Code, 612(47), 2019, 2015.

Huffman, G.J., E.F. Stocker, D.T. Bolvin, E.J. Nelkin, Jackson T.: GPM IMERG Final Precipitation L3 1 day 0.1 degree x 0.1 degree V06, Edited by Andrey Savtchenko, Greenbelt, MD, Goddard Earth Sciences Data and Information Services Center (GES DISC), Accessed: [2021-7-30], https://doi.org/10.5067/GPM/IMERGDF/DAY/06, 2019.

Jajarmizadeh, M., Harun, S., and Salarpour, M.: A review on theoretical consideration and types of models in hydrology, Journal of Environmental Science and Technology, 5(5), 249-261, https://doi.org/10.3923/jest.2012.249.261, 2012.

Jiang, L., and Bauer-Gottwein, P.: How do GPM IMERG precipitation estimates perform as hydrological model forcing? Evaluation for 300 catchments across Mainland China, Journal of Hydrology, 572, 486-500, https://doi.org/10.1016/j.jhydrol.2019.03.042, 2019.

Jiang, S., Zheng, Y., Wang, C., and Babovic, V.: Uncovering Flooding Mechanisms Across the Contiguous United States Through Interpretive Deep Learning on Representative Catchments, Water Resources Research, 58(1), e2021WR030185, https://doi.org/10.1029/2021WR030185, 2022.

Jnelson18.: jnelson18/pyquantrf: DOI release (v0.0.3doi), Zenodo [code], https://doi.org/10.5281/zenodo.5815105, 2022.

Kasraei, B., Heung, B., Saurette, D. D., Schmidt, M. G., Bulmer, C. E., and Bethel, W.: Quantile regression as a generic approach for estimating uncertainty of digital soil maps produced from machine-learning, Environmental Modelling and Software, 144, 105139, https://doi.org/10.1016/j.envsoft.2021.105139, 2021.

Kaune, A., Chowdhury, F., Werner, M., and Bennett, J.: The benefit of using an ensemble of seasonal streamflow forecasts in water allocation decisions, Hydrology and Earth System Sciences, 24(7), 3851-3870, https://doi.org/10.5194/hess-24-3851-2020, 2020.

Khakbaz, B., Imam, B., Hsu, K., and Sorooshian, S.: From lumped to distributed via semi-distributed: Calibration strategies for semi-distributed hydrologic models, Journal of Hydrology, 418, 61-77. https://doi.org/10.1016/j.jhydrol.2009.02.021, 2012.

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, https://doi.org/10.1016/j.jhydrol.2012.01.011, 2012.

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, 26(6), 1673-1693, https://doi.org/10.5194/hess-26-1673-2022, 2022.

Kobold, M., and Sušelj, K.: Precipitation forecasts and their uncertainty as input into hydrological models, Hydrology and Earth System Sciences, 9(4), 322-332, https://doi.org/10.5194/hess-9-322-2005, 2005.

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(11), 6005-6022, https://doi.org/10.5194/hess-22-6005-2018, 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(12), 11344-11354, https://doi.org/10.1029/2019WR026065, 2019a.

Kratzert, F., Klotz, D., Hochreiter, S., and Nearing, G. S.: A note on leveraging synergy in multiple meteorological data sets with deep learning for rainfall–runoff modelling, Hydrology and Earth System Sciences, 25(5), 2685-2703, https://doi.org/10.5194/hess-25-2685-2021, 2021.

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(12), 5089-5110, https://doi.org/10.5194/hess-23-5089-2019, 2019b.

Kratzert, F., Gauch, M., Nearing, G., and Klotz, D.: NeuralHydrology-A Python library for Deep Learning research in hydrology, Journal of Open Source Software, 7(71), 4050, https://doi.org/10.21105/joss.04050, 2022a

Kratzert, F., Nearing, G., Addor, N., Erickson, T., Gauch, M., Gilon, O., Gudmundsson, L., Hassidim, A., Klotz, D., and Nevo, S.: Caravan-A global community dataset for large-sample hydrology, EarthArXiv, [preprint], https://doi.org/10.31223/X50S70, 2022b.

Kubota, T., Aonashi, K., Ushio, T., Shige, S., Takayabu, Y. N., Kachi, M., Arai, Y., Tashima, T., Masaki, T., and Kawamoto, N.: Global Satellite Mapping of Precipitation (GSMaP) products in the GPM era, Satellite precipitation measurement, 1, 355-373, https://doi.org/10.1007/978-3-030-24568-9_20, 2020.

Kubota, T., Shige, S., Hashizume, H., Aonashi, K., Takahashi, N., Seto, S., Hirose, M., Takayabu, Y. N., Ushio, T., and Nakagawa, K.: Global precipitation map using satellite-borne microwave radiometers by the GSMaP project: Production and validation, IEEE Transactions On Geoscience and Remote Sensing, 45(7), 2259-2275, https://doi.org/10.1109/TGRS.2007.895337, 2007.

Lees, T., Buechel, M., Anderson, B., Slater, L., Reece, S., Coxon, G., and Dadson, S. J.: Benchmarking Data-Driven Rainfall-Runoff Models in Great Britain: A comparison of LSTM-based models with four lumped conceptual models, Hydrology and Earth System Sciences, 25(10), 5517–5534, https://doi.org/10.5194/hess-2021-127, 2021.

Li, A. H., and Martin, A.: Forest-type regression with general losses and robust forest, Proceedings of the 34th International Conference on Machine Learning, 70, 2091-2100, 2017.

Li, B., Friedman, J., Olshen, R., and Stone, C.: Classification and regression trees (CART), Biometrics, 40(3), 358-361, Retrieved from http://statweb.lsu.edu/faculty/li/IIT/tree1.pdf, 1984.

Li, D., Marshall, L., Liang, Z., and Sharma, A.: Hydrologic multi-model ensemble predictions using variational Bayesian deep learning, Journal of Hydrology, 604, 127221, https://doi.org/10.1016/j.jhydrol.2021.127221, 2022.

Li, D., Marshall, L., Liang, Z., Sharma, A., and Zhou, Y., Bayesian LSTM With Stochastic Variational Inference for Estimating Model Uncertainty in Process-Based Hydrological Models, Water Resources Research, 57(9), https://doi.org/10.1029/2021WR029772, 2021.

Li, M., Wang, Q. J., Bennett, J. C., and Robertson, D. E.: A strategy to overcome adverse effects of autoregressive updating of streamflow forecasts, Hydrology and Earth System Sciences, 19(1), 1-15, https://doi.org/10.5194/hess-19-1-2015, 2015.

Li, M., Wang, Q. J., Bennett, J. C., and Robertson, D. E.: Error reduction and representation in stages (ERRIS) in hydrological modelling for ensemble streamflow forecasting, Hydrology and Earth System Sciences, 20(9), 3561-3579, https://doi.org/10.5194/hess-20-3561-2016, 2016.

Li, W., Duan, Q., Miao, C., Ye, A., Gong, W., and Di, Z.: A review on statistical postprocessing methods for hydrometeorological ensemble forecasting, Wiley Interdisciplinary Reviews: Water, 4(6), e1246, https://doi.org/10.1002/wat2.1246, 2017.

Mai, J., Craig, J. R., Tolson, B. A., and Arsenault, R.: The sensitivity of simulated streamflow to individual hydrologic processes across North America, Nature Communications, 13(1), https://doi.org/10.1038/s41467-022-28010-7, 2022.

Mai, J., Shen, H., Tolson, B. A., Gaborit, É., Arsenault, R., Craig, J. R., Fortin, V., Fry, L. M., Gauch, M., Klotz, D., Kratzert, F., O'Brien, N., Princz, D. G., Rasiya Koya, S., Roy, T., Seglenieks, F., Shrestha, N. K., Temgoua, A. G. T., Vionnet, V., and Waddell, J. W.: The Great Lakes Runoff Intercomparison Project Phase 4: the Great Lakes (GRIP-GL), Hydrology and Earth System Sciences, 26(13), 3537-3572, https://doi.org/10.5194/hess-26-3537-2022, 2022.

Meinshausen, N., and Ridgeway, G.: Quantile regression forests, Journal of Machine Learning Research, 7(6), 983-999, https://www.jmlr.org/papers/volume7/meinshausen06a/meinshausen06a.pdf, 2006.

Miao, C., Gou, J., Fu, B., Tang, Q., Duan, Q., Chen, Z., Lei, H., Chen, J., Guo, J., and Borthwick, A. G.: High-quality reconstruction of China’s natural streamflow, Science Bulletin, 67(5), 547-556, https://doi.org/10.1016/j.scib.2021.09.022, 2022.

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

Nasreen, S., Součková, M., Vargas Godoy, M. R., Singh, U., Markonis, Y., Kumar, R., Rakovec, O., and Hanel, M.: A 500-year runoff reconstruction for European catchments, Earth System Science Data, 14, 4035–4056, https://doi.org/10.5194/essd-14-4035-2022, 2022.

Nearing, G. S., Tian, Y., Gupta, H. V., Clark, M. P., Harrison, K. W., and Weijs, S. V.: A philosophical basis for hydrological uncertainty, Hydrological sciences journal, 61(9), 1666-1678, https://doi.org/10.1080/02626667.2016.1183009, 2016.

Nguyen, P., Ombadi, M., Gorooh, V. A., Shearer, E. J., Sadeghi, M., Sorooshian, S., Hsu, K., Bolvin, D., and Ralph, M. F.: PERSIANN Dynamic Infrared–Rain Rate (PDIR-Now): A Near-Real-Time, Quasi-Global Satellite Precipitation Dataset, Journal of Hydrometeorology, 21(12), 2893-2906, https://doi.org/10.1175/JHM-D-20-0177.1, 2020a.

Nguyen, P., Shearer, E. J., Ombadi, M., Gorooh, V. A., Hsu, K., Sorooshian, S., Logan, W. S., and Ralph, M.: PERSIANN Dynamic Infrared–Rain Rate Model (PDIR) for High-Resolution, Real-Time Satellite Precipitation Estimation, Bulletin of the American Meteorological Society, 101(3), E286-E302, https://doi.org/10.1175/BAMS-D-19-0118.1, 2020b.

Pan, B., Anderson, G. J., Goncalves, A., Lucas, D. D., Bonfils, C. J., Lee, J., Tian, Y., and Ma, H. Y.: Learning to correct climate projection biases, Journal of Advances in Modelling Earth Systems, 13(10), e2021MS002509, https://doi.org/10.1029/2021MS002509, 2021.

Parrish, M. A., Moradkhani, H., and DeChant, C. M.: Toward reduction of model uncertainty: Integration of Bayesian model averaging and data assimilation, Water Resources Research, 48(3), https://doi.org/10.1029/2011WR011116, 2012.

Razavi, S.: Deep learning, explained: Fundamentals, explainability, and bridgeability to process-based modelling, Environmental Modelling and Software, 144, 105159, https://doi.org/10.1016/j.envsoft.2021.105159, 2021.

Schaake, J. C., Hamill, T. M., Buizza, R., and Clark, M.: HEPEX: the hydrological ensemble prediction experiment, Bulletin of the American Meteorological Society, 88(10), 1541-1548, https://doi.org/10.1175/BAMS-88-10-1541, 2007.

Shen, C., and Lawson, K.: Applications of deep learning in hydrology, Deep Learning for the Earth Sciences: A Comprehensive Approach to Remote Sensing, Climate Science, and Geosciences, 283-297, https://doi.org/10.1002/9781119646181.ch19, 2021.

Shen, Y., Ruijsch, J., Lu, M., Sutanudjaja, E. H., and Karssenberg, D.: Random forests-based error-correction of streamflow from a large-scale hydrological model: Using model state variables to estimate error terms, Computers and Geosciences, 159, 105019, https://doi.org/10.1016/j.cageo.2021.105019, 2022.

Shen, Z., Yong, B., Gourley, J. J., and Qi, W.: Real-time bias adjustment for satellite-based precipitation estimates over Mainland China, Journal of Hydrology, 596, 126133, https://doi.org/10.1016/j.jhydrol.2021.126133, 2021.

Sit, M., Demiray, B. Z., Xiang, Z., Ewing, G. J., Sermet, Y., and Demir, I.: A comprehensive review of deep learning applications in hydrology and water resources, Water Science and Technology, 82(12), 2635-2670, https://doi.org/10.2166/wst.2020.369, 2020.

Sittner, W. T., Schauss, C. E., and Monro, J. C.: Continuous hydrograph synthesis with an API-type hydrologic model, Water Resources Research, 5(5), 1007-1022, https://doi.org/10.1029/WR005i005p01007, 1969.

Sivapalan, M.: From engineering hydrology to Earth system science: milestones in the transformation of hydrologic science, Hydrology and Earth System Sciences, 22(3), 1665-1693, https://doi.org/10.5194/hess-22-1665-2018, 2018.

Sordo-Ward, Á., Granados, I., Martín-Carrasco, F., and Garrote, L.: Impact of Hydrological Uncertainty on Water Management Decisions, Water Resources Management, 30(14), 5535-5551, https://doi.org/10.1007/s11269-016-1505-5, 2016.

Staudemeyer, R. C., and Morris, E. R.: Understanding LSTM-a tutorial into long short-term memory recurrent neural networks, arXiv [preprint], arXiv:1909.09586, 2019.

Sun, Q., Miao, C., Duan, Q., Ashouri, H., Sorooshian, S., and Hsu, K. L.: A Review of Global Precipitation Data Sets: Data Sources, Estimation, and Intercomparisons, Reviews of Geophysics, 56(1), 79-107, https://doi.org/10.1002/2017RG000574, 2018.

Taillardat, M., Fougères, A., Naveau, P., and Mestre, O.: Forest-Based and Semiparametric Methods for the Postprocessing of Rainfall Ensemble Forecasting, Weather and Forecasting, 34(3), 617-634, https://dio.org/10.1175/WAF-D-18-0149.1, 2019.

Taillardat, M., Mestre, O., Zamo, M., and Naveau, P.: Calibrated Ensemble Forecasts Using Quantile Regression Forests and Ensemble Model Output Statistics, Monthly Weather Review, 144(6), 2375-2393, https://doi.org/10.1175/MWR-D-15-0260.1, 2016.

Tan, M. L., Gassman, P. W., Yang, X., and Haywood, J.: A review of SWAT applications, performance and future needs for simulation of hydro-climatic extremes, Advances in Water Resources, 143, 103662, https://doi.org/10.1016/j.advwatres.2020.103662, 2020.

Tian, Y., Peters-Lidard, C. D., Eylander, J. B., Joyce, R. J., Huffman, G. J., Adler, R. F., Hsu, K., Turk, F. J., Garcia, M., and Zeng, J.: Component analysis of errors in satellite-based precipitation estimates, Journal of Geophysical Research, 114(D24), https://doi.org/10.1029/2009JD011949, 2009.

Troin, M., Arsenault, R., Wood, A. W., Brissette, F., and Martel, J. L.: Generating Ensemble Streamflow Forecasts: A Review of Methods and Approaches Over the Past 40 Years, Water Resources Research, 57(7), https://doi.org/10.1029/2020WR028392, 2021.

Tsai, W., Feng, D., Pan, M., Beck, H., Lawson, K., Yang, Y., Liu, J., and Shen, C.: From calibration to parameter learning: Harnessing the scaling effects of big data in geoscientific modelling, Nature Communications, 12(1), 1-13, https://doi.org/10.1038/s41467-021-26107-z, 2021.

Tyralis, H., and Papacharalampous, G.: Quantile-based hydrological modelling, Water, 13(23), 3420, https://doi.org/10.3390/w13233420, 2021.

Tyralis, H., Papacharalampous, G., Burnetas, A., and Langousis, A.: Hydrological post-processing using stacked generalization of quantile regression algorithms: Large-scale application over CONUS, Journal of Hydrology, 577, 123957, https://doi.org/10.1016/j.jhydrol.2019.123957, 2019.

Wang, Q. J., Robertson, D. E., and Chiew, F. H. S.: A Bayesian joint probability modelling approach for seasonal forecasting of streamflows at multiple sites, Water Resources Research, 45(5), W05407, https://doi.org/10.1029/2008WR007355, 2009.

Wu, J., Yen, H., Arnold, J. G., Yang, Y. C. E., Cai, X., White, M. J., Santhi, C., Miao, C., and Srinivasan, R.: Development of reservoir operation functions in SWAT+ for national environmental assessments, Journal of Hydrology, 583, 124556, https://doi.org/10.1016/j.jhydrol.2020.124556, 2020.

Xia, J.: Identification of a constrained nonlinear hydrological system described by Volterra Functional Series, Water Resources Research, 27(9), 2415-2420, https://doi.org/10.1029/91WR01364, 1991.

Xia, J., Wang, G., Tan, G., Ye, A., and Huang, G. H.: Development of distributed time-variant gain model for nonlinear hydrological systems, Science in China Series D: Earth Sciences, 48(6), 713-723, https://doi.org/10.1360/03yd0183, 2005.

Xu, L., Chen, N., Moradkhani, H., Zhang, X., and Hu, C.: Improving Global Monthly and Daily Precipitation Estimation by Fusing Gauge Observations, Remote Sensing, and Reanalysis Data Sets, Water Resources Research, 56(3), https://doi.org/10.1029/2019WR026444, 2020.

Yang, Q., Wang, Q. J., and Hakala, K.: Achieving effective calibration of precipitation forecasts over a continental scale, Journal of Hydrology: Regional Studies, 35, 100818, https://doi.org/10.1016/j.ejrh.2021.100818, 2021.

Ye, A., Duan, Q., Schaake, J., Xu, J., Deng, X., Di, Z., Miao, C., and Gong, W.: Post-processing of ensemble forecasts in low-flow period, Hydrological Processes, 29(10), 2438-2453, https://doi.org/10.1002/hyp.10374, 2015.

Ye, A., Duan, Q., Yuan, X., Wood, E. F., and Schaake, J.: Hydrologic post-processing of MOPEX streamflow simulations, Journal of Hydrology, 508, 147-156, https://doi.org/10.1016/j.jhydrol.2013.10.055, 2014.

Ye, A., Duan, Q., Zeng, H., Li, L., and Wang, C.: A distributed time-variant gain hydrological model based on remote sensing, Journal of Resources and Ecology, 1(3), 222-230, https://doi.org/10.3969/j.issn.1674-764x.2010.03.005, 2010.

Ye, A., Duan, Q., Zhan, C., Liu, Z., and Mao, Y.: Improving kinematic wave routing scheme in Community Land Model, Hydrology Research, 44(5), 886-903, https://doi.org/10.2166/nh.2012.145, 2013.

Yilmaz, K. K., Gupta, H. V., and Wagener, T.: A process-based diagnostic approach to model evaluation: Application to the NWS distributed hydrologic model, Water Resources Research, 44(9), W09417, https://doi.org/10.1029/2007WR006716, 2008.

Zhang, X., Liu, P., Cheng, L., Liu, Z., and Zhao, Y, A back-fitting algorithm to improve real-time flood forecasting, Journal of Hydrology, 562, 140-150, https://doi.org/10.1016/j.jhydrol.2018.04.051, 2018.

Zhang, Y., and Ye, A.: Machine Learning for Precipitation Forecasts Postprocessing: Multimodel Comparison and Experimental Investigation, Journal of Hydrometeorology, 22(11), 3065-3085, https://doi.org/10.1175/JHM-D-21-0096.1, 2021.

Zhang, Y., Ye, A., Nguyen, P., Analui, B., Sorooshian, S., and Hsu, K.: New insights into error decomposition for precipitation products, Geophysical Research Letters, 48(17), e2021GL094092, https://doi.org/10.1029/2021GL094092, 2021a.

Zhang, Y., Ye, A., Nguyen, P., Analui, B., Sorooshian, S., and Hsu, K.: Error Characteristics and Scale Dependence of Current Satellite Precipitation Estimates Products in Hydrological Modelling, Remote Sensing, 13(16), 3061, https://doi.org/10.3390/rs13163061, 2021b.

Zhang, Y., Ye, A., Nguyen, P., Analui, B., Sorooshian, S., and Hsu, K.: QRF4P-NRT Probabilistic Post-processing of Near-real-time Satellite Precipitation Estimates using Quantile Regression Forests, Water Resources Research, 58(5), e2022WR032117, https://doi.org/10.1029/2022WR032117, 2022a.

Zhang, Y., Ye, A., Nguyen, P., Analui, B., Sorooshian, S., and Hsu, K.: Dataset and results for "Comparing machine learning and deep learning models for probabilistic post-processing of satellite precipitation-driven streamflow simulation" [Data set]. Zenodo. https://doi.org/10.5281/zenodo.7187505, 2022b.

Zhao, L., Duan, Q., Schaake, J., Ye, A., and Xia, J.: A hydrologic post-processor for ensemble streamflow predictions, Advances in geosciences, 29(29), 51-59, https://doi.org/10.5194/adgeo-29-51-2011, 2011.

Zhao, P., Wang, Q. J., Wu, W., and Yang, Q.: Extending a joint probability modelling approach for post-processing ensemble precipitation forecasts from numerical weather prediction models, Journal of Hydrology, 605, 127285, https://doi.org/10.1016/j.jhydrol.2021.127285, 2022.

Zhou, X., Polcher, J., and Dumas, P.: Representing Human Water Management in a Land Surface Model Using a Supply/Demand Approach, Water Resources Research, 57(4), https://doi.org/10.1029/2020WR028133, 2021.

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(9), 1313-1329, https://doi.org/10.1007/s00477-020-01766-4, 2020.

Zounemat-Kermani, M., Batelaan, O., Fadaee, M., and Hinkelmann, R: Ensemble machine learning paradigms in hydrology: A review, Journal of Hydrology, 598, 126266, https://doi.org/10.1016/j.jhydrol.2021.126266, 2021.