Water Resources Research

RESEARCH ARTICLE
10.1029/2022WR032117

QRF4P-NRT: Probabilistic Post-Processing of Near-Real-Time Satellite Precipitation Estimates Using Quantile Regression Forests

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

[1]State Key Laboratory of Earth Surface Processes and Resource Ecology, Faculty of Geographical Science, Beijing Normal University, Beijing, China, [2]Center for Hydrometeorology and Remote Sensing, Department of Civil and Environmental Engineering, University of California, Irvine, Irvine, CA, USA

Key Points:

- An ensemble postprocessor based on quantile regression forests for bias correction of satellite precipitation estimates is proposed
- The **P**robabilistic **P**ost-**P**rocessing of **N**ear-**R**eal-**T**ime Satellite **P**recipitation Estimates using **Q**uantile **R**egression **F**orests remarkably improved raw satellite precipitation estimates and provide reliable probabilistic outputs in a near-real-time way
- A static proxy of dynamic near-real-time predictor is an acceptable solution for operational application

Correspondence to:
A. Ye,
azye@bnu.edu.cn

Citation:
Zhang, Y., Ye, A., Nguyen, P., Analui, B., Sorooshian, S., & Hsu, K. (2022). QRF4P-NRT: Probabilistic post-processing of near-real-time satellite precipitation estimates using quantile regression forests. *Water Resources Research*, *58*, e2022WR032117. https://doi.org/10.1029/2022WR032117

Received 31 JAN 2022
Accepted 29 APR 2022

Author Contributions:
Conceptualization: Yuhang Zhang, Aizhong Ye
Data curation: Yuhang Zhang
Funding acquisition: Aizhong Ye
Methodology: Yuhang Zhang, Aizhong Ye, Phu Nguyen, Bita Analui
Software: Yuhang Zhang
Validation: Yuhang Zhang
Writing – original draft: Yuhang Zhang
Writing – review & editing: Aizhong Ye, Phu Nguyen, Bita Analui, Soroosh Sorooshian, Kuolin Hsu

Abstract Accurate and reliable near-real-time satellite precipitation estimation is of great importance for operational large-scale flood forecasting and drought monitoring. The state-of-the-art precipitation post-processing model is based on a deterministic approach to construct relationships between satellites estimates and ground observations. We propose a probabilistic postprocessor, the **P**robabilistic **P**ost-**P**rocessing of **N**ear-**R**eal-**T**ime Satellite **P**recipitation Estimates using **Q**uantile **R**egression **F**orests (QRF4P-NRT), based on quantile modeling, yielding both deterministic and probabilistic predictions. The experimental design incorporates different solutions of near-real-time predictors to further improve the model performance. Using the Integrated Multi-satellitE Retrievals Early Run for Global Precipitation Measurement Mission (IMERG-E) product as an example, we illustrate that the proposed method significantly improves the overall quality of the raw IMERG-E and is also superior to the bias-corrected product (IMERG Final Run, IMERG-F) at daily scale in a complex mountain basin. Evaluations of the corrected IMERG-E, raw IMERG-E, and IMERG-F using ground observation show that the corrected IMERG-E improves correlation coefficients (0.7), mean error (−0.14 mm/day) and root mean square error (3.3 mm/day) relative to the raw IMERG-E (0.31, −0.72 and 5.5 mm/day) and IMERG-F (0.34, −0.09 and 6.0 mm/day). The error decomposition further confirms that the QRF4P-NRT improves on the various deficiencies of the raw IMERG-E product. The ensemble assessment also demonstrates that the quantile outputs provide reliable prediction spread and sharp prediction intervals. The promising results indicate the great potential of the proposed method for probabilistic post-processing for near-real-time satellite precipitation estimates, and for further applications such as hydrological ensemble forecasting.

Plain Language Summary Errors in near-real-time satellite precipitation estimates limit their applications. The use of error correction models is better able to reduce the errors. However, current deterministic error correction models reduce errors while losing uncertain information. In this study, we propose a probabilistic error correction method that has been used in the field of ensemble numerical weather forecasts. While reducing the error, it is also possible to quantify the probabilistic information. Our method obtains the best score compared to both the raw product and bias-corrected product. This is of great interest for the application of near-real-time satellite precipitation estimates and can be further applied to operational flood forecasting and drought monitoring.

1. Introduction

Precipitation is a key climatic variable and an important element of the hydrological cycle. Real-time precipitation measurements provide a powerful tool for real-time drought and flood monitoring, forecasting, and warning (Hong et al., 2007; Nguyen et al., 2015; Qi et al., 2021; Zhou et al., 2014). Traditionally, several approaches are available for real-time precipitation measurements, but each of them has advantages and disadvantages (Sun et al., 2018). For example, gauge observations provide the most accurate estimates at the site scale, but it degrades at large scales because of sparse stations and sampling errors (Tang et al., 2018). Therefore, ground observations based on a limited gauge density do not provide sufficiently representative spatial precipitation distributions. Radar precipitation estimates suffer from similar issues and the method is also vulnerable to obstructions (Goudenhoofdt & Delobbe, 2009). High construction and maintenance costs also hinder the widespread rollout of both rain gauge and radar equipment, especially in remote areas (Nguyen et al., 2018).

Since the Tropical Rainfall Measuring Mission project, precipitation estimates based on satellite retrieval have been considered as a promising tool for large-scale precipitation measurements (Hou et al., 2013; Huffman et al., 2007). Multiple satellite precipitation estimates products are available (Sun et al., 2018; Y. Zhang et al., 2021a). Depending on whether they are bias-corrected and the record length of the data, they can be broadly classified into three categories: real-time or near-real-time products (RTs or NRTs), bias-corrected products (BCs), and climate data records (Y. Zhang et al., 2021a). Among them, NRTs (or RTs) with very short latency time are the ideal product for operational applications such as flood monitoring and forecasting (Hong et al., 2007; Zhou et al., 2014). For example, the latest Precipitation Estimate from Remotely Sensed Information using Artificial Neural Networks Dynamic Infrared Rain Rate near-real-time (PDIR-Now) issued by the Center for Hydrometeorology and Remote Sensing at the University of California, Irvine, which is primarily based on infrared imagery, has a delayed-release time of only 30 min (Nguyen, Ombadi, et al., 2020; Nguyen, Shearer, et al., 2020). Released by Global Precipitation Measurements Mission (GPM) led by the National Aeronautics and Space Administration (NASA), the latency time of IMERG early run is 4 hr (Huffman et al., 2015). Developed by Global Rainfall Watch led by the Japan Aerospace Exploration Agency, the latency time of Global Satellite Mapping of Precipitation real-time (GSMaP-Now) and near-real-time (GSMaP-NRT) is 0 and 4 hr, respectively (Kubota et al., 2020). However, due to the data source and retrieval algorithm, these raw precipitation estimates from NRTs (or RTs) normally suffer from large errors (Chen et al., 2020). This results in large uncertainty, which subsequently limits their operational applications.

Many studies have been attempted to develop effective bias correction methods for improving the quality of the satellite precipitation estimates (Dong et al., 2020; Gumindoga et al., 2019; W. Li et al., 2019). Typically, with ground observations, the satellite estimates are post-processed using regression-like or probability density function-based matching models. For example, many studies have developed multiple linear regression models (i.e., univariate or multivariate) based on the relationship between precipitation error and factors such as topography, seasonality, climate type, and rain rate (Chen et al., 2021). Distribution adjustment is mainly implemented by correcting the values of different quartiles to eventually match the observed and satellite-based cumulative density functions (CDFs; Maraun, 2013). However, precipitation post-processing is tagged with various difficulties such as nonlinearity, heteroskedasticity, and skewed distribution (A. H. Li et al., 2017). The existing methods mentioned above do not address these challenges well. For example, CDF-matching is a kind of "climatology-based" adjustment that is more effective for reducing systematic errors, but does not handle random errors well and does not improve the correlation coefficient between satellite estimates and ground observations (Shen et al., 2021). Moreover, regression models are mostly linear and do not address nonlinearity and heteroskedasticity well (Chen et al., 2021).

Recently, machine learning (ML) models have been commonly used in post-processing of the satellite precipitation estimates and related hydrometeorological applications because of their ability to better handle complex problems such as nonlinearity as well as big data (Sharifi et al., 2019; F. Wang et al., 2021; Y. Zhang & Ye, 2021). Among different ML models, the random forests (RF) model (Breiman, 2001) has been used for a variety of scenarios and showed superior model performance. For example, Ibarra-Berastegi et al. (2011) earlier applied the RF model for downscaling of precipitation and surface moisture flux. Subsequently, X. He et al. (2016) used and compared single RF and double RF models for GLDAS precipitation downscaling with different ratios of experiments. This RF-based approach to precipitation downscaling has also proven to be very effective in complex alpine terrain (Mei et al., 2020). Similarly, double RF models have been used in the study of merging multi-source satellite precipitation estimates products (L. Zhang et al., 2021). Baez-Villanueva et al. (2020) develop an RF-based MErging Procedure, which combines information from ground-based observations, state-of-the-art precipitation products, and topography-related features to improve the representation of the spatiotemporal distribution of precipitation in Chile. Herman and Schumacher (2018a, 2018b) conducted a series of experiments on the improvement of precipitation forecasts using the RF-based tree models and came to the very "valuable" conclusion that "*Money Does not Grow on Trees, but Forecasts Do*". The RF model was also successfully used to implement automatic detection and classification of low-level orographic precipitation processes from space-borne radars (Arulraj & Barros, 2021). Precipitation error correction models based on streamflow observations and RF models have also been experimented at the global scale (Beck et al., 2020).

Whether single-source or multi-source, corrected or merged, the studies mentioned above perform precipitation post-processing in a deterministic way. However, because of the imperfect nature of the model and the

ineradicable nature of the uncertainty, deterministic models do not always convey a consistent message and do not take full advantage of multi-source information (Donat et al., 2014; Henn et al., 2018). When these deterministic products are applied, the uncertainty may also be amplified with the model or over time (Cunha et al., 2012; Pan et al., 2010; Schreiner McGraw & Ajami, 2020). In contrast, probabilistic post-processing outputs not only provide deterministic information, but also quantify uncertainty and a basis to measure and manage the risk of decision making (Parrish et al., 2012; S. Wang et al., 2018). Few probabilistic products have been developed to achieve this objective, such as observation-based global temperature products (HadCRUT4, 100 ensemble members; Morice et al., 2012), high-resolution ensemble precipitation analysis (HREPA, 24 ensemble members; Khedhaouiria et al., 2020), and ensemble meteorological data set for North America (EMDNA, 100 ensemble members; Tang et al., 2021). For post-processing of near-real-time satellite precipitation estimates, its uncertainty information is even as important as deterministic information, yet almost no studies discussed this issue.

Quantile regression forests (QRF) model is a variant of the RF model that not only predicts the conditional mean of the predictand, but also provides the full conditional probability distributions (Meinshausen & Ridgeway, 2006). Thus, the QRF model inherits all the advantages of the RF model and provides additional probabilistic information. There have been few studies in ensemble post-processing of numerical weather forecasts based on the QRF model. Fox example, Taillardat et al. (2016) developed the QRF model and compared it with ensemble model statistics (EMOS) for post-processing surface temperature and wind speed. Results indicated that the QRF model performs better than EMOS and can bring additional value to the human forecaster. Evin et al. (2021) proposed using the QRF model to calibrate ensemble forecasts of the height of new snow, which also indicated that QRF could be applied to the correction of skewed distribution for variable similar to precipitation. In addition to the post-processing of MWPs, QRF has also been successfully applied to hydrological ensemble post-processing (Tyralis & Papacharalampous, 2021; Tyralis et al., 2019) and soil uncertainty mapping (Kasraei et al., 2021; Vaysse & Lagacherie, 2017). However, no studies have attempted to use it for near-real-time post-processing of satellite precipitation estimates.

In this study, we propose the development of a QRF-based postprocessor called QRF4P-NRT (**P**robabilistic **P**ost-**P**rocessing of **Near-real-time** Satellite **P**recipitation Estimates using QRF). We further demonstrate the proposed framework as a simple, multi-stage probabilistic post-processing method capable of providing dry-wet classification, deterministic and probabilistic prediction for near-real-time satellite precipitation estimates. Using the Yalong River basin in China as a case study, the more accurate dry-wet classification, deterministic adjustments, and reliable probabilistic outputs are confirmed by the comparison between the raw IMERG-E and IMERG-F products. The structure of this study is organized as follows in this manuscript: study area and data set are introduced in Section 2. Section 3 describes in details the architecture of the proposed method, its implementation, the inputs and outputs, and the final evaluation measures. Section 4 presents the analysis and results. Critical issues are discussed in Section 5. Main conclusions are summarized in the last section.

2. Study Area and Data

2.1. Study Area

The Yalong River, located in the western part of the Sichuan Basin and the eastern part of Qinghai-Tibet Plateau, is selected as the region of interest of this study. Yalong River is the largest tributary of the Jinsha River in the upper reaches of the Yangtze River of China (Figure 1a). The Yalong River basin (96°52'–102°48'E, 26°32'–33°58'N) spans 1,570 km, covers a total area of about 130,000 km^2 and presents significant topographic difference from north to south (from 7,148 to 115 m). The shape of the Yalong River basin is long and narrow, surrounded by high mountains and canyons. The upper reach of the basin is in a continental climate (cold and relatively dry), with a mean annual precipitation of 500–750 mm. The middle and lower reaches of the basin have relatively high temperatures and precipitation rates, with a mean annual precipitation of 750–1,500 mm (Figure 1b). Precipitation in Yalong River basin has a strong seasonal pattern, with dry periods in November–May and rainy seasons in June–October. The complexity of the terrain and the diversity of precipitation make the Yalong River basin an ideal study area to evaluate its corresponding satellite precipitation estimates and probabilistic post-processing.

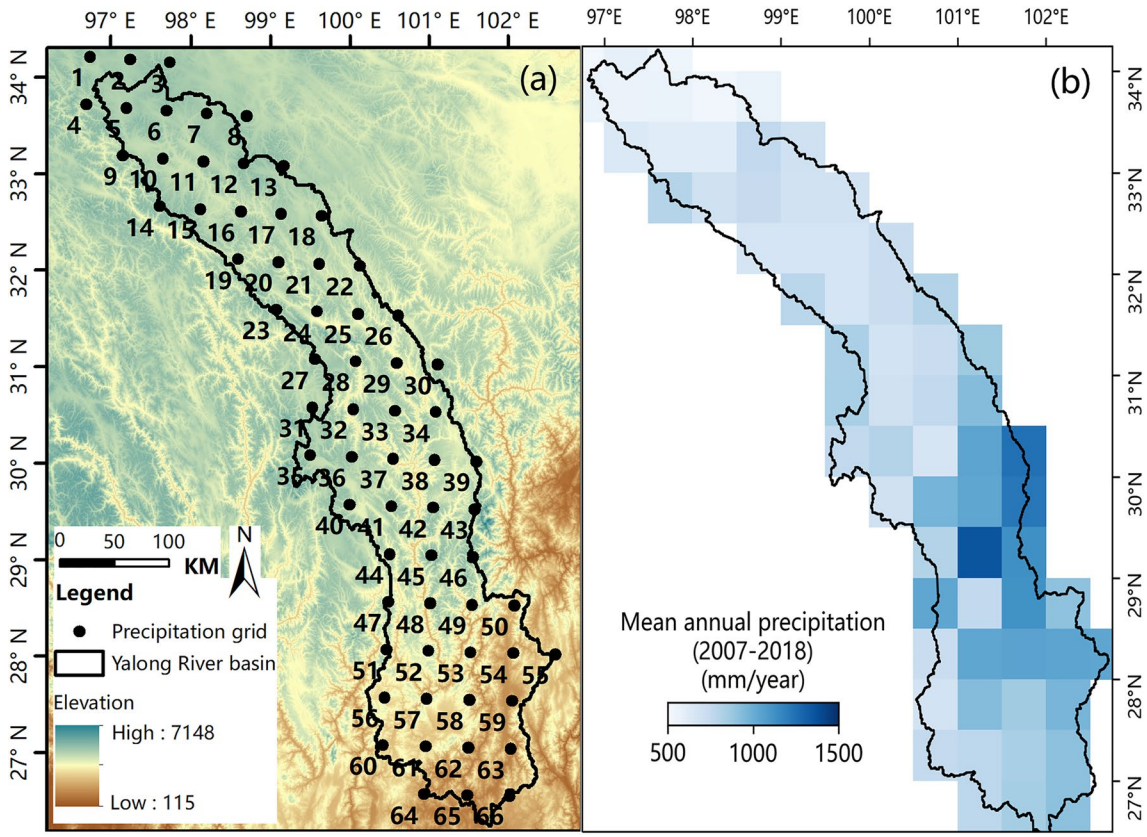

Figure 1. Study area. (a) Topography and precipitation grid cells and (b) precipitation climatology.

2.2. Data

The data used in this study include two types of satellite precipitation estimates products, one ground observation product, geographic data, intra-annual indicator, meteorological data, and subsurface state variable. They are described in details below.

2.2.1. Satellite Precipitation Estimates Products

The most commonly used Level 3 products in NASA's Global Precipitation Measurement project include three types of data sets, the IMERG Early Run (IMERG-E), IMERG Late Run (IMERG-L) and IMERG Final Run (IMERG-F; Huffman et al., 2015, 2019a, 2019b). Among them, IMERG-E and IMERG-L belong to near-real-time precipitation estimates. The minimum latency time of IMERG-E is 4 hr, and the minimum latency time of IMERG-L is 12 hr. IMERG-F is a bias-corrected, research-quality post-processing product with a 3.5 months latency time (https://gpm.nasa.gov/data/directory). In this study, we select the IMERG-E product as a representative for conducting probabilistic post-processing of near-real-time satellite precipitation estimates. And IMERG-F is selected as a reference for verifying the model performance of the proposed post-processing model. The quality of IMERG products has also been confirmed in our previous studies by comparing them with other satellite precipitation estimates products (Y. Zhang et al., 2021a, 2021b). The IMERG-E and IMERG-F data selected in this study are the latest IMERG version 6 products. A total of 12 yr of precipitation data (1 January 2007 to 31 December 2018) are used here for the analysis. The spatiotemporal resolution of the data is 1 day and 0.1°.

2.2.2. Reference Data

An observed gridded precipitation data set with a 0.5° spatial and daily temporal resolution from the National Meteorological Data Center of China Meteorological Administration (CMA) is selected as a reference. The data set with high accuracy and wide applications was developed by interpolating the high-quality precipitation gauge observations from more than 2,400 weather stations (approximately 50 gauge stations over our study area) over China using the Global 30 Arc Second Elevation Data Set (GTOPO30) and the thin plate smooth spline (TPS)

method (Y. Zhao et al., 2014). CMA gridded precipitation data set has also been used as reference data in many other studies in China (B. Guo et al., 2020; H. Guo et al., 2016; Qiang et al., 2016; Y. Zhang & Ye, 2021). In this study, we mainly investigate the reliability and effectiveness of the probabilistic post-processing model. So we taking the CMA observation as the ground "truth", and the sampling errors caused by the gauge network density and instruments are ignored. When more accurate and representative ground observations are available, we can retrain the model to obtain a more accurate model between satellite precipitation estimates and ground "truth". Similarly, because of the gauge density, the spatial resolution of the ground reference is 0.5°, so we upscale the satellite product to 0.5° and then train the model at a spatial scale of 0.5°. When finer resolution references are available, finer resolution models can be retrained directly. A total of 12 yr of precipitation reference data (1 January 2007 to 31 December 2018) are selected to maintain time overlap. There are 66 grid cells within the Yalong River basin, and their spatial distribution is shown in Figure 1a.

2.2.3. Auxiliary Data

Precipitation climatology varies with grid cells, which may be affected by latitude, longitude, and elevation (Basist et al., 1994). Therefore, we select the station number (ID), longitude (Lon), Latitude (Lat), and elevation (Ele) of each grid cell as the geographic information. The elevation data is obtained from NASA Shuttle Radar Topographic Mission digital elevation model, the original spatial resolution is 90 m. We extract the information at each grid center point as the representatives of the grid cell. Intra-annual variation in precipitation characteristics is also present. Day of year (DOY) is selected as an indicator.

Like geographic information, meteorological variables that correlate and interact with precipitation at different locations can reflect the local precipitation characteristics (Aleshina et al., 2021; Back & Bretherton, 2005; Déry & Wood, 2005; H. Li et al., 2021; Pendergrass & Hartmann, 2014; Sahin, 2012). Therefore, they can also be used as auxiliary data to retrieve precipitation estimates. Here, we selected six meteorological variables, including surface downward longwave radiation (Lard), surface downward shortwave radiation (Srad), near-surface air temperature (Temp), near-surface air pressure (Pres), near-surface air specific humidity (Shum), near-surface wind speed (Wind). They were obtained from the China meteorological forcing data set (CMFD) provided by the National Tibetan Plateau Data Center of China (J. He et al., 2020). This data set has been validated and widely used in many studies in China (e.g., Gou et al., 2021; Huang et al., 2020; Lu et al., 2020). These reanalysis data with a temporal resolution of 3 hr and a spatial resolution of 0.1° were aggregated to a daily scale and a spatial resolution of 0.5° using daily and spatial averaging. In addition, surface soil moisture (SM) is also used as an auxiliary variable to reflect the dry and wet state of the ground surface. The surface (0–7 cm) SM data selected for this study were obtained from the ERA5-land reanalysis products provided by the European Centre for Medium-Range Weather Forecasts (ECMWF), with a temporal resolution of 1 hr and a spatial resolution of 0.1° (Muñoz-Sabater et al., 2021). The ERA5 SM product has been validated in our earlier study (H. Li et al., 2021). Similarly, they are handled at daily scale and 0.5°. The time period for these data sets is 1 January 2007 to 31 December 2018.

3. Methodology

We first introduce the basics of QRF model in Sections 3.1 and 3.2. Then in Section 3.3 we introduce the workflow of our proposed method and conducted experiments. Finally, we describe a set of evaluation metrics in Section 3.4.

3.1. Decision Tree and Random Forests

Decision tree (DT) is a supervised ML algorithm based on top-down conditional judgments. A typical DT mainly consists of three types of nodes: root node, split node, and leaf node (Figure 2; Song & Ying, 2015). The DT grows up from the root node, classifies the samples according to their predictors (or features) and the node conditions, goes to the split node and selects the samples with a similar process in the root node again until to the leaf node. The growth of a typical DT is completed under the control of the leaf node. Finally, the prediction (classification or regression) is obtained at the leaf node.

Classification and regression tree (CART), proposed by B. Li et al. (1984), is a binary DT model. For the classification task, the principle of splitting any non-leaf node is based on the Gini index, which is a metric similar to

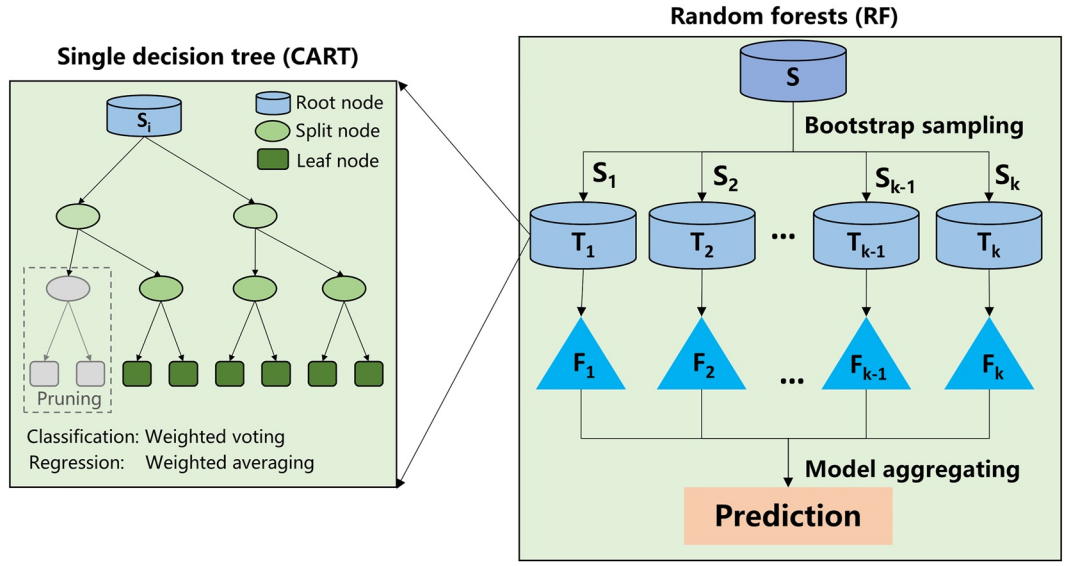

Figure 2. The link between the classification and regression tree (CART) and random forest (RF) model.

information entropy used to describe the purity of sample features. For the regression task, the splitting principle for any non-leaf node is based on minimizing the mean squared difference between the observations and predictions as the loss function, also known as the variance reduction criterion. The equations are given as follows:

$$\text{Gini}(S) = \sum_{i=1}^{m} P_i (1 - P_i) = 1 - \sum_{i=1}^{m} P_i^2 \tag{1}$$

where S is a specific sample; P_i represents the probability of positive or negative classes appearing in the sample set; m is the number of sample features.

$$\text{Loss} = \frac{1}{N} \sum_{j=1}^{N} \sqrt{(y_j - \hat{y}_j)^2} \tag{2}$$

where N is the number of samples, y_j is a specific observation, and $\hat{y}_j$ is the corresponding prediction.

CART is a simple non-parametric model with low computational complexity, no need for distribution assumptions and normalization of the samples. However, because of the growth process of the CART, it also has obvious drawbacks. For example, a CART may grow too deep, resulting in an over-complicated model, which leads to overfitting. CART is essentially a greedy algorithm, and there is randomness in the growth process of a single CART. Therefore, it is sensitive to the training samples and may converge to the local optimum. A pruning-like approach was proposed to mitigate the drawbacks of CART, but it is not a highly desirable solution.

Inspired by the idea of ensemble learning, Breiman (2001) proposed the RF model. RF model is to generate K individual CART to form huge "forests". Such an integrated idea is based on the collective decision of Bootstrap aggregating sampling (Bagging). The link between RF and CART is shown in Figure 2.

The implementation of the RF model consists of two steps:

1. **Determine K CARTs and use bootstrap sampling to generate the initial samples of each tree.** Given a data set S with N samples, N' samples ($N' \leq N$) are randomly selected as subsamples S_1, S_2, S_3,…,S_{k-1}, S_k of each CART. Therefore, this is where the first manifestation of the randomness of the RF model comes into play. It should be noted that the bootstrap sampling is not a simple replication of S, but a reconstruction that approximates the original sample space, which ensures both the difference and the similarity with the original feature space.

2. **Construct each CART using independent subsample S_i and form the final forests.** Another randomness of the RF model is mainly reflected in the feature selection when each non-leaf node splits. At any non-leaf node of an RF model containing K CARTs, a certain number of features are randomly selected from the whole feature space as the basis, and then node splitting is performed using Gini index (classification) and variance reduction criterion (regression). The random selection of feature space greatly reduces the similarity between different nodes and different CARTs in the same RF model, making it more robust.

The two types of randomness in the above steps make the RF model a comprehensive improvement compared with CARTs, which can effectively prevent overfitting and eliminate pruning in CARTs. In addition to the advantage of "forest" to prevent overfitting, the integration of multiple CARTs also solves the problem of single CART prediction falling into local optimum and greatly improves the ability of the algorithm to converge to the global optimum.

Following the previous notations (Breiman, 2001; A. H. Li & Martin, 2017), the prediction inference processes of the RF model are as follows:

Given a data set S with N samples, each sample in S consists of a predictor X_i containing p-dimensional features and a predictand (or target) Y_i. They are expressed as follows:

$$S = (Y_i,\ X_i)\,, i = 1, \ldots, N \tag{3}$$

$$X_i = x_i^1,\ x_i^2,\ x_i^3, \ldots,\ x_i^p \epsilon \mathbb{R}^p \tag{4}$$

In data set S, for the dry-wet classification task, Y_i is 0 or 1, that is, $Y_i = \{0, 1\}$, where 0 represents a dry event and 1 represents a wet event. For the regression task, Y_i is a continuous random variable representing the ground "truth" of precipitation on a specific day, which in this study is the CMA observation.

Define θ as the internal parameter that determines how the RF is generated, then the single CART can be expressed as $T(\theta)$. Suppose the single CART contains a total of l leaf nodes, R_l is the subspace obtained by the lth leaf node for the original factor space through a split. For any sample factor $x \in X$, an R_l can be found with $x \in R_l$, and the sample is noted as $l(x, \theta)$.

For each CART of RF model, given a new sample predictor $X = x$, solving the predictand $\hat{Y}$ means taking an equally weighted average of the sample values in all leaf nodes $l(x, \theta)$, which is expressed as follows:

$$\hat{Y}(x,\ \theta) = \sum_{i=1}^{n} \omega\,(X_i,\ x,\ \theta)\, Y_i \tag{5}$$

$$\omega\,(X_i, x, \theta) = \frac{1_{\{X_i \in R_{\ell(x,\theta)}\}}}{\#\{j\ :\ X_j \in R_{\ell(x,\theta)}\}} \tag{6}$$

$$\sum_{i=1}^{n} \omega\,(X_i,\ x,\ \theta) = 1 \tag{7}$$

Finally, extending single CART to K CARTs, the conditional mean $E(Y|X = x)$ is estimated by the averaged prediction of K CARTs, which is expressed as follows:

$$\hat{Y}(x) = \sum_{i=1}^{n} \omega\,(X_i,\ x)\, Y_i \tag{8}$$

$$\omega\,(X_i,\ x) = \frac{1}{K} \sum_{k=1}^{K} \omega\,(X_i,\ x,\ \theta_k) \tag{9}$$

$$\sum_{i=1}^{n} \omega\,(X_i,\ x) = 1 \tag{10}$$

where $\omega(X_i, x, \theta_k)$ is the weight matrix of the kth CART. For the classification task, all sample values are voted, while all sample values are averaged for the regression task.

3.2. Quantile and Quantile Regression Forests

Given any random variable X, whose distribution function is $F(x) = P(X \leqslant x)$. Then given any quantile $\tau \in (0, 1)$, we define the τ quantile function of the random variable X as:

$$Q(\tau) = F^{-1}(x) = \inf\{x : F(x) \geqslant \tau\} \tag{11}$$

According to the definition of the quantile function $Q(\tau)$, it can be seen that there exists a proportion of τ (or $1 - \tau$) samples larger (or smaller) than the quantile function $Q(\tau)$, respectively. Quantile regression models are built at specific quantile levels about the conditional distribution of the target variable, which can describe the global distribution characteristics of the target variable and therefore contain more uncertainty information (Meinshausen & Ridgeway, 2006). Conditional mean focuses on the mean state of the target variable, while quantile focuses on the global distribution; Empirical quantile modeling does not obey strict assumptions (e.g., normality, independence, and homoscedasticity). Therefore, quantile modeling is a more general framework of the conditional mean modeling.

The quantile loss function can be further defined as:

$$\rho_\tau(u) = \begin{cases} \tau \cdot u & \text{if } u \geq 0 \\ (\tau - 1) \cdot u & \text{if } u < 0 \end{cases} \tag{12}$$

where u can be regarded as samples larger (or smaller) than the quantile function, which ensures the quantile loss function to always be positive.

In regression inference using the RF model, the prediction is approximated by conditional mean, which is an equal-weighted averaging method that does not take full advantage of the CDF of the whole samples. Solving the RF model can be expressed as:

$$\hat{Y}(x) = \underset{\lambda \in \mathbb{R}}{\operatorname{argmin}} \sum_{i=1}^{n} \omega(X_i, x)(Y_i - \lambda)^2 \tag{13}$$

In the QRF model, the prediction uses the entire CDF information and is a non-equal-weight approach. Solving the QRF model can be expressed as:

$$\hat{Y}_\tau(x) = \underset{\lambda \in \mathbb{R}}{\operatorname{argmin}} \sum_{i=1}^{n} \omega(X_i, x)\rho_\tau(Y_i - \lambda) \tag{14}$$

Given any quantile $\tau \in (0, 1)$, the loss function $\text{Loss}(\tau)$ can be expressed as:

$$\text{Loss}(\tau) = \sum_{i=1}^{n} \omega(X_i, x)\left\{\sum_{i:u_i \geq 0}(\tau \cdot u_i)^2 + \sum_{i:u_i < 0}[(\tau - 1) \cdot u_i]^2\right\} \tag{15}$$

$$u_i = Y_i - \hat{Y}_\tau(x) \tag{16}$$

The above equation allows finding the predictand at different quantile levels. The prediction intervals (PIs) can also be obtained. For example, $\text{PI}_{50\%}$ and $\text{PI}_{90\%}$ can be given by the following equations:

$$\text{PI}_{50\%}(x) = [Y_{0.25}(x), Y_{0.75}(x)] \tag{17}$$

$$\text{PI}_{90\%}(x) = [Y_{0.05}(x), Y_{0.95}(x)] \tag{18}$$

3.3. Post-Processing Workflow

The workflow of the QRF4P-NRT is shown in Figure 3. Due to the resolution of the reference data, we build the core post-processing model on a coarse grid system. Therefore, the proposed workflow is to first upscale the raw satellite precipitation estimates and auxiliary predictors, and then perform a two-stage post-processing model

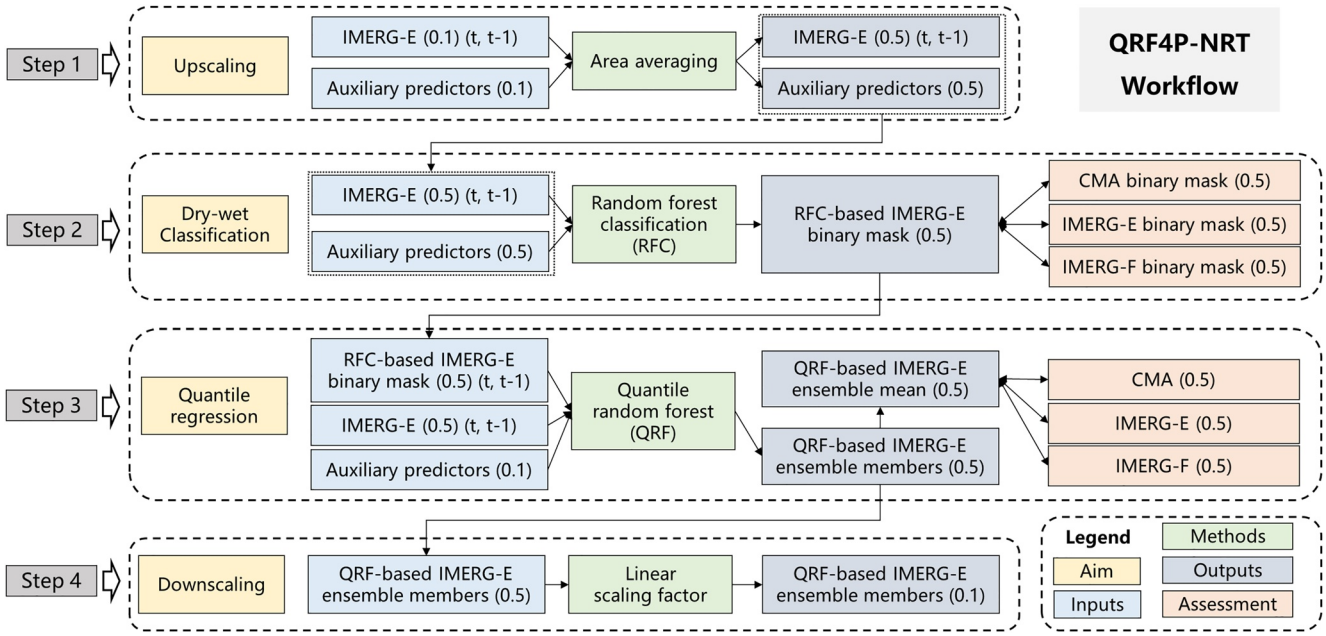

Figure 3. The workflow of the QRF4P-NRT. The 0.1 and 0.5 in the figure represent the 0.1° and 0.5°. Time step t and $t-1$ represent the target day and the previous day.

to generate the binary classification (wet-dry) mask and probabilistic outputs, respectively. In this process, the model is evaluated based on raw IMERG-E and bias-corrected IMERG-F satellite precipitation estimates. Finally, the fine-resolution satellite precipitation estimates are reconstructed using a simple scaling factor method.

3.3.1. Feature Selection and Feature Importance

For training an ML model, an indispensable step is feature engineering. To develop a near-real-time post-processing system for satellite precipitation estimates, auxiliary predictors acquired in parallel with satellite precipitation estimates are key to feature engineering. We first summarize the selected predictors for this study in Table 1, and then describe them in details below.

First, with respect to the main predictors (primary variable), the near-real-time satellite estimates (Sim^t) are considered to be the most appropriate and dominant predictors. Moreover, two situations in near-real-time satellite precipitation retrieval (e.g., temporal and spatial correlation) are considered (Y. Zhang & Ye, 2021). Therefore, for the main predictors, we not only selected the precipitation estimates at the central grid cell on a specific day (t), but also selected the rest eight grid cells in the neighboring 3×3 grid cells and all precipitation estimates in this 3×3 grid cells on the previous day ($t-1$), with a total of 18-dimensional features.

In terms of auxiliary predictors (or covariates), as mentioned in the data section of this article, we selected a variety of predictors, including geographic predictors, meteorological predictors, DOY index, and surface SM. Therefore, we also must consider the near-real-time accessibility of the auxiliary predictors. Among these predictors, the geographic factors (including station number, latitude, longitude, and elevation) and the DOY index are considered as permanent static predictors with five-dimensional features. The meteorological and surface SM data selected in this study were derived from reanalysis products. They are not available in near-real-time. We, therefore, focus on how to process and design experiments to incorporate these predictors into the proposed post-processing framework.

In order to increase the interpretability and physical significance of ML models, RF incidental feature importance analysis is attempted in this study (Breiaman, 2001). Quantification of the importance of features is mainly performed by out-of-bag sampling, which is a stepwise testing method. That is, the features are selectively controlled to observe the effect of feature variation on the results. And Gini index and the root-mean-square error are used to quantify the importance of each feature in classification and regression tasks, respectively.

Table 1
Predictors and Their Abbreviation Used in This Study

Variable name	Abbreviation	Unit
Precipitation in the target gird cell on the target day (t)	Sim^t	mm/day
Precipitation at adjacent grid cell (northwest) on the target day (t)	Sim_0^t	
Precipitation at adjacent grid cell (north) on the target day (t)	Sim_1^t	
Precipitation at adjacent grid cell (northeast) on the target day (t)	Sim_2^t	
Precipitation at adjacent grid cell (west) on the target day (t)	Sim_3^t	
Precipitation at adjacent grid cell (east) on the target day (t)	Sim_4^t	
Precipitation at adjacent grid cell (southwest) on the target day (t)	Sim_5^t	
Precipitation at adjacent grid cell (south) on target day (t)	Sim_6^t	
Precipitation at adjacent grid cell (southeast) on target day (t)	Sim_7^t	
Precipitation in the target gird cell on the previous day ($t-1$)	Sim^{t-1}	
Precipitation at adjacent grid cell (northwest) on the previous day ($t-1$)	Sim_0^{t-1}	
Precipitation at adjacent grid cell (north) on the previous day ($t-1$)	Sim_1^{t-1}	
Precipitation at adjacent grid cell (northeast) on the previous day ($t-1$)	Sim_2^{t-1}	
Precipitation at adjacent grid cell (west) on the previous day ($t-1$)	Sim_3^{t-1}	
Precipitation at adjacent grid cell (east) on the previous day ($t-1$)	Sim_4^{t-1}	
Precipitation at adjacent grid cell (southwest) on the previous day ($t-1$)	Sim_5^{t-1}	
Precipitation at adjacent grid cell (south) on the previous day ($t-1$)	Sim_6^{t-1}	
Precipitation at adjacent grid cell (southeast) on the previous day ($t-1$)	Sim_7^{t-1}	
Station number	ID	-
Longitude	Lon	-
Latitude	Lat	-
Elevation	Ele	m
Day of year index	DOY	-
Surface downward longwave radiation	Lrad	W/m^2
Surface downward shortwave radiation	Srad	
Near-surface air temperature	Temp	K
Near-surface air pressure	Pres	Pa
Near-surface air specific humidity	Shum	kg/kg
Near-surface wind speed	Wind	m/s
Soil moisture	SM	m^3/m^3

3.3.2. Experimental Design

Based on the discussion above, we first determined a total of 18-dimensional features for the main predictors and a total of five-dimensional features for the permanent static predictors. This is also our base experiment, containing 23-dimensional features. On this basis, we set up three different groups of experiments dealing with other predictors. They are described below and summarized in Table 2.

1. **Base experiment**: contains only 23-dimensional features with main predictors and permanent static predictors.
2. **Static experiment**: take the multiyear average (climatology) of the meteorological predictors and SM predictor as their real-time proxies that can be obtained during the operational post-processing system.
3. **Analog experiment**: during the operational post-processing, the most analogs meteorological and SM factors from the historical (training) period are selected as real-time proxies by the analog method (Hemri & Klein, 2017). This method is one of the most classic statistical post-processing methods for ensemble forecasting, and here we choose it as an alternative for finding near-real-time dynamic predictors (Zorita & Von Storch, 1999). The metric used here is implemented by calculating the Euclidean distance (Equation 19)

Table 2
Experimental Design

ID	Experiment	Predictors	Predictand	Dimensionality
1	Base	ID, Lon, Lat, Ele, DOY, $Sim^{t}_{3\times3}$, $Sim^{t-1}_{3\times3}$	O^t	23
2	Static	ID, Lon, Lat, Ele, DOY, $Sim^{t}_{3\times3}$, $Sim^{t-1}_{3\times3}$, Lrad, Pres, Shum, Srad, Temp, Wind, SM	O^t	30
3	Analog	ID, Lon, Lat, Ele, DOY, $Sim^{t}_{3\times3}$, $Sim^{t-1}_{3\times3}$, Lrad, Pres, Shum, Srad, Temp, Wind, SM	O^t	30
4	Dynamic	ID, Lon, Lat, Ele, DOY, $Sim^{t}_{3\times3}$, $Sim^{t-1}_{3\times3}$, Lrad, Pres, Shum, Srad, Temp, Wind, SM	O^t	30

Note. Predictors in the table can be referred to Table 1. Predictand (target variable) is the ground observation on target day (O^t).

between near-real-time precipitation estimates in 3 × 3 grid cells for two adjacent days (t and $t-1$) of the historical (training) and future (test) period.

$$\text{Euclidean Distance} = \sqrt{\sum_{i=1}^{18}\left(x_i^{\text{future}} - x_i^{\text{history}}\right)^2} \tag{19}$$

4. **Dynamic experiment**: The dynamic experiment is a hypothetical experiment in which it is assumed that we can obtain real-time dynamic factors for an operational post-processing system in future periods. In this study, it can also be seen as the upper bound of model performance.

3.3.3. Classification and Regression

In probabilistic post-processing experiments, the precipitation is criticized with characteristics and difficulties such as skewed, discrete-continuous distribution, and heteroskedasticity. To improve the skill of the proposed framework, probabilistic post-processing is divided into two steps. In the first step, dry-wet event discrimination based on random forest classification (RFC) is performed. Dry events were determined in this study using a threshold of 1 mm/day. In the second step, the dry events are seen as "truncated" values, then the QRF is implemented only using wet samples. In the QRF part, the quantile is equally sampled from 0.05 to 0.995, and a total of 100 set ensemble members are generated.

3.3.4. Downscale to the Original Resolution

To unify the spatial resolution of satellite precipitation estimates and observations, we upscaled the original satellite precipitation estimates to 0.5° in the probabilistic post-processing model. This degrades the original high spatial resolution. Therefore, after obtaining the calibrated post-processing ensemble members, we reconstruct them to the original resolution using a scaling factor as follows:

$$\widetilde{P_{0.1}} = P_{0.1} \times \frac{\widetilde{P_{0.5}}}{P_{0.5}} \tag{20}$$

where $\widetilde{P_{0.5}}$ and $P_{0.5}$ represent bias-corrected and raw precipitation estimates at 0.5° grid cells, mm/day; $\widetilde{P_{0.1}}$ and $P_{0.1}$ represent bias-corrected and raw precipitation estimates at 0.1° grid cells corresponding to the 0.5° grid cells, mm/day.

3.4. Performance Measures

3.4.1. Model Training and Validating

To train and test the model more fairly, the whole sample set (2007–2018) is first divided into a training set (2007–2016, 10 yr) and a test set (2017–2018, 2 yr). The training period is used to tune the model hyperparameters to obtain an offline model; the test set is used for future operationalization to test the online model. The training set is randomly divided into five parts (2 yr per part) for 5-fold cross-validation to fully train and tune the model hyperparameters. And at this stage, the test set (2017–2018) is not involved in any training or hyperparameter tuning process. After hyperparameters were tuned by cross-validation, we retrain the model on the entire training set (2007–2016) using the optimal hyperparameters. The tuned model is then tested and evaluated on the independent test set (2017–2018). The results of the training set are also not shown in the later section

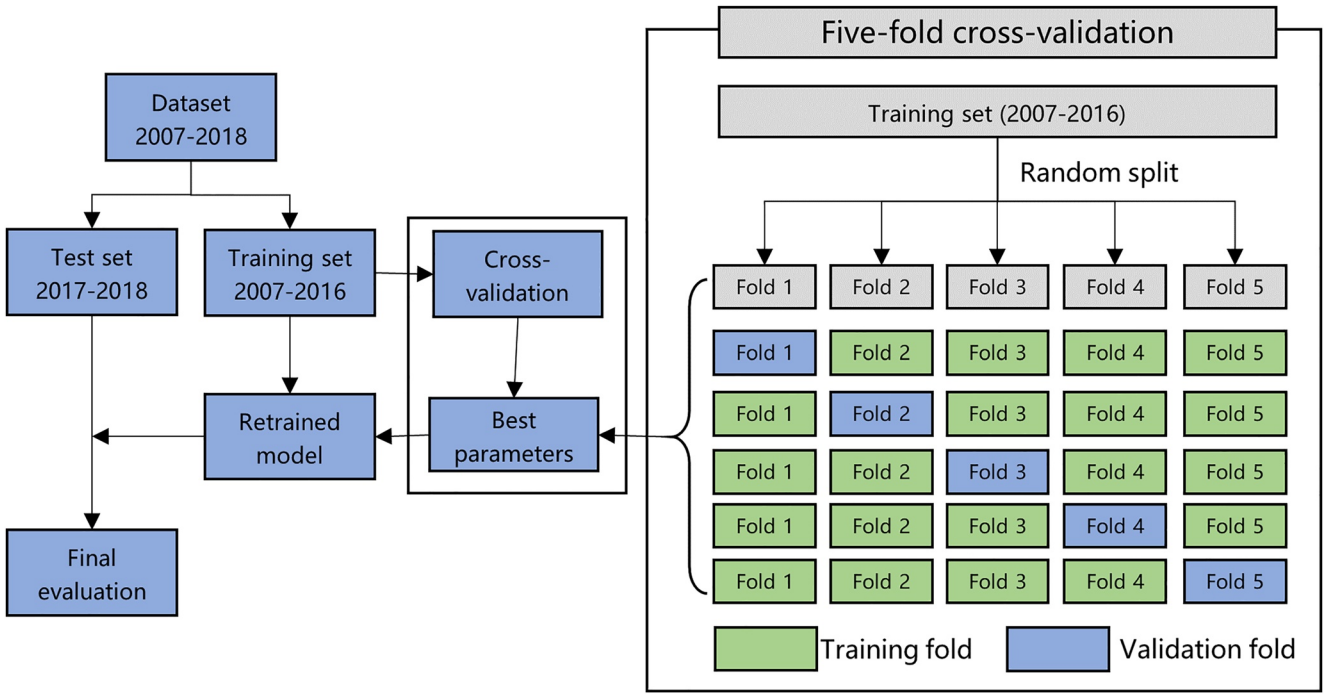

Figure 4. Data set split, model training, and evaluation. This figure is modified from Python scikit-learn package (https://scikit-learn.org/stable/modules/cross_validation.html#k-fold).

because they are not representative. The above process is recommended by scikit-learn ML package (Pedregosa et al., 2011). The data set split, cross-validation, model training, and evaluation are shown in Figure 4. Sensitive hyperparameters in the RF model include the number of trees (K), the number of predictors randomly sampled from total predictors (N) as candidates in non-leaf nodes (N_{try}), and the minimum number of samples in leaf nodes (N_{leaf}). Hyperparameter tuning is achieved through a combination of random grid search and grid search. For the classification task, K is 20, N_{try} is sqrt(N), and N_{leaf} is 1. For the regression task, K is 100, N_{try} is sqrt(N), and N_{leaf} is 5.

3.4.2. Probabilistic Metrics

According to the precipitation 2 × 2 contingency table, compared with the ground "truth", the satellite precipitation estimates can be classified into four types of events: hit events (TP), missed events (FN), false alarm events (FP) and correct negative events (TN). Then two commonly used probabilistic metrics probability of detection (POD) and false alarm ratio (FAR) can be expressed as follows:

$$\mathrm{POD} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{21}$$

$$\mathrm{FAR} = \frac{\mathrm{FP}}{\mathrm{TP} + \mathrm{FP}} \tag{22}$$

The values of POD and FAR range from 0 to 1. The larger the POD the better, and the smaller the FAR the better.

3.4.3. Continuous Metrics

Three commonly used continuous metrics are selected including Pearson correlation coefficient (PCC), mean error (ME, also known as bias), and root mean square error (RMSE). The equations are given as follows:

$$\mathrm{PCC} = \frac{\sum_{i=1}^{n}\left(O_i - \overline{O}\right)\left(S_i - \overline{S}\right)}{\sqrt{\sum_{i=1}^{n}\left(O_i - \overline{O}\right)^2}\sqrt{\sum_{i=1}^{n}\left(S_i - \overline{S}\right)^2}} \tag{23}$$

$$\mathrm{ME} = \frac{1}{n}\sum_{i=1}^{n}\left(S_i - O_i\right) \tag{24}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(S_i - O_i\right)^2} \tag{25}$$

where O_i and S_i represent observations and satellite precipitation estimates at timestep i, respectively. The value range of PCC is 0–1, the closer to 1 the better. ME $\in (-\infty, +\infty)$ and RMSE $\in [0, +\infty)$. The closer ME (or RMSE) is to 0, the smaller the total errors.

In addition to the above three commonly used metrics, we also selected a four-component error decomposition method (4CED), that is, the total bias of precipitation products can be decomposed into hit positive bias (HPB), hit negative bias (HNB), false alarm bias (FB), and missed bias (MB). The 4CED can be used to trace and measure the sources and magnitudes of individual error components (Y. Zhang et al., 2021b).

3.4.4. Ensemble Scoring Rules

The evaluation of ensemble post-processing uses metrics that describe the accuracy, reliability, and sharpness, including the continuous rank probability score, rank histogram, reliability diagram and PIs.

3.4.4.1. Continuous Rank Probability Score (CRPS)

The CRPS is an integrated scoring metric and one of the most used probabilistic tools for evaluating the ensemble skill in terms of accuracy, reliability, and sharpness (Bröcker, 2012). For given ensemble outputs, the CRPS corresponds to the integrated quadratic distance between the cumulative distribution function (CDF) of the ensemble outputs and the observation. According to previous studies (Gneiting & Ranjan, 2011; P. Zhao et al., 2022), we also use twCPRS to assess the model performance on predicting heavy precipitation events. CRPS and twCRPS can be expressed as:

$$\mathrm{CRPS} = \int_{-\infty}^{\infty}\left(F\left(P_t\right) - O\left(P_t\right)\right)^2 dP_t \tag{26}$$

$$tw\mathrm{CRPS} = \int_{-\infty}^{\infty}\left(F\left(P_t\right) - O\left(P_t\right)\right)^2 \omega(P) dP_t \tag{27}$$

where $\omega(P)$ is a weight function that equals 1 (or 0) if $P \geq q$ (or $P < q$); and q (95% in this study) is a given threshold. P_t is a specific threshold of precipitation; $F(P_t)$ is the CDF obtained from the ensemble outputs for day t; $O(P_t)$ is the Heaviside function, and it can be expressed as:

$$O\left(P_t\right) = \begin{cases} 1, & O_i > P_t \\ 0, & O_i \leq P_t \end{cases} \tag{28}$$

The closer the CRPS (twCRPS) value is to 0, the better the ensemble forecast is, and the CPRS (twCRPS) value equals 0 for a perfect ensemble post-processing system.

3.4.4.2. Rank Histogram

Rank histogram, as known as the Talagrand diagram, is a commonly used tool to assess the reliability of the ensemble forecast systems (Hamill, 2001). If the ensemble outputs obtained by the ensemble post-processing model are sufficiently reliable, the ranking of the observation in the CDFs of the ensemble outputs should be uniformly distributed. In addition to being able to assess the reliability of ensemble outputs, the rank histogram can also be used to measure the ensemble dispersion (spread skill) and the presence of systematic bias in the ensemble post-processing system by its distribution. If the rank histogram is more uniform and flatter, the closer the ensemble system is to perfection. Conversely, a "U" shaped rank histogram indicates that the ensemble outputs are under-dispersed, while a bell-shaped rank histogram indicates that the model is over-dispersed. The other two rank histogram shapes "\" and "/" indicate the existence of systematic overestimation and systematic underestimation of the ensemble post-processing model, respectively. In this study, 20 ensemble members in the rank histogram are sampled equally from the original 100 ensemble quantiles. Following the proposed method of Bröcker and Ben Bouallègue (2020), we combine the ensemble members and observations in the same rank histogram. To eliminate the effect of truncated values, only wet events are selected to calculate the displayed rank histograms.

3.4.4.3. Prediction Intervals (PIs)

In addition to CRPS and rank histograms, the PI is another scoring rule that is very intuitive and commonly used to describe the sharpness of the ensemble outputs. Here, we use the two most used PIs (e.g., 50% and 90%; Gneiting et al., 2007).

3.4.4.4. Reliability Diagram

The reliability diagram is also applied in this study to assess how well the predicted probabilities of an event correspond to their observed frequencies (Hartmann et al., 2002). To better assess the reliability of extreme events, three commonly used thresholds (i.e., 80%, 90%, 95%) are selected according to the previous study (Yang et al., 2021). We then pool ensemble outputs of different gird cells and days to calculate the probability. Perfect reliable predictions in the reliability diagram will fit the diagonal (1:1) line. Predictions plots above (or under) this line indicates underestimation (or overestimation).

4. Results

This section first shows the results of the dry-wet classification (Section 4.1), and then compares the skill of the ensemble members obtained from different experiments (Section 4.2) and the deterministic results of the ensemble mean (Section 4.3). Finally, the reconstructed fine-resolution spatial distribution of precipitation is briefly shown (Section 4.4).

4.1. Dry-Wet Classification

First, based on the ground reference, the performance for dry-wet events classification of the raw IMERG-E, different RFC experiments outputs, and IMERG-F obtained are compared during the test period (2017–2018). Here, accuracy is a preferred indicator for the classification task. All data set displays presentable classification accuracy, with 76% for IMERG-E, 75.2% for IMERG-F, 82.94% for RFC-base, 83.58% for RFC-static, 80.21% for RFC-analog, and 85.22% for RFC-dynamic, respectively. The designed RFC experiments with four different configurations show better classification accuracy relative to the raw IMERG-E product. Among them, RFC-dynamic is the ranked one, followed by RFC-static, RFC-base, and finally RFC-analog. Unexpectedly, IMERG-F exhibits worse performance compared to IMERG-E. To help diagnose the performance of different data set s, we further counted the number of occurrences of different precipitation events and calculated the POD and FAR indicators. Their spatial distribution as well as boxplots are displayed in Figure 5. The displayed results of the RF model here is RFC-static, which is also the selected model configuration we use for subsequent analysis. Other RFC models are not shown, and their differences are fully compared in the discussion part. In Figures 5a–5c and 5g, one can see that the regional averaged POD of IMERG-F (0.592) is slightly better than that of IMERG-E (0.528), while the regional averaged POD of IMERG-E is substantially improved after post-processing with RFC-static model (0.76). In terms of regional averaged FAR (Figures 5d–5g), IMERG-F (0.272) is degraded compared to raw IMERG-E (0.253); in contrast, the use of RFC-static postprocessor improves the

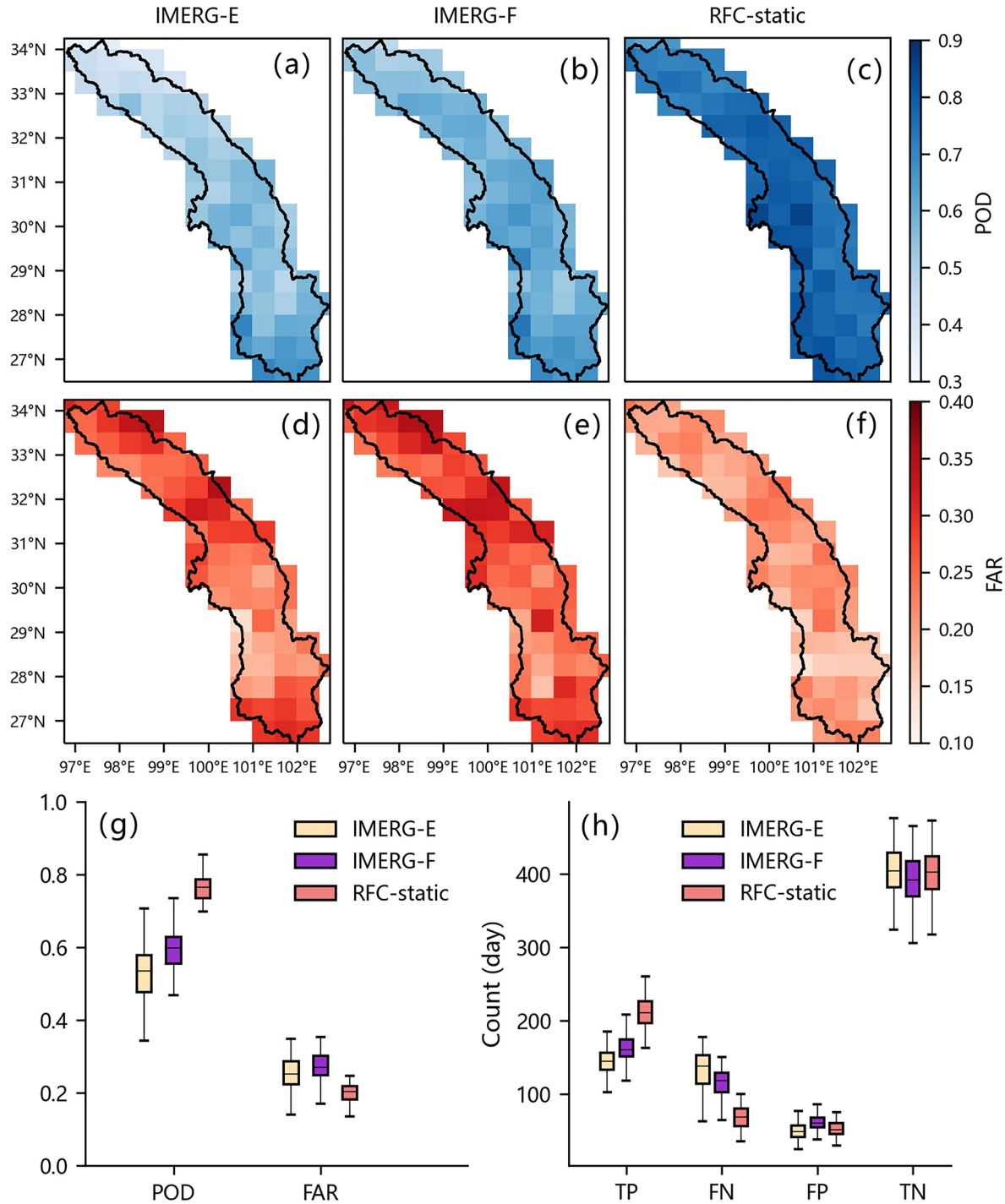

Figure 5. Wet-dry events classification performance of China Meteorological Administration (CMA) observation and the IMERG-E, IMERG-F, and RFC-static experiment during the test period (2017–2018). (a–c) Probability of detection (POD, higher is better), (g) their boxplot of all grid cells of the basin, (d–f) false alarm ratio (FAR, smaller is better), (h) binary events (TP: hit events, FN: missed events, FP: false alarm events, TN: correct negative events). RFC-static experiment in the figure represents the random forest classification with static predictors. The results in (g, h) are calculated by flattening data of all grid cells into one-dimensional vectors.

overall FAR value (0.2). The decomposition of the four events (Figure 5h) can further explain these results. The significant improvement in the discrimination of hit events has led to a remarkable increase in the POD of the RFC-static experiment. And, compared to minor increase in false alarm events, the significant improvement in the discrimination of hit events also results in a smaller FAR. It can also be seen that the improvement in hit

events is mainly due to the transformation of missed events, which means that the RFC model greatly reduces the probability of a missed precipitation event. A downside is that missed events may also be converted to false alarms events. The information on the occurrence of precipitation is increased by introducing auxiliary variables, thus reducing the missed events and increasing the hit events. The spatial plots also confirm that the RF-based ML algorithm can better serve as a simple regional model, which has been recognized in previous studies (X. He et al., 2016; Y. Zhang & Ye, 2021).

4.2. Probabilistic Post-Processing Assessment

Figures 6a and 6c compare the CPRS and twCRPS for the four QRF experiments with different configurations across the basin, respectively. Similar to the classification task, QRF-dynamic remains the unbeatable one (1.147 and 6.128), followed by QRF-static (1.248 and 6.638), QRF-base (1.269 and 6.689), and finally QRF-analog (1.357 and 8.271). Figures 6b and 6d show the spatial distribution of CPRS and twCPRS of QRF-static, respectively. The accuracy of ensemble outputs is positively correlated with the precipitation scale, where the ensemble forecasts are relatively poor for sites with large daily precipitation rates and relatively better for sites with slight daily precipitation rates.

We further compared the reliability of ensemble outputs obtained from QRF experiments with different configurations using rank histograms. The ensemble members in Figure 7 are the 20 ensemble members extracted by the equal sampling of the original ensemble quantiles. Except for QRF-analog, all the other three experiments (Figures 7a, 7b and 7d) yield almost flat rank histograms. Among them, QRF-base and QRF-static are equally best (RMSD = 0.0055), followed by QRF-dynamic (RMSD = 0.0068). But they are not significantly different, and all of them obtained relatively reliable ensemble members. The less-than-flat histogram distribution may be due to sampling errors. A relatively pronounced "/" pattern appears in QRF-analog experiment, indicating underestimation of the ensemble post-processing outputs. One possible reason is that Euclidean distances in high-dimensional feature space may lead to sparse solutions, eventually leading to biased (underestimated) analogs samples.

We randomly selected three grid cells from the upper to lower reaches of the Yalong River basin (No. 16, No. 33, No. 58), respectively. Figure 8 depicts their time series during the rainy season of the test period (June–October 2017 and 2018). The other seasons were not selected because almost no precipitation events occurred. The observations are indicated by black dots and the two PIs are the 50% and 90% intervals, respectively. The 50% PI inherently encloses precipitation events of 5–10 mm, and the 90% PI mostly encloses a broader range of precipitation events of 1–30 mm, for any grid in the upper, middle, or lower reaches of the basin. Some of the extreme precipitations are not captured because relatively few extreme precipitation samples in the training period are sought at the leaf nodes of the quantile forests. In addition, the "zero" precipitation events are treated as "truncated" values in the ensemble post-processing scheme and in the first step they are removed by the RFC algorithm. We train the QRF model on wet days. Therefore, for the treatment of "zero" precipitation events, we assign a value of "zero" to all ensemble members and no ensemble intervals are obtained. This approach also resulted in some minor rain events being misclassified as dry days in the first step, and no PIs were obtained about these misjudged events, resulting in underestimation and under-dispersion.

Reliability diagrams of ensemble outputs from different experiments during the test period (2017–2018) using 80th, 90th, and 95th percentiles of observations as the thresholds are shown in Figure 9. Based on the three selected thresholds, the QRF-analog experiment is still the worst performer, while the remaining three experiments do not differ significantly. The QRF-analog experiment exhibits relative low probabilities at all three threshold conditions. At lower threshold (80%), the QRF-base, QRF-static, and QRF-dynamic almost all fitted the 1:1 line. As the threshold increases, they all produce underestimates, indicating limitations for extreme precipitation events.

4.3. Deterministic Post-Processing Assessment

Using the ground "truth" as a reference, Figure 10 shows the spatial plots of the three commonly used deterministic metrics for IMERG-E, IMERG-F and QRF-static ensemble mean. For PCC, both IMERG-E and IMERG-F exhibit relatively low correlations (0.31 and 0.34), and IMERG-F barely improves the correlation coefficients of the raw IMERG-E estimates, while the QRF-static ensemble mean significantly improves the correlation between near-real-time satellite estimates and ground observations (0.7). For ME, the raw

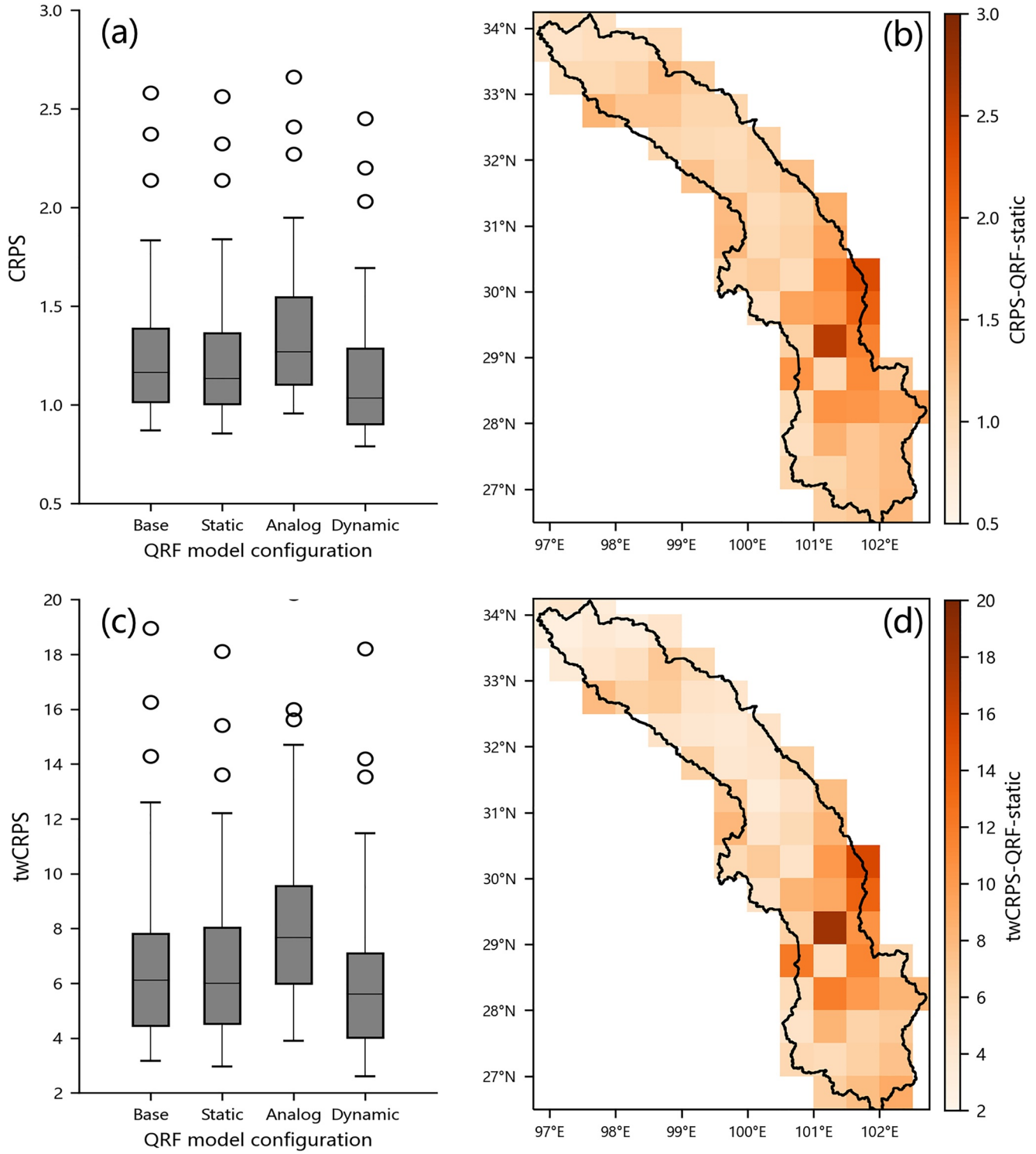

Figure 6. Continuous rank probability score (CRPS, smaller is better) of Quantile regression forests (QRF)-based ensemble outputs during the test period (2017–2018). (a, c) CRPS and twCRPS for all grid cells with different QRF experiments. (b, d) The spatial distribution of CRPS and twCRPS using the QRF-static experiment. Circles "o" in (a, c) represent outliers. The threshold for calculating twCRPS is the 95% quantile of observed precipitation values in each case. The results in (a, c) are calculated by flattening data of all grid cells into one-dimensional vectors.

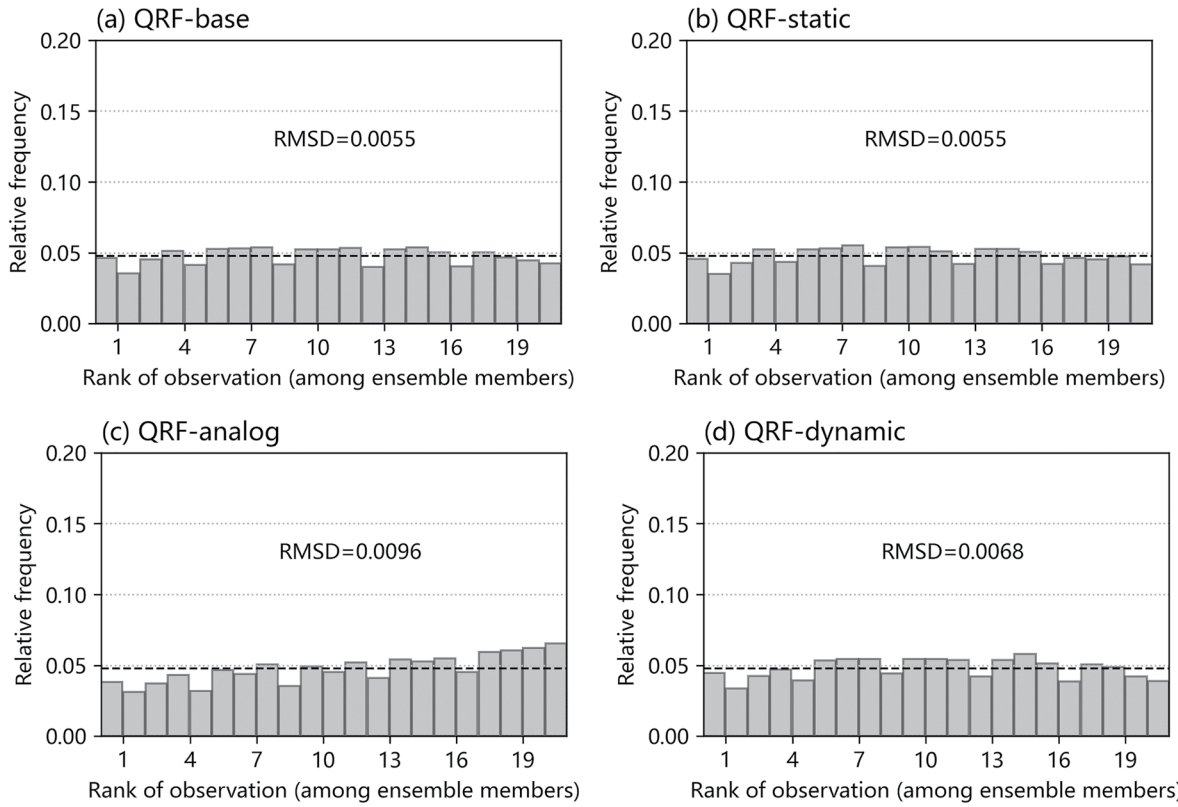

Figure 7. Rank histograms of the ensemble outputs from (a) QRF-base, (b) QRF-static, (c) QRF-analog, and (d) QRF-dynamic experiments during the test period (2017–2018). The results in this figure are calculated by flattening data of all grid cells into one-dimensional vectors.

IMERG-E presents underestimation in most parts of the basin, except at the downstream area of the basin, with significant underestimation in the midstream region. Both IMERG-F and QRF-static ensemble mean significantly reduce the bias (overestimation or underestimation), keeping the total bias in a small interval. QRF-static ensemble mean reaches a smaller total bias compared to IMERG-F, with ME value to be concentrated around "zero". This is mainly due to the further improvement of the overestimation in the downstream region of the basin. For RMSE, the error range of IMERG-E is roughly 2–12 mm/day, and IMERG-F even shows degradation compared to IMERG-E, which is related to the bias correction procedure of IMERG-F and the calculation of RMSE, where sporadic extreme values may lead to a large total RMSE. The performance of the QRF-static ensemble mean continues to be impressive, greatly reducing the RMSE across the whole basin. The possible reasons for the gap between IMERG-F and QRF-static ensemble mean are inferred as follows. The post-processing method of IMERG-F is mainly quantile mapping based on observations, which mainly adjusts the quantiles and does not improve the correlation, thus obtaining a lower correlation coefficient; while improving the quantiles results in the total smaller bias (ME); but the quantile adjustment process unreasonably introduces few extreme values, making it possible to obtain a relatively larger RMSE. The significant improvement in QRF-static is attributed to the additional predictors introduced that boost the correlation coefficient and result in smaller ME and RMSE.

ME and RMSE metrics are two commonly used, but relatively "unfair", scoring metrics. The calculation of ME is the result of accumulating the positive and negative bias of the time series, which may add up to a smaller total bias for each larger daily bias with different signs (positive and negative bias); and may add up to a larger total bias for smaller daily bias with same signs (positive or negative bias). The RMSE is calculated by accumulating the sum of squares of the differences of the two time series, which may lead to some sporadic extremes affecting the overall score, since the sum of squares gives more weight to these extremes, resulting in a relatively larger RMSE. To obtain relatively "fair" assessments, we introduce a four-components error decomposition (4CED) method to separate the total bias in the precipitation estimates into four independent parts: HPB, HNB, false alarm bias (FB) and MB. The four error components are uncorrelated and symbolically distinct, allowing not

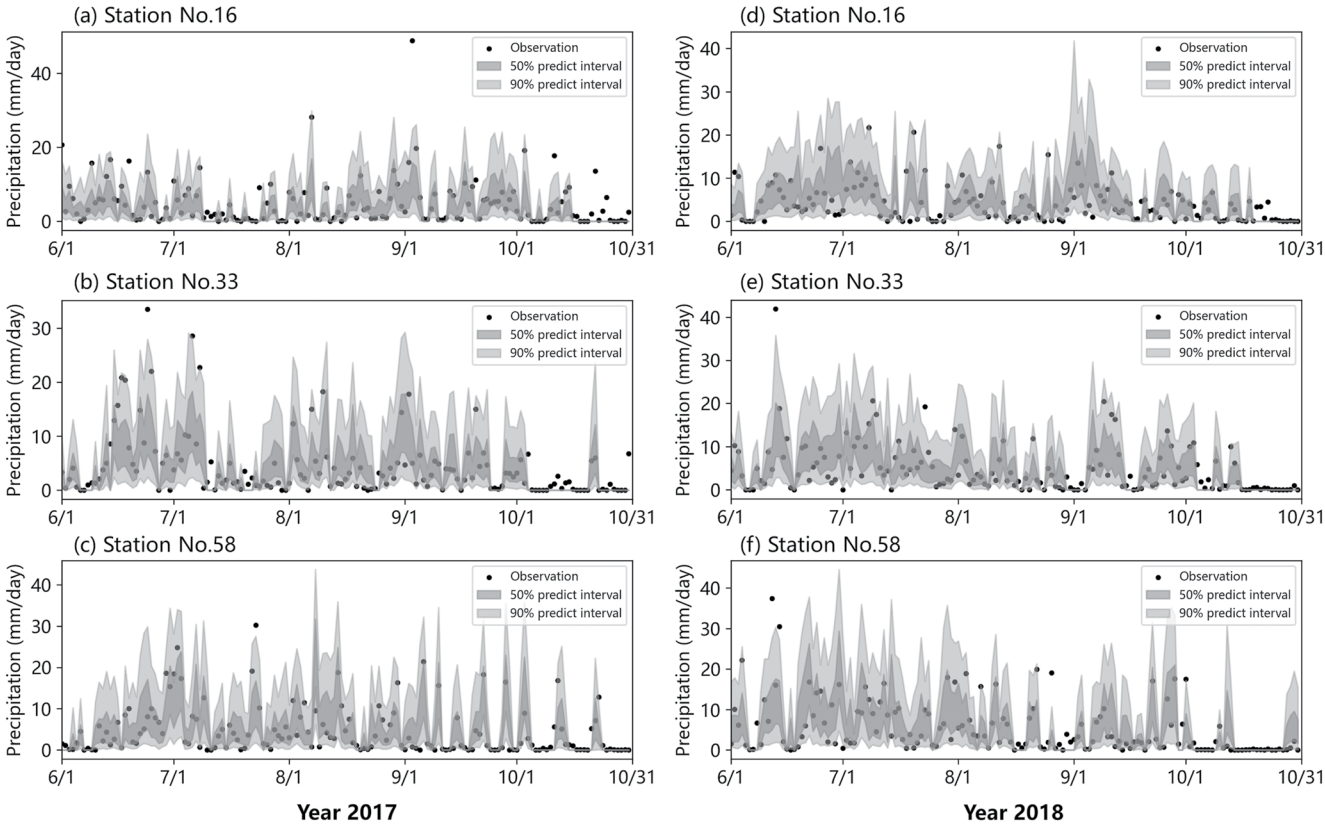

Figure 8. Time series of the CMA observations and predict intervals (shaded parts, 50% and 90%) for three different grid cells ((a, d): Station No.16; (b, e): Station No.33; (c, f): Station No.58) during the rainy season (June–October) in 2017 (a–c) and 2018 (d–f).

only to compare the magnitude of individual error components between different data sets, but also to analyze the dominant error factors. The use of 4CED can further answer a key question "Which error components are effectively post-processed by IMERG-F and QRF experiment?"

Using 4CED, the four individual error components of IMERG-E, IMERG-F, QRF-static ensemble mean during 2017–2018 are displayed in Figure 11. Based on the four error components, it is clearly observed that dominant factors and synergetic factors jointly shape the characteristics of the spatial plots of MEs. For IMERG-E, the

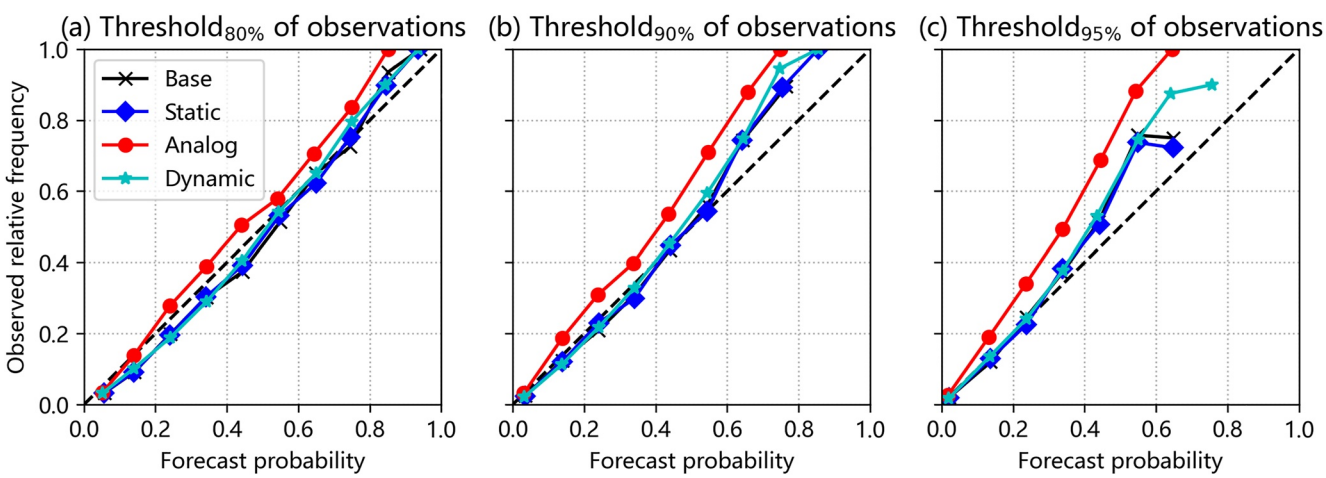

Figure 9. Reliability diagrams of ensemble outputs from different experiments during the test period (2017–2018) using 80th, 90th, and 95th percentiles of observations as the thresholds. The results in this figure are calculated by flattening data of all grid cells into one-dimensional vectors.

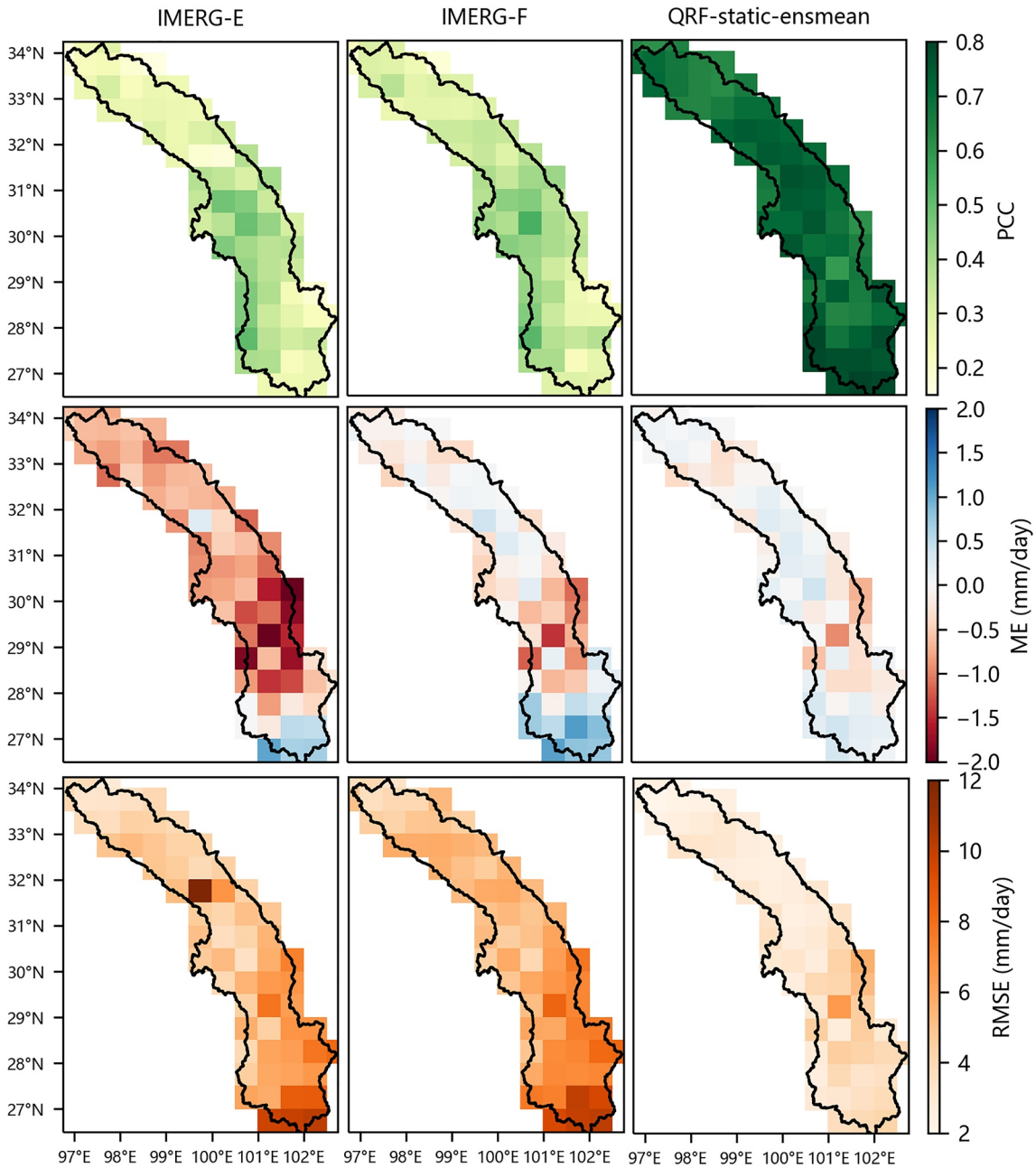

Figure 10. Pearson correlation coefficient (PCC, higher is better), mean error (ME, closer to 0 is better) and root mean square error (RMSE, smaller is better) of China Meteorological Administration (CMA) observations and the results of the IMERG-E, IMERG-F, and QRF-static-ensmean (ensemble mean) during the test period (2017–2018).

negative ME in the upstream region of the basin is mainly due to HNB, while the negative ME in the middle and lower reaches comes from the synergetic effect of HNB and MB. The positive ME near the outlet of the basin comes from the synergetic effect of HPB and FB. IMERG-F degrades the raw IMERG-E performance in terms of HPB and FB. There is an only slight improvement for HNB and MB, which finally presents a relatively balanced positive and negative bias. It further explains the main reason for the better performance of IMERG-F in ME and the worse performance of RMSE. Conversely, the precipitation obtained by the QRF model improves all aspects of the raw IMERG-E estimates, with the most remarkable improvement being the HNB, followed by the MB. For HPB and FB, QRF shows relatively small improvements, mainly in regions with higher precipitation climatology

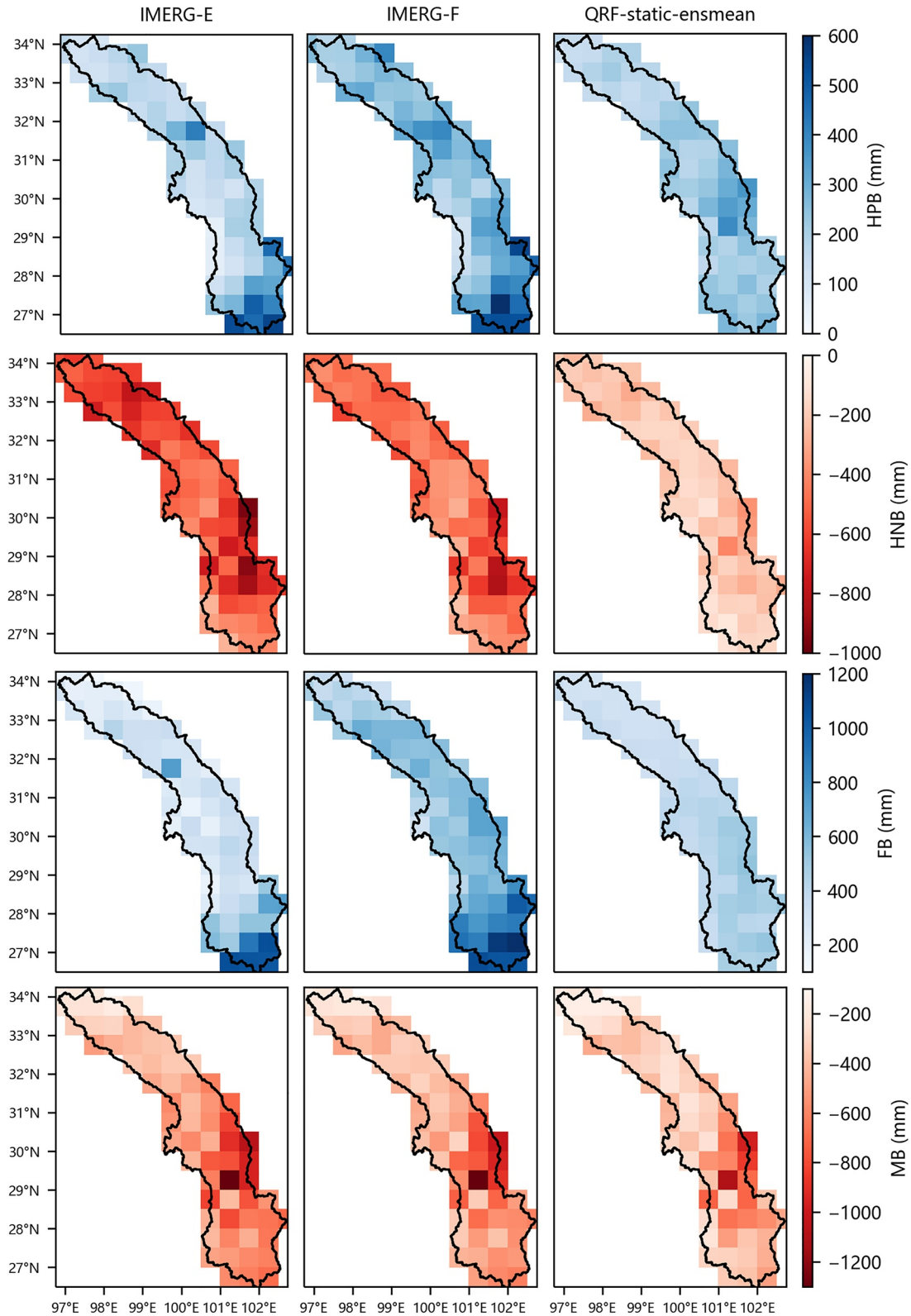

Figure 11. Hit positive bias (HPB), hit negative bias (HNB), false bias (FB), and missed bias (MB) of China Meteorological Administration (CMA) observations and the results of the IMERG-E, IMERG-F, and QRF-static-ensmean (ensemble mean) during the test period (2017–2018). The smaller HPB, HNB, FB, and MB are better.

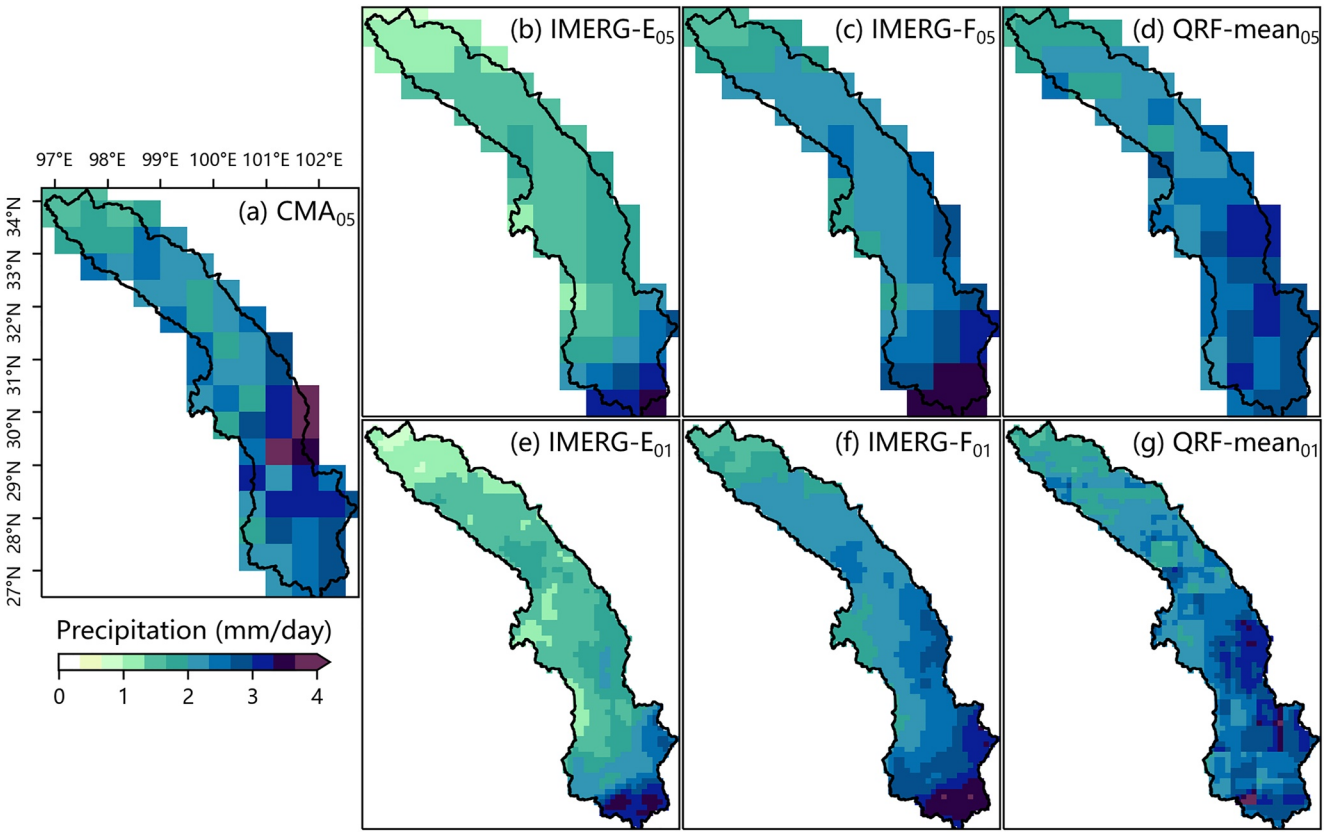

Figure 12. The precipitation spatial distribution of (a) China Meteorological Administration (CMA) with 0.5°, IMERG-E with (b) 0.5° and (e) 0.1°, IMERG-F with (c) 0.5° and (f) 0.1°, and QRF-mean (static ensemble mean) with (d) 0.5° and (g) 0.1° during the test period (2017–2018).

(e.g., downstream of the basin). The overall improvement of four individual error components is also responsible for the smaller ME and RMSE of QRF estimates.

4.4. Downscaling to Original Resolution

Due to the spatial resolution (0.5°) of the reference data, we implement the probabilistic post-processing at a coarse resolution. To reconstruct the original fine resolution (0.1°) of the IMERG product, we use a simple linear scaling factor method. Depending on the ratio between the coarse IMERG-E and the QRF outputs, we map these scaling factors to the original IMERG-E product to obtain the post-processed fine resolution IMERG-E product. Once a finer reference is available, we can build the fine-resolution post-processing model directly and skip this step.

Figure 12 shows the precipitation spatial distribution of CMA observation, IMERG-E, IMERG-F, and QRF-static-mean (ensemble mean) at 0.5° and 0.1°. As the previous results (ME in Figure 10) demonstrate, we can see that IMERG-E (Figure 12b) and IMERG-F (Figure 12c) are underestimated in the upstream region and overestimated near the basin outlet. After QRF post-processing, these biases are mitigated to some extent. Both the original fine IMERG-E and IMERG-F are relatively smooth. However, the reconstructed fine-resolution precipitation distribution is not smooth enough. One possible reason for this is that the scaling factor varies between grids.

5. Discussions

5.1. Feature Importance

Among the ML models, one advantage of the RF model is that it can be used to measure the relative importance of input variables, thus increasing the knowledge of the interpretability of the "black-box" model. Here, we show

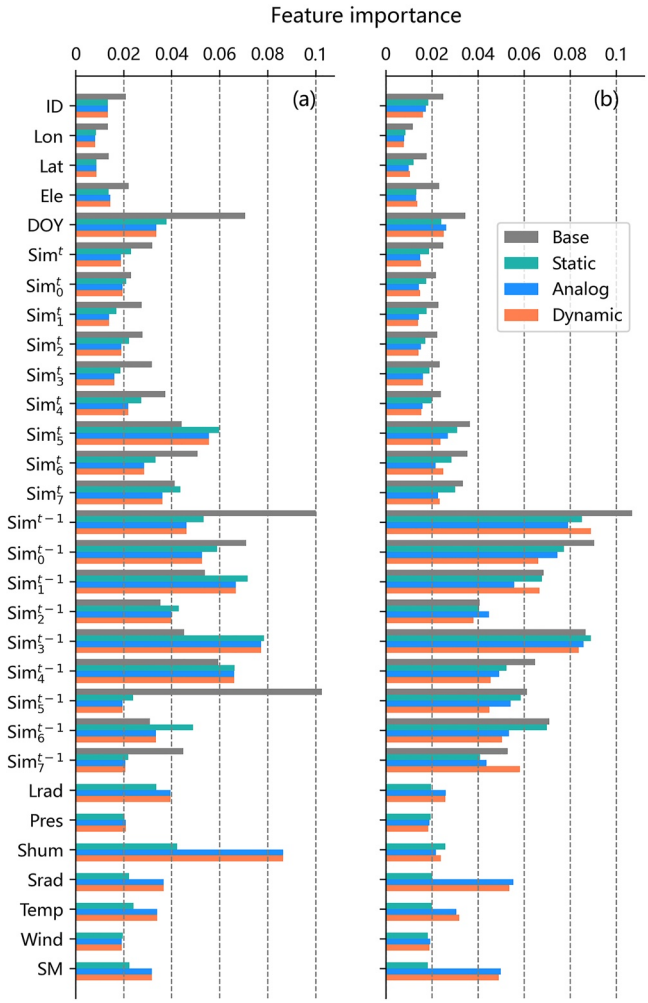

Figure 13. Feature importance of random forest (a) classification and (b) regression with different model configurations.

the relative importance of the 30 predictors (23 predictors for the base model) chosen for the different model configurations (Figure 13). In general, the importance of the predictors differs between the classification and regression task and, in addition, across model configurations. For the classification task, the dependence on the 3 × 3 grid precipitation estimates on the previous is high in all models. It is worth noting that DOY plays a relatively important role in the RFC-base experiment. The additional introduced meteorological predictors also play compelling role in RFC-static, RFC-analog, and RFC-dynamic experiments. Except for near-surface air pressure (Pres) and near-surface wind speed (Wind), other predictors also reflect different gaps in static and dynamic experiments. Of all the meteorological auxiliary variables, near-surface air specific humidity (Shum) plays the most important role in the classification task. This indicates that humidity is a discerning signal to discriminate dry-wet events classification. For the regression task, the relative importance of the different predictors is roughly similar, but there are differences. The model performance is more dependent on the 3 × 3 grid precipitation estimates on the previous day. The DOY becomes less important, while among the meteorological auxiliary predictors, near-surface air specific humidity (Shum), which previously was indicated as important in the classification task, becomes almost insensitive to the model configuration and less important. The importance of surface downward shortwave radiation (Srad) and SM increased subsequently. Among the predictors describing geographic attributes, grid number (ID) and elevation (Ele) are more discriminative compared to latitude and longitude. The comparison of the precipitation estimates on the previous day seems to show the ability of the infrared images to reflect the precipitation signal earlier.

In the proposed experimental design, four different model configurations were used to explore and address the problem of the unavailability of dynamic meteorological auxiliary predictors in near-real-time. The practicality of these predictors was also indicated by the improvement in the QRF-static and QRF-dynamic models. But the auxiliary variables we used in the present study are still limited. And they are mostly extracted from reanalysis products, which are not accessible in real-time. Therefore, we propose two solutions: (a) static experiments use station-based, multiyear averaged climatology to show the specificity of each grid cell; (b) analog experiments create surrogate "fake" near-real-time accessible predictors by searching for historical samples. Theoretically, the analogs experiment is the more "dynamic" one. However, the Static experiments perform better than the Analog experiment in all situations. Because although static experiments use fixed climatology as predictors, the samples are relatively evenly distributed (i.e., homogeneous). However, the homogeneity of Analog experiments was degrading heavily (Evin et al., 2021). Dynamic experiments exhibited the best performance and played an unbeatable role, which confirms the importance of near-real-time predictors. Therefore, obtaining more targeted and near-real-time accessible predictors that match the release time of near-real-time satellite retrieval are both prerequisites to ensure the optimal performance of probabilistic post-processing models. In addition, there are other predictors (e.g., time lag NDVI, etc.) that require more discussion and selection to maintain a trade-off balance between model efficiency and accuracy (Mei et al., 2020).

5.2. Limitations and Future Work

Although the above results demonstrate the great potential of the proposed method, there are still some issues. First, although the proposed method can significantly improve the accuracy of IMERG-E precipitation estimates, the error decomposition reveals that these improvements are mainly focused on the reduction of HNB and MB. But for HPB and FB, the present method expresses limited ability. On the one hand, this is because the raw IMERG-E performs relatively well in these two aspects (HNB and MB); while for HPB and FB, especially for the latter, the final observations in leaf node obtained based on RF node split are not accurate enough. Meaning

that, it is difficult to correct a false wet event as a dry event. Another possible reason is that the increase of the hit events is the result of the conversion of the original miss events, which leads to partial HPB. Nevertheless, in regions with relatively large precipitation climatology (near to outlet), the proposed model still expresses effective adjustments.

Another issue is the handling of zero precipitation events by the probabilistic post-processing model. Within the proposed framework, the post-processing process is divided into two steps: classification and regression. In the first step, we use RF classification to separate samples into wet and dry events. And in the second step, we generate ensemble outputs by training the QRFs model only for the samples of wet days. The dry events with "zero" precipitation were handled as truncated values. Then, in the generated ensemble estimates, the adjusted samples of dry events do not have PIs. This approach is acceptable for those correctly classified samples. However, for misclassified samples, the truncated events lead to inevitable errors. Therefore, reliable ensemble members cannot be obtained either. The solution that has been tried is to use QRF directly, without classification. However, a tricky challenge prevented us from trying further. The main reason was that the large number of zero (or near zero) precipitation events made it difficult for the algorithm to converge. We even spent dozens of hours with smaller hyperparameters (e.g., fewer trees) and did not get ensemble outputs. In traditional statistical post-processing models, the truncated values are replaced by proxy samples, which are generated by randomly sampling the samples under the truncated threshold (Scheuerer & Hamill, 2015). The proxy samples are then brought into the explicit conditional probability distribution function to generate the ensemble outputs. The reason for not using this method in the present study is that although the proxy samples can be obtained by random sampling, the other predictors of the sample cannot be obtained using the same method. The feature space of the new prediction sample is incomplete, so no inference can be made either. Another reason is that QRF is built based on a finite number of quantiles instead of an explicit conditional probability distribution function. Conditional probability inference under the truncated threshold is inaccurate. One possible solution is to use parametric methods (e.g., logistic regression) to model the probability of precipitation separately, replacing the original RFC, then generating ensemble forecasts for truncated events.

A common issue with models based on historical analogy search, including RF, is the strong dependence on historical observations (A. H. Li & Martin, 2017). For example, the analog method, although effective, requires a very large number of historical samples to guarantee perfect prediction (A. H. Li et al., 2017). The same is true for the RF model. By dissecting the inference process of these models, one can be found is that the model architecture is built on a set of analogs historical observations. Then the inference is calculated by averaging these analogs observations. The different predictors are mainly used as the basis for the growth and split of the RF model and are not involved in the final prediction inference. Thus, for a specific quantile, when not enough analogs historical observations are searched, the prediction will be further biased after weight averaging only a few biased analogs samples. For example, biased analogs samples also result in underestimated targets (e.g., Figure 6c) in our QRF-analog experiments. In addition, as we mentioned above, QRF is built based on a finite number of quantiles instead of an explicit conditional probability distribution function. And the prediction of heavy precipitation events is also based on the inference or extrapolation of historical samples. As a result, this may lead to two different degrees of underestimated bias. The first scenario is heavy precipitation events where the analogs samples are historically available but the sample size is not enough. In this case, the averaging of analogs samples leads to a slightly negative bias (Figure 8). The second scenario is extreme precipitation events where there are no analogs samples in the training set and only completely underestimated samples can be found. In this case, the averaging of these samples leads to a severe negative bias (Figure 8). Such models cannot anticipate the case of future extreme events that have not occurred in the past. It further hinders its application in future prediction in the context of climate change. One possible solution is to combine parametric methods (e.g., extreme-value distribution) with non-parametric methods (e.g., QRF). The parametric distribution allows a better fit for the case of extreme events. The challenge here, however, is to choose the appropriate distribution function. A recent study has attempted to combine quantile RFs and extended generalized Pareto distributions (Taillardat et al., 2019). Their research was inspired by Naveau et al. (2016). However, this hybrid approach requires that different distribution functions are first selected and then serval parameters need to be calibrated to fit extreme precipitation events, which greatly increases the complexity of the modeling. The choice between model complexity and model efficiency is a trade-off. How to more succinctly combine the estimation of explicit CDFs with ML algorithms is also a future research direction (Gnecco et al., 2022). Another possible solution is to use neural networks to learn the parameters of the distribution function directly (W. Li et al., 2022; Rasp & Lerch, 2018; Schulz &

Lerch, 2021). In addition, other parametric methods (e.g., Bayesian approach based on copula functions; Khajehei et al., 2018; Khajehei & Moradkhani, 2017) can be used as alternative techniques. Our model in this study will be able to act as a benchmark to compare and enhance possible effective models in future studies.

Finally, it is important to note that the ensemble outputs we obtain are only probabilistic estimates, not spatial ensembles, and cannot be used directly for downstream applications such as hydrological ensemble simulations. A necessary step is to perform spatiotemporal reconstructions, that is, a reordering of the ensemble members for different stations. Methods that can be used include Schaake shuffle (Clark et al., 2004), variants of the Schaake shuffle (Schefzik, 2016; Scheuerer et al., 2017) and ensemble copula coupling (Schefzik et al., 2013).

In this study, we selected the Yalong River basin in China as our region of interest, and we built a regional model for all target grid cells. It is an advantage of ML modeling, being able to extract local features and use knowledge from neighboring regions to improve overall model performance. This is very promising for areas with poor gauge networks. In future studies, we can test training the model at specific grid cells and validate it at other grid cells to see how the model performs (Ma et al., 2021). Although the Yalong River basin covers a wide range of topographic conditions and climate types, its basin area is still relatively small compared to the continental scale or global scale. How to extend the regional model to the continental model or even to the global model and maintain relatively high computational efficiency is also a great challenge. It will be of great significance and value to conduct the global near-real-time satellite precipitation estimates post-processing, and further improve the global near-real-time flood monitoring and other operational applications.

6. Conclusions

In this study, a QRF-based probabilistic post-processor, QRF4P-NRT, is proposed to calibrate the near-real-time satellite precipitation estimates. Unlike commonly used deterministic correction models, QRF4P-NRT can perform both deterministic mean correction and probabilistic calibration. We designed several experiments to apply the proposed method to the operational post-processing routine of IMERG-E and compared their outputs with the raw IMERG-E and the officially released bias-corrected IMERG-F product. The ensemble mean based on QRF4P-NRT is significantly improved relative to the raw IMERG-E and exceeds the high quality of IMERG-F. The ensemble outputs of QRF models with different dynamic predictor solutions also provide reliable probabilistic information. These promising results are evidence of the potential of the QRF4P-NRT for operational applications. However, the selection of more predictors and the near-real-time accessibility of these predictors are still challenging to further enhance the skill of probabilistic post-processing for near-real-time satellite precipitation estimates.

Data Availability Statement

The GPM IMERG (https://gpm.nasa.gov/data/directory) Early Run (IMERG-E) L3 1 day 0.1° × 0.1° V06 data set and Final Run (IMERG-F) L3 1 day 0.1° × 0.1° V06 data set are publicly available at (https://disc.gsfc.nasa.gov/datasets/GPM_3IMERGDE_06/summary?keywords="IMERG_Early" and https://disc.gsfc.nasa.gov/datasets/GPM_3IMERGDF_06/summary?keywords="IMERG_final%22), respectively. The CMA reference data is available at (http://data.cma.cn/en) provided by the National Meteorological Information Center of China Meteorological Administration. The China meteorological forcing data set is publicly available at the National Tibetan Plateau Data Center of China via (http://data.tpdc.ac.cn/en/data/8028b944-daaa-4511-8769-965612652c49/). The ERA5-land soil moisture data is publicly available at (https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-land?tab=overview; Muñoz-Sabater, 2019) provided by the ECMWF. The Shuttle Radar Topographic Mission data is publicly available at (https://www2.jpl.nasa.gov/srtm/). The QRF4P-NRT codes are accessible from Zenodo repository (Jnelson18, 2022; Y. Zhang et al., 2022).

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

References

Aleshina, M. A., Semenov, V. A., & Chernokulsky, A. V. (2021). A link between surface air temperature and extreme precipitation over Russia from station and reanalysis data. *Environmental Research Letters*, *16*(10), 105004. https://doi.org/10.1088/1748-9326/ac1cba

Arulraj, M., & Barros, A. P. (2021). Automatic detection and classification of low-level orographic precipitation processes from space-borne radars using machine learning. *Remote Sensing of Environment*, *257*, 112355. https://doi.org/10.1016/j.rse.2021.112355

Back, L. E., & Bretherton, C. S. (2005). The relationship between wind speed and precipitation in the Pacific ITCZ. *Journal of Climate*, *18*(20), 4317–4328. https://doi.org/10.1175/JCLI3519.1

Baez-Villanueva, O. M., Zambrano-Bigiarini, M., Beck, H. E., McNamara, I., Ribbe, L., Nauditt, A., et al. (2020). RF-MEP: A novel random forest method for merging gridded precipitation products and ground-based measurements. *Remote Sensing of Environment*, *239*, 111606. https://doi.org/10.1016/j.rse.2019.111606

Basist, A., Bell, G. D., & Meentemeyer, V. (1994). Statistical relationships between topography and precipitation patterns. *Journal of Climate*, *7*(9), 1305–1315. https://doi.org/10.1175/1520-0442(1994)007<1305:SRBTAP>2.0.CO;2

Beck, H. E., Wood, E. F., McVicar, T. R., Zambrano-Bigiarini, M., Alvarez-Garreton, C., Baez-Villanueva, O. M., et al. (2020). Bias correction of global high-resolution precipitation climatologies using streamflow observations from 9,372 catchments. *Journal of Climate*, *33*(4), 1299–1315. https://doi.org/10.1175/JCLI-D-19-0332.1

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

Bröcker, J. (2012). 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

Bröcker, J., & Ben Bouallègue, Z. (2020). Stratified rank histograms for ensemble forecast verification under serial dependence. *Quarterly Journal of the Royal Meteorological Society*, *146*(729), 1976–1990. https://doi.org/10.1002/qj.3778

Chen, H., Yong, B., Gourley, J. J., Wen, D., Qi, W., & Yang, K. (2021). A novel real-time error adjustment method with considering four factors for correcting hourly multi-satellite precipitation estimates. *IEEE Transactions on Geoscience and Remote Sensing*, *60*, 1–11. https://doi.org/10.1109/TGRS.2021.3131238

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

Clark, M., Gangopadhyay, S., Hay, L., Rajagopalan, B., & Wilby, R. (2004). The Schaake shuffle: A method for reconstructing space-time variability in forecasted precipitation and temperature fields. *Journal of Hydrometeorology*, *5*(1), 243–262. https://doi.org/10.1175/1525-7541(2004)005<0243:TSSAMF>2.0.CO;2

Cunha, L. K., Mandapaka, P. V., Krajewski, W. F., Mantilla, R., & Bradley, A. A. (2012). 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

Déry, S. J., & Wood, E. F. (2005). Observed 20th-century land surface air temperature and precipitation covariability. *Geophysical Research Letters*, *32*(21), L21414. https://doi.org/10.1029/2005GL024234

Donat, M. G., Sillmann, J., Wild, S., Alexander, L. V., Lippmann, T., & Zwiers, F. W. (2014). Consistency of temperature and precipitation extremes across various global gridded in situ and reanalysis data sets. *Journal of Climate*, *27*(13), 5019–5035. https://doi.org/10.1175/JCLI-D-13-00405.1

Dong, J., Crow, W. T., & Reichle, R. (2020). 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

Evin, G., Lafaysse, M., Taillardat, M., & Zamo, M. (2021). 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

Gnecco, N., Terefe, E. M., & Engelke, S. (2022). *Extremal Random Forests*. arXiv preprint arXiv:2201.12865.

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

Gneiting, T., & Ranjan, R. (2011). Comparing density forecasts using threshold-and quantile-weighted scoring rules. *Journal of Business & Economic Statistics*, *29*(3), 411–422. https://doi.org/10.1198/jbes.2010.08110

Gou, J., Miao, C., Samaniego, L., Xiao, M., Wu, J., & Guo, X. (2021). CNRD v1. 0: A high-quality natural runoff data set 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

Goudenhoofdt, E., & Delobbe, L. (2009). Evaluation of radar-gauge merging methods for quantitative precipitation estimates. *Hydrology and Earth System Sciences*, *13*(2), 195–203. https://doi.org/10.5194/hess-13-195-2009

Gumindoga, W., Rientjes, T. H. M., Haile, A. T., Makurira, H., & Reggiani, P. (2019). Performance of bias-correction schemes for CMORPH rainfall estimates in the Zambezi River basin. *Hydrology and Earth System Sciences*, *23*(7), 2915–2938. https://doi.org/10.5194/hess-23-2915-2019

Guo, B., Zhang, J., Meng, X., Xu, T., & Song, Y. (2020). Long-term spatio-temporal precipitation variations in China with precipitation surface interpolated by ANUSPLIN. *Scientific Reports*, *10*(1), 1–17. https://doi.org/10.1038/s41598-019-57078-3

Guo, H., Bao, A., Liu, T., Chen, S., & Ndayisaba, F. (2016). Evaluation of PERSIANN-CDR for meteorological drought monitoring over China. *Remote Sensing*, *8*(5), 379. https://doi.org/10.3390/rs8050379

Hamill, T. M. (2001). Interpretation of rank histograms for verifying ensemble forecasts. *Monthly Weather Review*, *129*(3), 550–560. https://doi.org/10.1175/1520-0493(2001)129<0550:IORHFV>2.0.CO;2

Hartmann, H. C., Pagano, T. C., Sorooshian, S., & Bales, R. (2002). 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

He, J., Yang, K., Tang, W., Lu, H., Qin, J., Chen, Y., & Li, X. (2020). The first high-resolution meteorological forcing data set for land process studies over China. *Scientific Data*, *7*(1), 1–11. https://doi.org/10.1038/s41597-020-0369-y

He, X., Chaney, N. W., Schleiss, M., & Sheffield, J. (2016). Spatial downscaling of precipitation using adaptable random forests. *Water Resources Research*, *52*(10), 8217–8237. https://doi.org/10.1002/2016WR019034

Hemri, S., & Klein, B. (2017). Analog-based postprocessing of navigation-related hydrological ensemble forecasts. *Water Resources Research*, *53*(11), 9059–9077. https://doi.org/10.1002/2017WR020684

Henn, B., Newman, A. J., Livneh, B., Daly, C., & Lundquist, J. D. (2018). An assessment of differences in gridded precipitation data sets in complex terrain. *Journal of Hydrology*, *556*, 1205–1219. https://doi.org/10.1016/j.jhydrol.2017.03.008

Herman, G. R., & Schumacher, R. S. (2018b). "Dendrology" in numerical weather prediction: What random forests and logistic regression tell us about forecasting extreme precipitation. *Monthly Weather Review*, *146*(6), 1785–1812. https://doi.org/10.1175/MWR-D-17-0307.1

Herman, G. R., & Schumacher, R. S. (2018a). Money does not grow on trees, but forecasts do: Forecasting extreme precipitation with random forests. *Monthly Weather Review*, *146*(5), 1571–1600. https://doi.org/10.1175/MWR-D-17-0250.1

Hong, Y., Adler, R. F., Negri, A., & Huffman, G. J. (2007). Flood and landslide applications of near real-time satellite rainfall products. *Natural Hazards*, *43*(2), 285–294. https://doi.org/10.1007/s11069-006-9106-x

Hou, A. Y., Kakar, R. K., Neeck, S., Azarbarzin, A. A., Kummerow, C. D., Kojima, M., et al. (2013). 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

Huang, Q., Qin, G., Zhang, Y., Tang, Q., Liu, C., Xia, J., et al. (2020). Using remote sensing data-based hydrological model calibrations for predicting runoff in ungauged or poorly gauged catchments. *Water Resources Research*, *56*(8), e2020WR028205. https://doi.org/10.1029/2020WR028205

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

Huffman, G. J., Bolvin, D. T., Nelkin, E. J., Wolff, D. B., Adler, R. F., Gu, G., et al. (2007). The TRMM Multisatellite Precipitation Analysis (TMPA): Quasi-global, multiyear, combined-sensor precipitation estimates at fine scales. *Journal of Hydrometeorology*, *8*(1), 38–55. https://doi.org/10.1175/JHM560.1

Huffman, G. J., Stocker, E. F., Bolvin, D. T., Nelkin, E. J., & Tan, J. (2019a). GPM IMERG early precipitation L3 1 day 0.1° × 0.1° V06. In A. Savtchenko, & M. D. Greenbelt (Eds.), *Goddard Earth Sciences Data and Information Services Center (GES DISC)*, Accessed: [30 July 2021]. https://doi.org/10.5067/GPM/IMERGDE/DAY/06

Huffman, G. J., Stocker, E. F., Bolvin, D. T., Nelkin, E. J., & Tan, J. (2019b). GPM IMERG final precipitation L3 1 day 0.1° × 0.1° V06. In A. Savtchenko, & M. D. Greenbelt (Eds.), *Goddard Earth Sciences Data and Information Services Center (GES DISC)*, Accessed: [30 July 2021]. https://doi.org/10.5067/GPM/IMERGDF/DAY/06

Ibarra-Berastegi, G., Saénz, J., Ezcurra, A., Elías, A., Diaz Argandoña, J., & Errasti, I. (2011). Downscaling of surface moisture flux and precipitation in the Ebro Valley (Spain) using analogues and analogues followed by random forests and multiple linear regression. *Hydrology and Earth System Sciences*, *15*(6), 1895–1907. https://doi.org/10.5194/hess-15-1895-2011

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

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

Khajehei, S., Ahmadalipour, A., & Moradkhani, H. (2018). An effective post-processing of the North American Multi-Model Ensemble (NMME) precipitation forecasts over the continental U.S. *Climate Dynamics*, *51*(1–2), 457–472. https://doi.org/10.1007/s00382-017-3934-0

Khajehei, S., & Moradkhani, H. (2017). Towards an improved ensemble precipitation forecast: A probabilistic post-processing approach. *Journal of Hydrology*, *546*, 476–489. https://doi.org/10.1016/j.jhydrol.2017.01.026

Khedhaouiria, D., Bélair, S., Fortin, V., Roy, G., & Lespinas, F. (2020). High-resolution (2.5 km) ensemble precipitation analysis across Canada. *Journal of Hydrometeorology*, *21*(9), 2023–2039. https://doi.org/10.1175/JHM-D-19-0282.1

Kubota, T., Aonashi, K., Ushio, T., Shige, S., Takayabu, Y. N., Kachi, M., et al. (2020). 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

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

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

Li, H., Ye, A., Zhang, Y., & Zhao, W. (2021). InterComparison and evaluation of multiSource soil moisture products in China. *Earth and Space Science*, *8*(10). https://doi.org/10.1029/2021EA001845

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

Li, W., Pan, B., Xia, J., & Duan, Q. (2022). Convolutional neural network-based statistical post-processing of ensemble precipitation forecasts. *Journal of Hydrology*, *605*, 127301. https://doi.org/10.1016/j.jhydrol.2021.127301

Li, W., Sun, W., He, X., Scaioni, M., Yao, D., Chen, Y., et al. (2019). Improving CHIRPS daily satellite-precipitation products using coarser ground observations. *IEEE Geoscience and Remote Sensing Letters*, *16*(11), 1678–1682. https://doi.org/10.1109/LGRS.2019.2907532

Lu, H., Zheng, D., Yang, K., & Yang, F. (2020). Last-decade progress in understanding and modeling the land surface processes on the Tibetan Plateau. *Hydrology and Earth System Sciences*, *24*(12), 5745–5758. https://doi.org/10.5194/hess-24-5745-2020

Ma, K., Feng, D., Lawson, K., Tsai, W. P., Liang, C., Huang, X., et al. (2021). Transferring hydrologic data across continents-leveraging data-rich regions to improve hydrologic prediction in data-sparse regions. *Water Resources Research*, *57*(5), e2020WR028600. https://doi.org/10.1029/2020WR028600

Maraun, D. (2013). Bias correction, quantile mapping, and downscaling: Revisiting the inflation issue. *Journal of Climate*, *26*(6), 2137–2143. https://doi.org/10.1175/JCLI-D-12-00821.1

Mei, Y., Maggioni, V., Houser, P., Xue, Y., & Rouf, T. (2020). A nonparametric statistical technique for spatial downscaling of precipitation over high mountain Asia. *Water Resources Research*, *56*(11), e2020WR027472. https://doi.org/10.1029/2020WR027472

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

Morice, C. P., Kennedy, J. J., Rayner, N. A., & Jones, P. D. (2012). Quantifying uncertainties in global and regional temperature change using an ensemble of observational estimates: The HadCRUT4 data set. *Journal of Geophysical Research: Atmospheres*, *117*(D8), D08101. https://doi.org/10.1029/2011JD017187

Muñoz-Sabater, J. (2019). ERA5-Land hourly data from 1981 to present. Copernicus Climate Change Service (C3S) Climate Data Store (CDS). (Accessed on 30 July 2021). https://doi.org/10.24381/cds.e2161bac

Muñoz-Sabater, J., Dutra, E., Agustí-Panareda, A., Albergel, C., Arduini, G., Balsamo, G., et al. (2021). ERA5-Land: A state-of-the-art global reanalysis data set for land applications. *Earth System Science Data*, *13*, 4349–4383. https://doi.org/10.5194/essd-13-4349-2021

Naveau, P., Huser, R., Ribereau, P., & Hannart, A. (2016). Modeling jointly low, moderate, and heavy rainfall intensities without a threshold selection. *Water Resources Research*, *52*(4), 2753–2769. https://doi.org/10.1002/2015wr018552

Nguyen, P., Ombadi, M., Gorooh, V. A., Shearer, E. J., Sadeghi, M., Sorooshian, S., et al. (2020). PERSIANN Dynamic Infrared-Rain Rate (PDIR-Now): A near-real-time, quasi-global satellite precipitation data set. *Journal of Hydrometeorology*, *21*(12), 2893–2906. https://doi.org/10.1175/JHM-D-20-0177.1

Nguyen, P., Ombadi, M., Sorooshian, S., Hsu, K., AghaKouchak, A., Braithwaite, D., et al. (2018). The PERSIANN family of global satellite precipitation data: A review and evaluation of products. *Hydrology and Earth System Sciences*, *22*(11), 5801–5816. https://doi.org/10.5194/hess-22-5801-2018

Nguyen, P., Shearer, E. J., Ombadi, M., Gorooh, V. A., Hsu, K., Sorooshian, S., et al. (2020). 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

Nguyen, P., Thorstensen, A., Sorooshian, S., Hsu, K., & AghaKouchak, A. (2015). Flood forecasting and inundation mapping using HiResFlood-UCI and near-real-time satellite precipitation data: The 2008 Iowa flood. *Journal of Hydrometeorology*, *16*(3), 1171–1183. https://doi.org/10.1175/JHM-D-14-0212.1

Pan, M., Li, H., & Wood, E. (2010). Assessing the skill of satellite-based precipitation estimates in hydrologic applications. *Water Resources Research*, *46*(9), W09535. https://doi.org/10.1029/2009WR008290

Parrish, M. A., Moradkhani, H., & DeChant, C. M. (2012). Toward reduction of model uncertainty: Integration of Bayesian model averaging and data assimilation. *Water Resources Research*, *48*(3), 3519. https://doi.org/10.1029/2011WR011116

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research*, *12*, 2825–2830. Retrieved from https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf?ref=https://githubhelp.com

Pendergrass, A. G., & Hartmann, D. L. (2014). The atmospheric energy constraint on global-mean precipitation change. *Journal of Climate*, *27*(2), 757–768. https://doi.org/10.1175/JCLI-D-13-00163.1

Qi, W., Yong, B., & Gourley, J. J. (2021). Monitoring the super typhoon Lekima by GPM-based near-real-time satellite precipitation estimates. *Journal of Hydrology*, *603*, 126968. https://doi.org/10.1016/j.jhydrol.2021.126968

Qiang, F., Zhang, M., Wang, S., Liu, Y., Ren, Z., & Zhu, X. (2016). Estimation of areal precipitation in the Qilian Mountains based on a gridded data set since 1961. *Journal of Geographical Sciences*, *26*(1), 59–69. https://doi.org/10.1007/s11442-016-1254-7

Rasp, S., & Lerch, S. (2018). Neural networks for postprocessing ensemble weather forecasts. *Monthly Weather Review*, *146*(11), 3885–3900. https://doi.org/10.1175/MWR-D-18-0187.1

Sahin, S. (2012). An aridity index defined by precipitation and specific humidity. *Journal of Hydrology*, *444*, 199–208. https://doi.org/10.1016/j.jhydrol.2012.04.019

Schefzik, R. (2016). A similarity-based implementation of the Schaake shuffle. *Monthly Weather Review*, *144*(5), 1909–1921. https://doi.org/10.1175/MWR-D-15-0227.1

Schefzik, R., Thorarinsdottir, T. L., & Gneiting, T. (2013). Uncertainty quantification in complex simulation models using ensemble copula coupling. *Statistical Science*, *28*(4), 616–640. https://doi.org/10.1214/13-STS443

Scheuerer, M., & Hamill, T. M. (2015). Statistical postprocessing of ensemble precipitation forecasts by fitting censored, shifted gamma distributions. *Monthly Weather Review*, *143*(11), 4578–4596. https://doi.org/10.1175/MWR-D-15-0061.1

Scheuerer, M., Hamill, T. M., Whitin, B., He, M., & Henkel, A. (2017). A method for preferential selection of dates in the Schaake shuffle approach to constructing spatiotemporal forecast fields of temperature and precipitation. *Water Resources Research*, *53*(4), 3029–3046. https://doi.org/10.1002/2016WR020133

Schreiner McGraw, A. P., & Ajami, H. (2020). Impact of uncertainty in precipitation forcing data sets on the hydrologic budget of an integrated hydrologic model in mountainous terrain. *Water Resources Research*, *56*(12), e2020WR027639. https://doi.org/10.1029/2020WR027639

Schulz, B., & Lerch, S. (2021). Machine learning methods for postprocessing ensemble forecasts of wind gusts: A systematic comparison. *Monthly Weather Review*, *150*(1), 235–257. https://doi.org/10.1175/MWR-D-21-0150.1

Sharifi, E., Saghafian, B., & Steinacker, R. (2019). Downscaling satellite precipitation estimates with multiple linear regression, artificial neural networks, and spline interpolation techniques. *Journal of Geophysical Research: Atmospheres*, *124*(2), 789–805. https://doi.org/10.1029/2018JD028795

Shen, Z., Yong, B., Gourley, J. J., & Qi, W. (2021). 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

Song, Y., & Ying, L. U. (2015). Decision tree methods: Applications for classification and prediction. *Shanghai Archives of Psychiatry*, *27*(2), 130. https://doi.org/10.11919/j.issn.1002-0829.215044

Sun, Q., Miao, C., Duan, Q., Ashouri, H., Sorooshian, S., & Hsu, K. L. (2018). 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

Taillardat, M., Fougères, A., Naveau, P., & Mestre, O. (2019). Forest-based and semiparametric methods for the postprocessing of rainfall ensemble forecasting. *Weather and Forecasting*, *34*(3), 617–634. https://doi.org/10.1175/waf-d-18-0149.1

Taillardat, M., Mestre, O., Zamo, M., & Naveau, P. (2016). 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

Tang, G., Behrangi, A., Long, D., Li, C., & Hong, Y. (2018). Accounting for spatiotemporal errors of gauges: A critical step to evaluate gridded precipitation products. *Journal of Hydrology*, *559*, 294–306. https://doi.org/10.1016/j.jhydrol.2018.02.057

Tang, G., Clark, M. P., Papalexiou, S. M., Newman, A. J., Wood, A. W., Brunet, D., & Whitfield, P. H. (2021). EMDNA: An ensemble meteorological data set for North America. *Earth System Science Data*, *13*(7), 3337–3362. https://doi.org/10.5194/essd-13-3337-2021

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

Tyralis, H., Papacharalampous, G., Burnetas, A., & Langousis, A. (2019). 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

Vaysse, K., & Lagacherie, P. (2017). Using quantile regression forest to estimate uncertainty of digital soil mapping products. *Geoderma*, *291*, 55–64. https://doi.org/10.1016/j.geoderma.2016.12.017

Wang, F., Tian, D., Lowe, L., Kalin, L., & Lehrter, J. (2021). Deep learning for daily precipitation and temperature downscaling. *Water Resources Research*, *57*(4), e2020WR029308. https://doi.org/10.1029/2020WR029308

Wang, S., Ancell, B. C., Huang, G. H., & Baetz, B. W. (2018). Improving robustness of hydrologic ensemble predictions through probabilistic pre- and post-processing in sequential data assimilation. *Water Resources Research*, *54*(3), 2129–2151. https://doi.org/10.1002/2018WR022546

Yang, Q., Wang, Q. J., & Hakala, K. (2021). 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

Zhang, L., Li, X., Zheng, D., Zhang, K., Ma, Q., Zhao, Y., & Ge, Y. (2021). Merging multiple satellite-based precipitation products and gauge observations using a novel double machine learning approach. *Journal of Hydrology*, *594*, 125969. https://doi.org/10.1016/j.jhydrol.2021.125969

Zhang, Y., & Ye, A. (2021). Machine learning for precipitation forecasts post-processing—Multi-model comparison and experimental investigation. *Journal of Hydrometeorology*, *22*(11), 3065–3085. https://doi.org/10.1175/JHM-D-21-0096.1

Zhang, Y., Ye, A., Nguyen, P., Analui, B., Sorooshian, S., & Hsu, K. (2021a). Error characteristics and scale dependence of current satellite precipitation estimates products in hydrological modeling. *Remote Sensing*, *13*(16), 3061. https://doi.org/10.3390/rs13163061

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

Zhang, Y., Ye, A., Nguyen, P., Analui, B., Sorooshian, S., & Hsu, K. (2022). QRF4P-NRT (Probabilistic Post-Processing of Near-Real-Time Satellite Precipitation Estimates using Quantile Regression Forests) (v1.0). [Software]. Zenodo. https://doi.org/10.5281/zenodo.5917926

Zhao, P., Wang, Q. J., Wu, W., & Yang, Q. (2022). Extending a joint probability modeling 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

Zhao, Y., Zhu, J., & Xu, Y. (2014). Establishment and assessment of the grid precipitation data sets in China for recent 50 yr. *Journal of the Meteorological Sciences*, *34*(4), 414–420. https://doi.org/10.3969/2013jms.0008

Zhou, T., Nijssen, B., Huffman, G. J., & Lettenmaier, D. P. (2014). Evaluation of real-time satellite precipitation data for global drought monitoring. *Journal of Hydrometeorology*, *15*(4), 1651–1660. https://doi.org/10.1175/JHM-D-13-0128.1

Zorita, E., & Von Storch, H. (1999). The analog method as a simple statistical downscaling technique: Comparison with more complicated methods. *Journal of Climate*, *12*(8), 2474–2489. https://doi.org/10.1175/1520-0442(1999)012<2474:TAMAAS>2.0.CO;2