A Novel mRMR-RFE-RF Method for Enhancing Medium- and Long-Term Hydrological Forecasting: A Case Study of the Danjiangkou Basin

Tiantian Tang [ID], *Member, IEEE*, Tao Chen [ID], and Guan Gui [ID], *Fellow, IEEE*

Abstract—In machine learning (ML)-based hydrological forecasting, particularly in medium- and long-term prediction, judicious predictor selection is paramount, as it ultimately determines the forecast accuracy. This study pioneered an advanced predictor-screening method that synergizes the mutual information (MI) and random forest (RF) technologies through minimum-redundancy-maximum-relevance-recursive feature elimination-random forest (mRMR-RFE-RF) method, blending both filtering and wrapping techniques. This method was rigorously tested through a detailed case study in the Danjiangkou basin, where a comprehensive analysis of 1560 meteorological factors was conducted. Employing three sophisticated ML algorithms—RF, eXtreme Gradient Boosting (XGB), and Light Gradient Boosting (LGB)—we developed precipitation forecasting models. Furthermore, we performed an in-depth rationality analysis of high-frequency predictors. The findings from our study show that this novel hybrid screening strategy markedly outperformed conventional singular predictor-screening methods in enhancing the accuracy of precipitation forecasting when integrated into these forecasting models. Moreover, it assured the validity of the high-frequency forecast factors employed. Therefore, this innovative method not only elevates the accuracy of medium- and long-term precipitation forecasting but also contributes a novel perspective to the methodology of predictor selection in hydrological forecasting models.

Index Terms—Machine learning (ML), medium- and long-term hydrological forecasting, screening of predictors.

I. INTRODUCTION

MEDIUM- and long-term hydrological forecasting performed monthly is crucial in hydrological science,

Tiantian Tang and Guan Gui are with the College of Telecommunications and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210003, China (e-mail: tangtt@njupt.edu.cn; guiguan@njupt.edu.cn).

Tao Chen is with Nanjing Hydraulic Research Institute, Hydrology and Water Resources Department, Nanjing 210029, China (e-mail: chent@nhri.cn).

significantly contributing to flood control and disaster mitigation [1]. Historically, these forecasts have relied on process-driven methods that utilize hydrological models known for their explicit structural clarity and solid physical basis. Nevertheless, these process-driven models demand extensive data, which can hinder their applicability in data-scarce regions [2]. In contrast, while data-driven approaches also require substantial data, they can sometimes leverage alternative data sources or imputation techniques to address data limitations in such regions. With the ongoing advancements in computer science and the robust enhancement of digital infrastructure in water management, employing machine learning (ML) techniques for medium- and long-term hydrological forecasting has emerged as a cutting-edge area [3], [4], [5], [6], [7], [8], [9], [10], [11], [12]. In recent years, the field of hydrological forecasting has benefited significantly from the integration of hybrid ML algorithms. Researchers have successfully integrated various ML techniques to enhance the accuracy and efficiency of predicting hydrological events such as rainfall, streamflow, and water quality [13], [14]. These hybrid models often combine the strengths of different ML strategies, such as support vector machines, neural networks, and decision trees to capture the complex nonlinear relationships inherent in hydrological data [15], [16]. The hybrid ML algorithmic process in hydrological forecasting mainly includes predictor selection and the establishment of a forecast model. The choice of predictors is crucial as it shapes the input data for the forecasting-model construction. Moreover, predictor choice is crucial, as it shapes the input data for the forecasting model, directly influencing the forecast outcome. Thus, enhancing medium- and long-term precipitation forecasting accuracy through the development of a predictor selection technology is essential. Such technology should efficiently eliminate redundant factors while ensuring that the selected predictors accurately capture the complex dynamics of the variables being forecasted. This focus on optimizing predictor selection has been a long-standing area of interest among hydrologists and remains an active area in hydrological research [17].

Currently, as ML becomes increasingly integrated into hydrological studies, the techniques for selecting predictors are diversifying. Traditionally, there are three such methods. The first category is the filtering method, which involves selecting predictors from sample datasets prior to training the model. Filtering methods use specific evaluation criteria to "filter" initial features. Features scoring above a certain threshold are selected,

while those below are discarded, and the retained features are then used to train the model. Prominent filtering techniques include mutual information (MI) [18] and partial MI [19]. The second method, the wrapper method, selects predictors based on the forecast accuracy achieved with the forecasting model. The selection is integrated or "wrapped" around the model training process, exemplified by techniques such as the Las Vegas Wrapper and recursive feature elimination (RFE) [20]. The third method is the embedding approach, where feature selection and model training occur simultaneously, with features being selected during the training process. This method is commonly employed in hydrology, with techniques such as random forest (RF) [21] being prevalent. Each method has its strengths and weaknesses. The filtering method is highly efficient and effective at managing large-scale data sets, although it might not always yield the highest performance across all contexts. The wrapper and embedding methods typically offer improved performance and manage complexity well but may lead to overfitting. To leverage the benefits of each, researchers have proposed hybrid feature-selection methods, aiming to optimize the feature-selection process by combining the strengths of different techniques [22]. Hybrid methods for feature selection are extensively applied across various fields to identify relevant factors. For instance, researchers in bioinformatics have used a hybrid feature-selection method based on the minimum-redundancy-maximum-relevance (mRMR) method coupled with binary differential evolution for gene selection, which is crucial for disease diagnosis and drug development [23]. Additionally, hybrid feature-selection techniques are employed in single-cell transcriptomics to support detailed analyses of biological systems [24]. A benchmark study of feature-selection strategies for multiomics data also extensively employed hybrid methods to enhance the accuracy of disease risk-prediction accuracy [25]. Furthermore, optimized hybrid approaches combining chi-square and particle swarm optimization algorithms have been applied in intelligent systems for tasks such as intrusion detection and cancer detection, showcasing their effectiveness in improving ML models [26]. At the same time, hybrid methods for selecting forecasting factors are also widely used in hydrological forecasting. For example, Liu et al. [27] developed a runoff forecasting model based on a predictor screening method CC-PCA, which outperformed the other predictors-screening methods. Schmidt et al. [19] used RFE [28] and feature-importance ranking [29] to select forecast predictors from dynamic factors such as rainfall, soil texture, and air temperature as the input of RF, and carried out flood forecast research in some areas of Germany.

In this study, we introduce an advanced predictor-screening method, minimum-redundancy-maximum-relevance-recursive feature elimination-random forest (mRMR-RFE-RF), which integrates MI and the RF's recursive elimination capabilities. Please note that the initial work has been presented in part in [30]. This method employs the mRMR technique to initially sift through a wide array of predictors. Subsequently, the recursive elimination process, anchored in RF algorithms, is used to ascertain the most relevant predictors. To validate the efficacy of this innovative screening tool, we applied it in the Danjiangkou basin, situated in the Yangtze River's upper and middle reaches. Here, we constructed and compared three sophisticated ML models for precipitation forecasting against models derived from a singular predictor-screening method. This comparative analysis demonstrated our proposed method's superior accuracy and effectiveness. Furthermore, we conducted a thorough, grid-point-specific investigation, meticulously identifying and quantifying high-frequency influential factors, followed by an exploratory analysis of their physical mechanisms affecting precipitation. This comprehensive evaluation affirmed the validity and ingenuity of our novel predictor-selection methodology.

This article significantly expands upon the preliminary findings initially discussed at the 2023 IEEE 23rd International Conference on Communication Technology. The conference paper served as an introduction to the novel mRMR-RFE-RF predictor screening method, showcasing its potential through initial applications and early results. It primarily focused on the introduction of the methodology and provided a basic validation of the model's effectiveness in medium- and long-term hydrological forecasting. In contrast, this article extends these discussions by presenting a more comprehensive validation and a deeper analysis. It offers comparisons involving more ML models and incorporates additional evaluation metrics such as the mean absolute percentage error (MAPE) to assess forecast accuracy more rigorously. Furthermore, the article delves into a detailed examination of the rationality of predictors and the physical mechanisms affecting their performance, which were only briefly mentioned in the conference article. It also includes a new section on the reliability analysis of factors, providing a systematic evaluation of the stability and dependability of the selected predictors. Moreover, the journal version elaborates on the theoretical advancements and practical applications brought about by the mRMR-RFE-RF method. It explores the effectiveness of this method across different hydrological scenarios and assesses its impact on enhancing forecast accuracy. This not only reinforces the original findings but also contributes new research dimensions to the field of hydrological science. This progression from a conference setting to a detailed journal publication is emblematic of the academic pathway of rigorous investigation and peer review, which helps us to understand more deeply and apply innovative scientific methods in real-world scenarios.

II. STUDY AREA AND DATASET

A. Study Area

The Danjiangkou Basin (see Fig. 1), located above the control section of the Danjiangkou Reservoir in the central and upper reaches of the Yangtze River basin, occupies a strategic position. It is located between latitudes $31° \sim 34°$ N and longitudes $106° \sim 112°$ E. This basin, encompassing an expansive area of approximately $95\,217$ km^2, constitutes about 60% of the entire Han River basin. Characterized by a temperate climate, the basin experiences an average annual temperature ranging between $15°$ and 17 °C. The region also exhibits significant evaporative patterns, with annual rates varying from 900 to
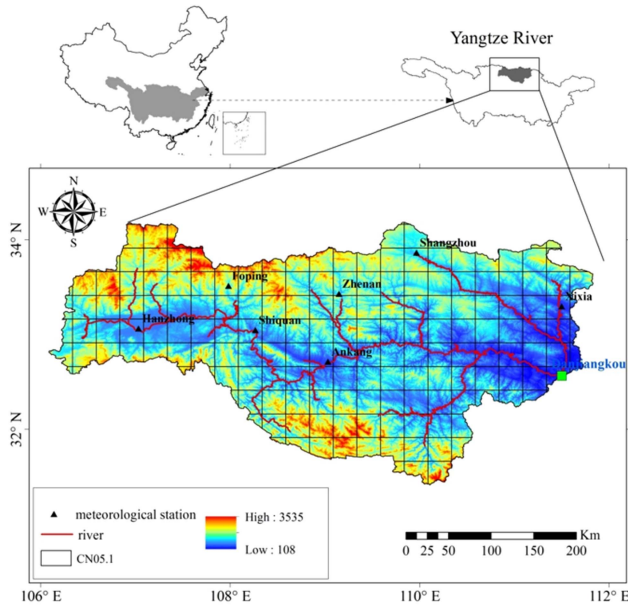
Fig. 1. Map of the Danjiangkou basin.

1500 mm. Notably, the basin benefits from copious precipitation and ample water vapor. However, this bounty is not uniformly distributed throughout the year. The bulk of the rainy season is concentrated in the months from May to October, during which period over 80% of the yearly precipitation occurs. The average annual precipitation in the Danjiangkou Basin is estimated to be 700∼1100 mm, highlighting the region's substantial contribution to the river's hydrology.

B. Datasets

1) Precipitation Datasets: We employed the CN05.1 gridded precipitation dataset, illustrated in Fig. 1. Developed by Wu et al. [31], CN05.1 is based on the methodology used for the CN05 dataset [32]. It offers a spatial resolution of 0.25×0.25. The data span from 1982 to 2015.

2) Predictors Datasets: The study gathered 130 meteorological and climatic index data from 1982 to 2015, which comprised 88 atmospheric circulation indexes, 26 monthly sea surface temperature (SST) indexes, and 16 other indexes. The data were sourced from the website of the National Climate Center (NCC) of China. At the same time, considering that the large-scale meteorological-climatic index has a certain lag in the formation of precipitation, there might be a remote correlation between the two. This study considered the meteorological-climatic index 12 months in advance as the candidate predictors set. When the lead time was 1 month, the meteorological-climatic index used in forecasting the precipitation in January of that year ranged from January to December of the previous year. The meteorological-climatic index time was from February of the previous year to January of that year, when the precipitation in February of that year was predicted. When the lead time was 2 months, the meteorological-climatic index time was from January to December of the previous year, when the precipitation in February of that year was predicted in early January, and from

February of last year to January of that year, when the precipitation in March of that year was predicted in early February, and so on. Therefore, the set of predictors to be selected for the precipitation forecast in each month contained 1560 (12×130) predictors.

III. METHODOLOGY

To validate the predictor-screening method's accuracy, we developed three precipitation forecasting models using ML techniques: RF, eXtreme Gradient Boosting (XGB), and Light Gradient Boosting (LGB). The effectiveness and precision of the predictor-screening method's accuracy were verified by comparing precipitation forecasting accuracy before and after the predictor-screening implementation. Consequently, this section is structured into two main parts: the first introduces the three ML models, and the second details the process of predictor screening.

A. Ml-Based Precipitation Forecasting Method

1) Random Forest: This robust ML approach[29] integrates Bagging ensemble learning theory with the random subspace method [33]. It employs Bootstrap technology to sample from the original dataset, enabling the creation of new training subsets. Each subset then uses the random subspace method to randomly select feature attributes for constructing a decision tree. The final prediction is determined by aggregating the outputs of these trees, typically through voting or averaging [17]. The application of RF in precipitation forecasting involves several key steps.

1) The selected M predictors and corresponding precipitation data are combined to form the training dataset $D = \{(x_i, y_i), x_i \in X, y_i \in Y, i = 1, 2, \ldots, N\}$, where X represents an M-dimensional vector of predictors, Y denotes the series of precipitation measurements, and N is the total number of samples.

2) Using the Bootstrap resampling technique, generate k new subsets of the training dataset D, each matching the size of the original dataset.

3) For each of the k training subsets, construct a CART decision tree. During this process, $m = \sqrt{M}$ indicators are randomly selected from the set of predictors according to the random subspace method to determine the attributes at the nodes of the tree.

4) Each decision tree is developed recursively from the top down. Upon completion, each tree produces a predicted output. The results from all the CART trees are then aggregated through voting (for classification) or averaging (for regression) to produce the final forecasted value.

2) eXtreme Gradient Boosting: This is an ensemble of boosting algorithms introduced by Chen and Guestrin [34]. It integrates K CART decision trees into a model. For the sample set $D = \{(x^{(i)}, y^{(i)})\} (|D| = N, x^{(i)} \in \mathbb{R}^m, y^{(i)} \in \mathbb{R})$, XGB linearly combines these K weak learners as follows:

$$\widehat{y^{(i)}} = \phi(x^{(i)}) = \sum_{k=1}^{K} f_k \left(x^{(i)} \right). \tag{1}$$

The loss function for XGB is expressed as follows:

$$L_t = \sum_{i=1}^{N} L\left(y^{(i)}, f_{t-1}(x^{(i)}) + h_t(x^{(i)})\right) + \Omega(h_t) \quad (2)$$

where $f_{t-1}(x^{(i)})$ represents the output of a learner from a previous iteration and $h_t(x^{(i)})$ is the tree introduced in the tth iteration to minimize the loss function. The regularization term $\Omega(h_t)$ helps prevent overfitting

$$\Omega(h_t) = \gamma J + \frac{\lambda}{2} \sum_{j=1}^{J} \omega_{tj}^2 \quad (3)$$

where J is the number of leaf nodes, ω_{tj} represents the value at the jth leaf, and γ and λ are coefficients adjusted during practical applications. Our goal is to minimize the loss function and obtain the corresponding model. The first- and second- derivatives of the ith sample in the tth weak learner in the first round are, respectively, recorded as follows:

$$g_{ti} = \frac{\partial L\left(y^{(i)}, f_{t-1}\left(x^{(i)}\right)\right)}{\partial f_{t-1}\left(x^{(i)}\right)} \quad (4a)$$

$$h_{ti} = \frac{\partial^2 L\left(y^{(i)}, f_{t-1}\left(x^{(i)}\right)\right)}{\partial f_{t-1}^2\left(x^{(i)}\right)}. \quad (4b)$$

Since the value of the jth leaf node of each regression tree is ω_{tj}, it can be simplified as follows:

$$L_t \approx \sum_{i=1}^{K} g_{ti} h_t\left(x^{(i)}\right) + \frac{1}{2} h_{ti} h_t^2\left(x^{(i)}\right) + \gamma J + \frac{\lambda}{2} \sum_{j=1}^{J} \omega_{tj}^2$$

$$= \sum_{j=1}^{J} \left(\left(\sum_{x^{(i)} \in R_{tj}} g_{ti} \right) \omega_{tj} + \frac{1}{2} \left(\sum_{x^{(i)} \in R_{tj}} h_{ti} + \lambda \right) \omega_{tj}^2 \right) + \gamma J. \quad (5)$$

Define $G_{tj} = \sum_{x^{(i)} \in R_{tj}} g_{ti}$, $H_{tj} = \sum_{x^{(i)} \in R_{tj}} h_{ti}$, then the final loss function can be expressed as:

$$L_t = \sum_{j=1}^{J} \left(G_{tj} \omega_{tj} + \frac{1}{2} \left(H_{tj} + \lambda \right) \omega_{tj}^2 \right) + \gamma J. \quad (6)$$

At this point, the optimization problem of the loss function is transformed into solving the optimal J leaf nodes of the regression tree and the optimal value ω_{tj} corresponding to each leaf node. First, the derivative of ω_{tj} is taken based on the loss function to make the derivative 0. Then, the following is obtained:

$$\omega_{tj} = -\frac{G_{tj}}{H_{tj} + \lambda}. \quad (7)$$

Substituting the loss function, we can obtain the following:

$$L_t = -\frac{1}{2} \sum_{j=1}^{J} \frac{G_{tj}^2}{H_{tj} + \lambda} + \gamma J. \quad (8)$$

The score function described by (8) is employed to assess the efficacy of the tree's structure, with Fig. 2 illustrating the calculation of this score. A lower score indicates a more optimal
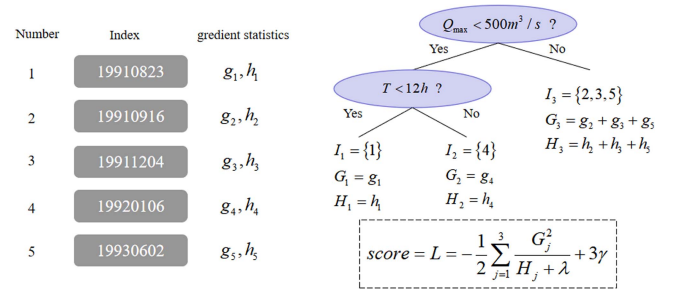

Fig. 2. Structure score calculation.

tree structure. The focus then shifts to the division of nodes. Recall that the sums of the first and second derivatives for the left and right subtrees at a given node are denoted as G_L, H_L, G_R, and H_R, respectively. The function L_{split} serves to evaluate potential segmentation points for the nodes

$$L_{split} = \frac{1}{2} \frac{G_L^2}{H_L + \lambda} + \frac{1}{2} \frac{G_R^2}{H_R + \lambda} - \frac{1}{2} \frac{(G_L + G_R)^2}{H_L + H_R + \lambda} - \gamma. \quad (9)$$

Finally, the goal is to find a feature and its corresponding optimal value, so that the above formula can obtain the maximum value.

3) Light Gradient Boosting: LGB is a type of boosting algorithm introduced by Microsoft in 2017 [35]. Traditional boosting algorithms require traversing all features to select the optimal split, which can be inefficient and time-consuming with large-scale, high-dimensional data. Addressing these challenges, LGB incorporates two innovative techniques: gradient-based one-side sampling (GOSS) and exclusive feature bundling (EFB). These methods enhance the efficiency of the model by reducing the data size needed for training without significantly sacrificing accuracy and by efficiently handling feature interactions, respectively.

The GOSS algorithm is advantageous because it does not use all the sample points to calculate the gradient but, rather, only samples the original sample to calculate the gradient. It reserves data with large errors, resulting in large absolute gradient values. It takes a random sample of data with small errors, resulting in small absolute gradient values. Owing to the improvement of the GOSS algorithm, no sample has a large loss error, and it does not change the data distribution while reducing the training data. Therefore, the GOSS algorithm increases the training speed of the model without affecting the accuracy. Furthermore, the mutually exclusive characteristic of the LGB selection is the histogram algorithm, which first discretizes the continuous forecast into the subvalue into the K integer while constructing a histogram, with the width K. When traversing the forecasting factor of the training set, the histogram is used to accumulate the values of the statistical discreteness. After the data are traversed, the histogram accumulates the statistics required. Then, according to the discrete value of the histogram, the optimal segmentation point can be found.

Differently from XGB, which utilizes a levelwise algorithm for constructing decision trees, LGB employs a leafwise growth strategy with depth limitation. In this approach, all leaf nodes are

traversed before making a split, and the node with the maximum split gain is chosen for division. This leaf-growth strategy tends to reduce splitting errors and achieve higher simulation accuracy when the number of splits remains constant. However, because deeper decision trees are more susceptible to overfitting, LGB implements strict maximum depth restrictions on the decision trees to mitigate this risk.

B. Predictor Selection Based on MI and RF Recursive Elimination

During predictor selection, it is crucial to eliminate redundant factors while ensuring that the selected predictors effectively capture the complex dynamics of the forecasted object. Initially, the mRMR method is employed to choose 50 predictors. The screening method of recursive elimination forecasting factors based on RF (RFE-RF) is used to further narrow down the selection of 50 predictors to 20 predictors, which are deemed the final set of predictors.

1) Minimum-Redundancy-Maximum-Relevance: MI serves as an information measure, quantifying the amount of shared information between variables. It signifies the degree of correlation between variables and can capture nonlinear relationships beyond linear correlations. Higher MI values denote stronger correlations between variables. Peng et al. [36] introduced mRMR method, which builds upon the principles of MI. mRMR aims to enhance the effectiveness of measuring the correlation between predictors and predictands by eliminating redundant information between predictors and optimizing predictor selection. This is achieved through the calculation of information differences and information entropy to identify the "optimal" predictor.

For the subset S of predictors for which $m\{x_i\}$ predictors need to be found, the maximum correlation between predictor x_i and forecast variable c can be expressed as follows:

$$\max D(S,c), \ D = \frac{1}{|S|} \sum_{x_i \in S} I(x_i; c). \tag{10}$$

The minimum redundancy between predictor x_i and predictor x_j can be expressed as follows:

$$\min R(S), \ R = \frac{1}{|S|^2} \sum_{x_i, x_j \in S} I(x_i; x_j). \tag{11}$$

Integrating the maximum correlation degree and the minimum redundancy degree to make it optimal aims to solve

$$\max \Phi(D,R), \ \Phi = D - R. \tag{12}$$

Therefore, assuming that we have a subset of forecast factors S_{m-1}, our goal is to find the mth predictor from the remaining factor subset $X - S_{m-1}$. Through predictor screening, the formula of mRMR method is expressed as follows:

$$\max_{x_j \in X - S_{m-1}} \left[I(x_j; c) - \frac{1}{m-1} \sum_{x_i \in S_{m-1}} I(x_i; x_j) \right]. \tag{13}$$

2) RFE-RF: The predictor-screening method based on RF comprises two main parts. First, RF is utilized to select a subset of predictors. Subsequently, RFE is applied recursively to refine the predictors selected by RF. The primary steps of predictor screening based on RF are outlined below.

1) *Step 1:* Calculate the values of importance S of each predictor and sort them in order. Assuming that there are M trees in the forest, the importance of the predictor X can be expressed as follows:

$$S = \frac{\sum_{i=1}^{M} (\mathrm{errOOB2}_i - \mathrm{errOOB1}_i)}{M} \tag{14}$$

where errOOB1 is the out-of-bag error of the decision tree, and errOOB2 is the out-of-bag error after adding noise interference.

2) *Step 2:* Determine the proportion to be eliminated, and obtain a new predictor set to remove the corresponding proportion based on the importance of the predictors.

3) *Step 3:* Set the number m of predictors that must be screened, and repeat the above process with a new predictor set until m predictors remain.

4) *Step 4:* Based on the out-of-bag error rates corresponding to the various predictor sets in the above process, select the predictor set with the lowest external error rate.

RFE is categorized as a wrapping method in feature selection [37]. Its core concept involves employing a base ML model for iterative training rounds. After each iteration, several predictors with their corresponding weight coefficients are eliminated, leading to the next round of training with the updated predictor set. This method is instrumental in obtaining a more efficient set of predictors and operates as a greedy algorithm for discovering the optimal feature subset. In this article, the chosen base learner is an RF model. Consequently, the main steps of RFE-RF can be summarized as follows.

1) *Step 1:* Measure the importance of predictors by using the RF model.

2) *Step 2:* Select the index with the greatest importance of the predictors (factor elimination).

3) *Step 3:* Repeat steps 1 and 2 on the remaining predictor sets, and repeatedly build an RF model until all predictors are traversed.

4) *Step 4:* The order in which the predictors are eliminated is the order of the predictors until the specified k factors are screened out.

C. Accuracy Evaluation Index

1) Anomaly Correlation Coefficient (ACC): ACC is a statistical measure used primarily in meteorology to assess geographical and temporal forecasts' accuracy by comparing predicted anomalies to observed anomalies. An anomaly in this context refers to the deviation of a parameter (e.g., temperature or precipitation) from its long-term average. ACC is widely used to evaluate forecasting models in terms of their ability to capture the variability of weather or climate patterns. The formula for ACC is expressed as follows:

$$\mathrm{ACC}_p = \frac{\sum_{k=1}^{K} \left(\delta r_{k,p} - \overline{\delta r_p}\right) \cdot \left(\delta g_{k,p} - \overline{\delta g_p}\right)}{\sqrt{\sum_{k=1}^{K} \left(\delta r_{k,p} - \overline{\delta r_p}\right)^2 \cdot \sum_{k=1}^{K} \left(\delta g_{k,p} - \overline{\delta g_p}\right)^2}} \tag{15}$$

where $k = 1, 2, \ldots, K$ represents the number of grid points within the catchment; $p = 1, 2, \ldots, P$ is the total number of data points in the precipitation series; $r_{k,p}$ denotes the actual precipitation measurement and $g_{k,p}$ denotes the predicted precipitation value; $\overline{\delta r_k}$ is the mean observed precipitation at grid point k, and $\delta r_{k,p}$ are the observed precipitation deviations at instance p; $\overline{\delta r_p}$ denotes the mean observed precipitation deviations across all grid points at instance p; $\overline{\delta g_k}$ denotes the long-term mean forecasted precipitation at grid point k, and $\delta g_{k,p}$ denote the forecasted precipitation deviations at instance p; $\overline{\delta g_p}$ denotes the mean forecasted precipitation deviations across all grid points in the catchment.

2) Mean Absolute Percentage Error: MAPE is a statistical measure used to assess the forecasting model's accuracy. It quantifies the size of the error in percentage terms and is commonly used because it is easy to interpret and provides a clear indication of how off the predictions are, on average, from the actual values. The formula for MAPE is given as follows:

$$\text{MAPE} = \left(\frac{1}{n} \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right| \right) \times 100\% \qquad (16)$$

where y_i is the actual value, $\hat{y}_i$ is the predicted value, and n is the number of observations.

IV. RESULTS AND DISCUSSIONS

A. Construction of ML Precipitation Forecasting Model

The ML model developed in this study relies on the scikit-learn ML library in the Python programming language. In addition to the model's source code, the key methodologies encompass leave-one-out cross-validation (LOOCV) and model parameter tuning. Cross-validation enhances the model's accuracy through iterative training and helps optimize model parameters by evaluating cross-validation accuracy.

1) LOOCV: When training the three ML models (RF, XGB, LGB), we employed LOOCV, which effectively mitigates overfitting and assesses the generalization capability of the models. In cases where sample datasets are limited, traditional methods of dividing data into training and validation sets may lead to reduced training data, hindering the construction of models with robust predictive capabilities. LOOCV maximizes the utility of available training data by iteratively training on $N - 1$ samples and validating on the remaining sample, repeating this process N times until each sample has been used as a validation set. The validation error is then averaged to obtain the final validation result.

However, it is important to acknowledge the limitations of LOOCV. Since nearly all data points are used for training in each iteration, the models trained on each fold are very similar to those trained on the entire dataset. This can lead to high variance in performance metrics, causing significant fluctuations between iterations. Such variability can result in unreliable estimates of the model's true performance. Additionally, LOOCV can contribute to overfitting, where the model performs well on training data but poorly on unseen data. The small test set in each iteration may not provide a robust estimate of the model's generalization

TABLE I
METHODS FOR TUNING PARAMETERS AND THE SETTINGS OF ML MODELS

Model	Method	Parameter	Interval
RF	grid search	n_estimators	$50 \sim 500$
		max_depth	$0 \sim 10$
		max_features	$0 \sim 15$
		min_samples_leaf	$0 \sim 15$
XGB	grid search	n_estimators	$50 \sim 500$
		learning_rate	$0 \sim 1$
		max_depth	$0 \sim 10$
		α	$0 \sim 1$
		λ	$0 \sim 1$
		γ	positive number
LGB	grid search	n_estimators	$50 \sim 500$
		learning_rate	$0 \sim 1$
		num_leaves	$0 \sim 2^{\text{max\_depth}}$
		max_depth	$0 \sim 10$
		min_child_weight	$0 \sim 10$
		α	$0 \sim 1$
		λ	$0 \sim 1$
		γ	positive number

ability. Despite these limitations, LOOCV was chosen in this research due to the relatively small size of the dataset, ensuring maximum utilization of available data for training.

2) Parameter Adjustment of ML Model: The ML models (RF, XGB, and LGB) used in this research, generally have adjustable parameters that vary widely, often from dozens to several hundreds. Table I outlines the various methods for tuning these parameters, the names of the parameters themselves, and the ranges within which they can be adjusted for each model. Nevertheless, owing to the diversity in forecasting lead times, the months considered, and the geographical granularity of the data, the precise settings of these parameters are not detailed in this study.

B. Results of ML Precipitation Forecasting Model

In this analysis, we validate and explore the effectiveness of the mRMR-RFE-RF method for selecting predictors. The selection process yields 20 predictors, combined with a precipitation dataset covering the years 1982 to 2015, which are then fed into three different ML models: RF, XGB, and LGB. To benchmark the performance of the mRMR-RFE-RF method, it is compared against two other methods: one based solely on MI and another utilizing a RF approach. For simplicity, this study refers to the mRMR-RFE-RF method as MRR.

1) ACC Scores for ML-Based Precipitation Forecasting: The ACC score measures the correlation between predicted and observed precipitation data across all grid points, with possible values ranging from -1 to 1, where higher scores indicate more accurate forecasting. Figs. 3–5 illustrate the performance of three different models (XGB, LGB, RF) using three predictor-screening methods (mRMR-RFE-RF, MI, RF) for precipitation forecasting. The figures display the ACC values for each month from 1985 to 2015, showcasing the predictive accuracy of different methods and models.
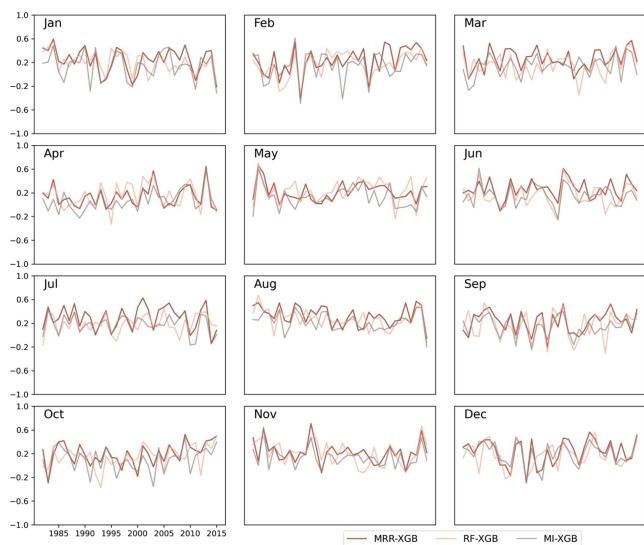
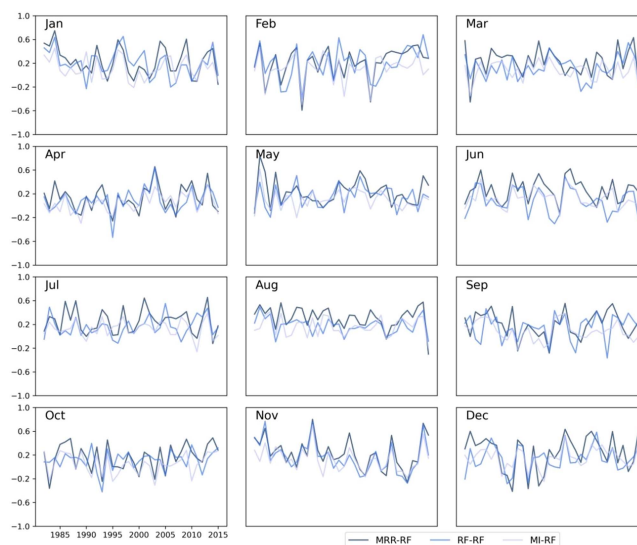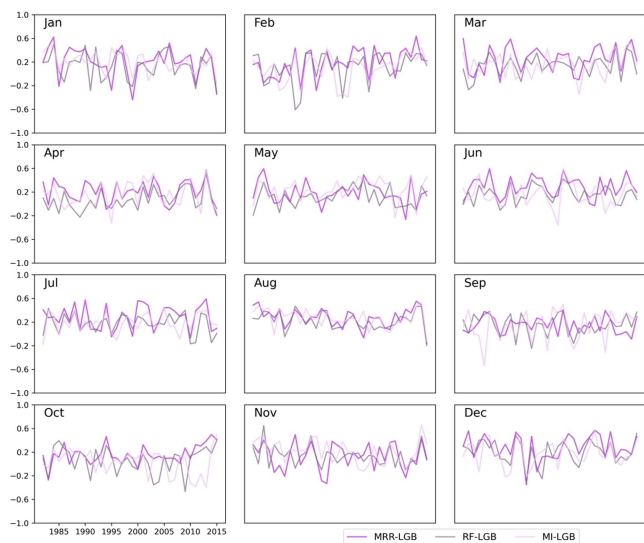Fig. 3. ACC of the XGB model forecast by three predictor-screening methods.

Fig. 4. ACC of the LGB model forecast by three predictor-screening methods.

Fig. 3 shows the ACC values for precipitation forecasts using the XGB model with the same three predictor-screening methods. The figure highlights the ACC values from 1985 to 2015 for each month. The mRMR-RFE-RF method again shows superior performance in several specific months, notably in March, May, August, and November. For example, in March, the ACC value for mRMR-RFE-RF is approximately 0.35, while the values for RF and MI are around 0.2 and 0.25, respectively. Moreover, in August, the ACC value for mRMR-RFE-RF is about 0.3, compared to around 0.2 for the other methods.

Fig. 4 illustrates the ACC values for precipitation forecasts using the LGB model with three predictor-screening methods. The figure presents the ACC values from 1985 to 2015 for each month. The mRMR-RFE-RF method consistently outperforms the other methods in many months, particularly in February,

Fig. 5. ACC of RF model forecast by three predictor-screening methods.

June, September, and December. For instance, in February, the ACC value for mRMR-RFE-RF is approximately 0.3, while the values for RF and MI are below 0.2. Additionally, in September, the ACC value for mRMR-RFE-RF is around 0.25, significantly higher than the other methods, which range from 0.1 to 0.15. These results highlight the superior predictive performance of the MRR method across different months and models.

Fig. 5 presents the ACC values for precipitation forecasting using the RF model combined with three different predictor-screening methods (mRMR-RFE-RF, MI, RF). The data reveal that the mRMR-RFE-RF method outperforms the other methods in several specific months, particularly in January, April, July, and October. For instance, in January, the ACC value for mRMR-RFE-RF is approximately 0.4, while the values for RF and MI are below 0.3. This indicates that the mRMR-RFE-RF method provides higher accuracy and stability when using the RF model for precipitation forecasting. Additionally, in October, the ACC value for mRMR-RFE-RF is around 0.35, significantly higher than the other methods, which hover around 0.2.

The results from Figs. 3–5 collectively indicate that the mRMR-RFE-RF method demonstrates significant advantages in precipitation forecasting across different models and various months. The mRMR-RFE-RF method consistently shows higher ACC values, indicating greater predictive accuracy and stability compared to other predictor-screening methods. This emphasizes the practicability and reliability of mRMR-RFE-RF method in precipitation forecasting.

Based on the results depicted in Fig. 6, the predictors selected from the mRMR-RFE-RF predictor-screening method, when input into the XGB, LGB, and RF precipitation forecast models, exhibit distinct patterns in terms of monthly precipitation forecasting accuracy index (ACC) values. The XGB model demonstrates relatively higher ACC values across most months, with minimal fluctuations. Notably, in months such as January, March, June, and December, the XGB model's ACC values are more stable and consistent compared to the other models.
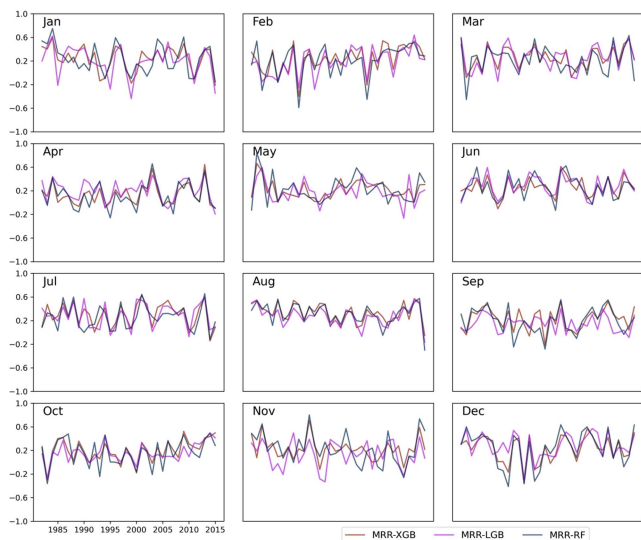
Fig. 6.   ACC of three models predicted by MRR predictor-screening method.

For instance, in January, the ACC value of the XGB model ranges from −0.2 to 0.4, in March it is between 0 and 0.4, in June it fluctuates between −0.2 and 0.4, and in December it varies from −0.2 to 0.6, all of which indicate a high degree of stability and consistency. In contrast, the ACC values for the LGB and RF models show greater variability in these months, and their performance is less stable than that of the XGB model. Consequently, the XGB model exhibits superior performance in monthly precipitation forecasting in the Danjiangkou Basin.

2) MAPE of ML Precipitation Forecast Model: As mentioned above, MAPE measures the average absolute percentage difference between the forecasted values and the observed values, expressed as a percentage. MAPE values range from 0% to $+\infty$, with lower values indicating higher accuracy. A MAPE of 0% would indicate a perfect forecast, where the forecasted values exactly match the actual values. Figs. 7–9 show a series of heatmaps for different models over the months of the year, indicating MAPE for precipitation forecasting. In each heatmap, color intensity represents the average relative error, with warmer colors (reds) indicating higher errors and cooler colors (blues) suggesting lower errors. The intensity of the color correlates with the magnitude of the error, the deeper the red, the higher the error, and the deeper the blue, the lower the error.

Fig. 7 illustrates the spatial distribution of MAPE after applying three different predictor-screening methods to the XGB precipitation forecasting model. In January, the prediction accuracy is poorer in the left and upper right regions of the basin. By February, these areas show improved accuracy, with March seeing even more significant improvements. In September and October, the right side of the basin experiences lower prediction accuracy, and the entire basin shows unsatisfactory precision in November and December. Conversely, from May to August, the prediction accuracy across the basin is relatively high. Analyzing the average MAPE data from 167 grid points within the basin,

it is evident that the mRMR-RFE-RF method performs exceptionally well in most months, significantly outperforming the MI and RF single predictor-screening methods. For instance, in January, the MAPE of mRMR-RFE-RF is 0.57, approximately 11.4% lower than MI's 0.64 and 3.6% lower than RF's 0.59. Similarly, in March, the mRMR-RFE-RF' MAPE is 0.46, about 13.2% lower than MI and 1.2% lower than RF. In May, the mRMR-RFE-RF error is 0.32, approximately 7.8% lower than MI's 0.34 and 1% lower than RF's 0.32. A similar trend is observed in July, where the mRMR-RFE-RF error is 0.31, significantly lower than MI and RF by about 10.9% and 7.8%, respectively.

Fig. 8 shows the spatial distribution of MAPE after applying the same three predictor-screening methods to the LGB precipitation forecast model. From January to April, the prediction accuracy in the upper and middle parts of the basin is lower than in the downstream regions, with improvements as the months progress. From May to August, the entire basin exhibits higher prediction accuracy. In September and October, the right side of the basin has lower accuracy compared to the left side. November and December show the lowest prediction accuracy for the entire year. For the LGB method, the mRMR-RFE-RF also outperforms MI and RF in most months. For example, in January, the mRMR-RFE-RF error is 0.65, about 15.5% lower than MI's 0.78. In May, the mRMR-RFE-RF error is 0.33, approximately 19.1% lower than MI's 0.41. In July, the mRMR-RFE-RF error is 0.33, about 18.1% lower than MI's 0.40. In August, the mRMR-RFE-RF error is 0.35, approximately 17.2% lower than MI's 0.42.

The results in Fig. 9 align with the previous figures, showing lower prediction accuracy in months with less precipitation and higher accuracy in months with more precipitation. Similarly, the mRMR-RFE-RF method proves to be more accurate than the other two single methods. For the RF method, mRMR-RFE-RF performs better than MI and RF in most months. For instance, in January, the mRMR-RFE-RF error is 0.58, about 23.5% lower than MI's 0.76 and 17.4% lower than RF's 0.71. In March, the mRMR-RFE-RF error is 0.48, about 21.1% lower than MI's 0.61. In May, the mRMR-RFE-RF error is 0.32, approximately 18.9% lower than MI's 0.39. In July, the mRMR-RFE-RF error is 0.32, significantly lower than both MI and RF by approximately 20.9% and 19.4%, respectively.

A comprehensive analysis of the data from these three figures reveals that the mRMR-RFE-RF method exhibits significant superiority in predictor-screening. This method greatly enhances precipitation prediction accuracy by accurately identifying key predictors. In evaluating the performance of different precipitation forecasting models, the XGB model stands out with its high prediction accuracy throughout the year and across the entire basin, whereas the LGB model shows less satisfactory overall accuracy. Further monthly analysis indicates that during May to August, when precipitation is higher and data distribution is relatively uniform, the prediction accuracy is generally higher. However, during the dry season in November and December, the MAPE values are generally higher due to the smaller observed values, which is a characteristic of the MAPE calculation
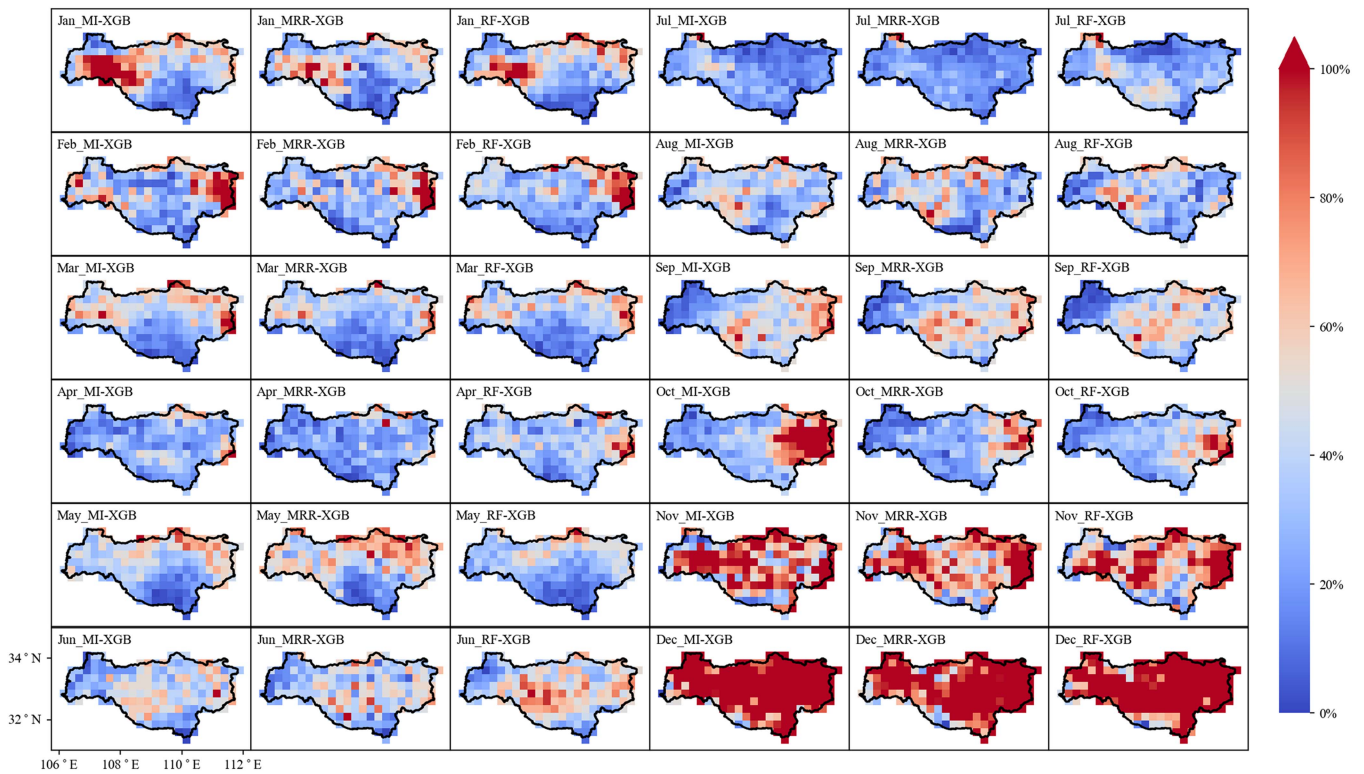
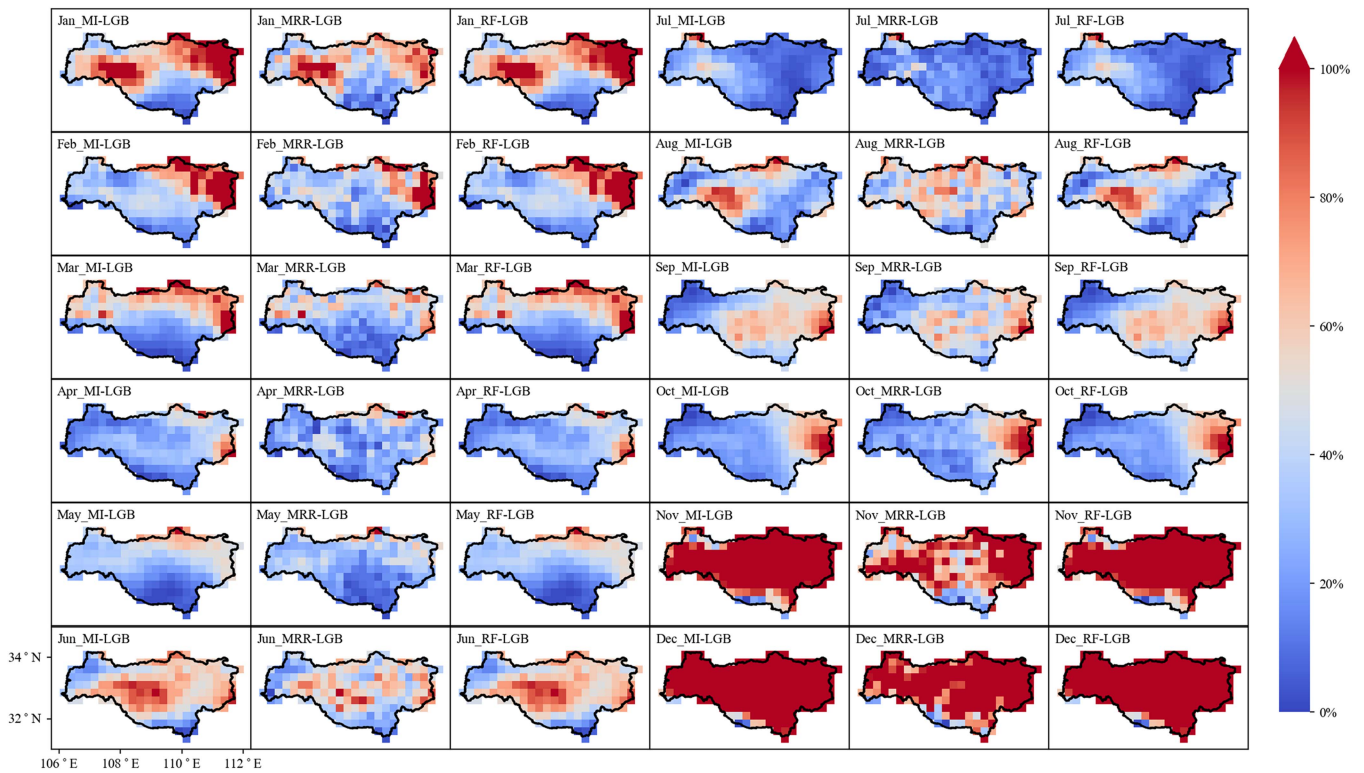Fig. 7.   MAPE of XGB model forecast by three predictor-screening methods.

Fig. 8.   MAPE of the LGB model forecast by three predictor-screening methods.
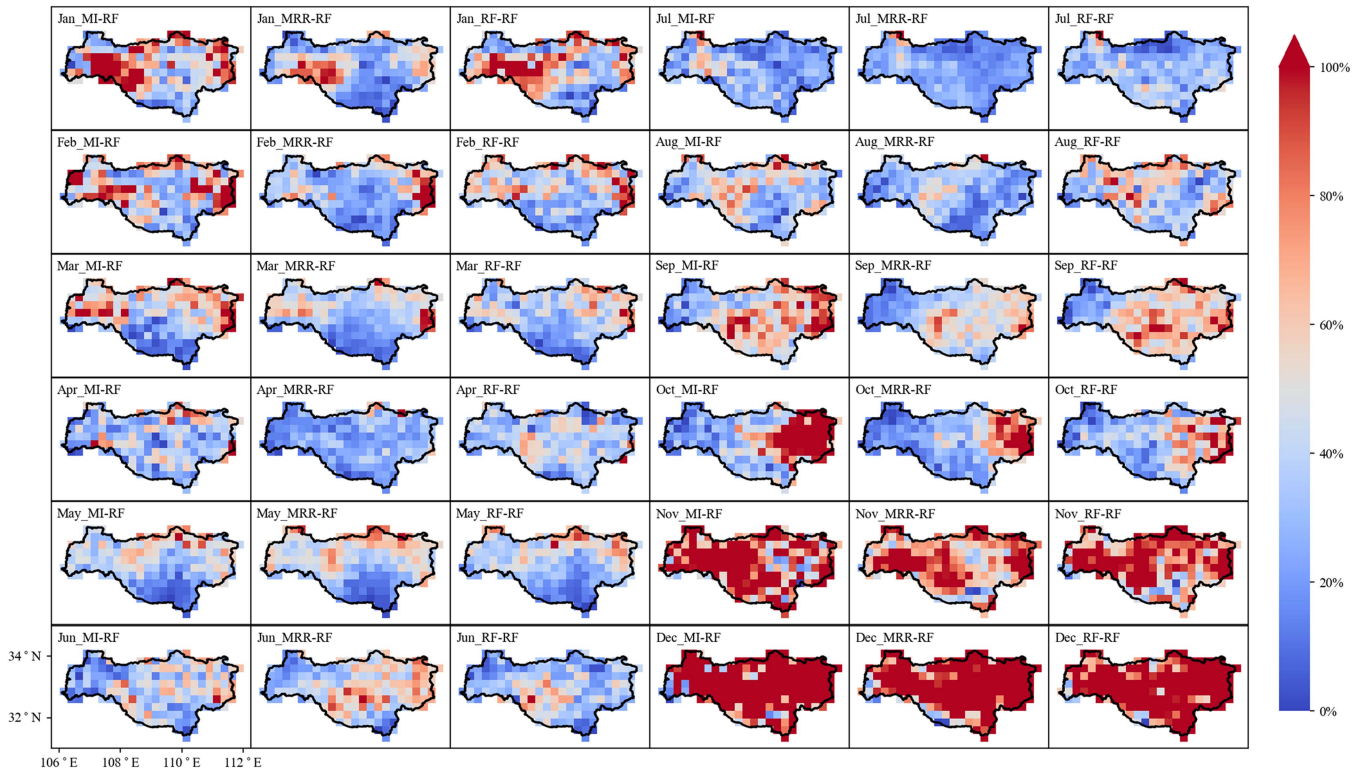
Fig. 9.   MAPE of the RF model forecast by three predictor-screening methods.

formula, amplifying error proportions when observed values are low. Additionally, the study notes significant spatial dependence in precipitation prediction effectiveness across different geographic regions. This phenomenon suggests that precipitation patterns may be influenced by neighboring areas, highlighting the importance of spatial correlation in precipitation prediction. This finding offers a new perspective for future research, exploring how spatial dependence affects precipitation prediction and considering how to leverage this characteristic to optimize prediction models.

In summary, the mRMR-RFE-RF method consistently outperforms the single predictor-screening methods MI and RF across all models, particularly in months with stable data. The mRMR-RFE-RF method demonstrates remarkable performance, especially in the first half of the year, with errors consistently lower than other methods. Even in November and December, despite increased errors due to seasonal variations in precipitation data, the mRMR-RFE-RF method maintains good stability, showcasing its superiority and potential as a novel method. Therefore, in practical applications, the mRMR-RFE-RF method has broad prospects, especially suitable for months with less data fluctuation, providing more accurate prediction results.

C. Discussions of Reliability of Selected Predictors

This article introduces a recursive elimination method for factor selection, utilizing MI and RF. The approach explores the distant correlation between CN05.1 grid precipitation data and

the Danjiangkou basin climate factors to discern and analyze crucial forecasting factors. The study encompasses 167 grid points within the basin, and each point constructs an individual forecasting factor selection model. The outcomes of high-frequency forecasting factor selection are outlined in Tables II and III. Specifically, Table II compiles results that account for the temporal lag effects of predictors and predictands, while Table III showcases results without considering such temporal lag effects.

In recent decades, an increasing number of researchers have turned their attention to the impact of large-scale climate factors on precipitation. From a climatological perspective, precipitation reflects changes in atmospheric heat sources, making it closely related to large-scale regional circulation and sea surface temperatures. It is well known that the most significant large-scale climate factor associated with sea surface temperatures is the El Niño-Southern Oscillation (ENSO). ENSO is a periodic climate phenomenon occurring near the equatorial Pacific, which can be divided into El Niño and La Niña. El Niño is characterized by warmer sea surface temperatures in the eastern Pacific cold tongue, while La Niña is characterized by cold temperatures in the eastern Pacific accompanied by low sea-level pressure in the western Pacific. The impact of ENSO on precipitation in China can be summarized as follows: it indirectly affects precipitation in China by influencing wind fields, moisture transport, and the trajectory of the western Pacific subtropical high. This ultimately affects the occurrence of droughts or floods. Therefore, this article will provide a simple physical attribution explanation and validation of the

TABLE II
HIGH-FREQUENCY PREDICTORS CONSIDERING THE LAG TIME

Month	Number	Index(Considering the lag time)	Selected times
January	1	Pre- April- North Atlantic Triple index	77
	2	Pre- July- Northern Hemisphere Polar Vortex Area Index	71
	3	Pre- August- West Pacific 850mb Trade Wind Index	67
	4	Pre- August- Cold- tongue ENSO Index	56
	5	Pre- September- West Wind Drift Current SST Index	48
February	1	Pre- February- North African -North Atlantic -North American Subtropical High Ridge Position Index	117
	2	Pre- January- Indian Subtropical High Ridge Position Index	79
	3	Pre- May- Tibet Plateau Region 2 Index	70
	4	Pre- June- Northern Hemisphere Polar Vortex Central Intensity Index	62
	5	Pre- May- Tibet Plateau Region 1 Index	53
March	1	Pre- January- Central Pacific 850mb Trade Wind Index	63
	2	Pre- September- Indian-Burma Trough Intensity Index	42
	3	Pre- November- Asian Meridional Ciculation Inddex	42
	4	Pre- November- Central Pacific 850mb Trade Wind Index	41
	5	Pre- March- Western North Pacific Typhoon number	38
April	1	Pre- January- Indian-Burma Trough Intensity Index	92
	2	Pre- January- Asia Polar Vortex Intensity Index	80
	3	Pre- January- East Atlantic Pattern, EA	75
	4	Pre- August- Mid-Eastern Pacific 200mb Zonal Wind Index	66
	5	Pre- March- Indian Subtropical High Ridge Position Index	51
May	1	Cur- February- Tibet Plateau Region 1 Index	80
	2	Pre- June- North American Polar Vortex Intensity Index	59
	3	Cur- February- Tibet Plateau Region 2 Index	58
	4	Pre- October- Northern Hemisphere Subtropical High Northern Boundary Position Index	54
	5	Pre- May- East Asian Trough Position Index	51
June	1	Pre- October- Northern Hemisphere Polar Vortex Intensity Index	64
	2	Pre- October- Scandinavia Pattern , SCA	38
	3	Cur- February- Antarctic Oscillation, AAO	35
	4	Pre- September- North American-North Atlantic Subtropical High Ridge Position Index	33
	5	Pre- June- North African-North Atlantic-North American Subtropical High Ridge Position Index	32
July	1	Cur- June- NINO Z SSTA Index	55
	2	Pre- July- Tropical Southern Atlantic SST Index	51
	3	Pre- October- East Atlantic Pattern, EA	43
	4	Cur- January- Atlantic Sub Tropical High Ridge Position Index	40
	5	Cur- April- NINO Z SSTA Index	33
August	1	Pre- December- Mid-Eastern Pacific 200mb Zonal Wind Index	50
	2	Pre- August- Scandinavia Pattern , SCA	42
	3	Pre- October- Asia Polar Vortex Intensity Index	41
	4	Pre- November- Kuroshio Current SST Index	41
	5	Cur- May- Pacific/ North American Pattern , PNA	38
September	1	Pre- December- Tibet Plateau Region-1 Index	72
	2	Pre- September- Asia Polar Vortex Area Index	66
	3	Cur- August- North American Subtropical High Northern Boundary Position Index	64
	4	Pre- September- South Indian Ocean Dipole Index	54
	5	Cur- February- Quasi-Biennial Oscillation Index	47
October	1	Pre- October- North American-Atlantic Subtropical High Northern Boundary Position Index	67
	2	Cur- March- East Atlantic Pattern, EA	45
	3	Pre- October- North African-North Atlantic-North American Subtropical High Northern Boundary Position Index	44
	4	Pre- November- 30hPa zonal wind Index	43
	5	Pre- November- East Atlantic-West Russia Pattern, EA/WR	42
November	1	Cur- August- North American Subtropical High Northern Boundary Position Index	97
	2	Pre- November- Eurasian Meridional Circulation Index	70
	3	Cur- March- Quasi-Biennial Oscillation Index	69
	4	Pre- November- East Atlantic-West Russia Pattern, EA/WR	67
	5	Pre- November- West Pacific Pattern, WP	64
December	1	Cur- June- Pacific Subtropical High Intensity Index	69
	2	Cur- March- South China Sea Subtropical High Ridge Position Index	67
	3	Cur- September- Equatorial Pacific 130E-80W Upper 300m temperature averaged anomaly index	60
	4	Cur- October- Atlantic Subtropical High Northern Boundary Position Index	55
	5	Cur- January- India-Burma Trough Intensity Index	49

relationship between the main forecasting factors selected in the article and precipitation in the study area based on various studies on the relationship between large-scale climate factors and precipitation. This aims to validate the rationality of the forecasting factor selection method.

1) Discussions on the Influence of the North Atlantic Triple Index: The results of the mRMR-RFE-RF forecasting factor selection method reveal that the North Atlantic Triple index (NATI) was selected 93 times for January grid precipitation data and 113 times for June grid precipitation data, without considering the short-term or long-term temporal lag effects between forecast factors and precipitation. Notably, the North Atlantic SSTI with an eight-month lag was selected a remarkable 77 times. These results strongly emphasize the substantial influence of the NATI on precipitation in the study area, highlighting its pivotal role in forecasting.

TABLE III
HIGH-FREQUENCY PREDICTORS WITHOUT CONSIDERING THE LAG TIME

Month	Number	Index(Without considering the lag time)	Selected times
January	1	West Wind Drift Current SST Index	200
	2	Cold-tongue ENSO Index	153
	3	West Pacific 850mb Trade Wind Index	125
	4	Atlantic-European Polar Vortex Area Index	110
	5	North Atlantic Triple index	93
February	1	Antarctic Oscillation, AAO	146
	2	Indian Subtropical High Ridge Position Index	125
	3	Northern Hemisphere Polar Vortex Central Intensity Index	102
	4	Tibet Plateau Region 2 Index	95
	5	North African-North Atlantic-North American Subtropical High Ridge Position Index	93
March	1	Central Pacific 850mb Trade Wind Index	122
	2	East Asian Trough Position Index	88
	3	Scandinavian Teleconnection Pattern	71
	4	Indian-Burma Trough Intensity Index	65
	5	Western Pacific Subtropical High Ridge Position Index	65
April	1	Asia Polar Vortex Intensity Index	123
	2	Indian-Burma Trough Intensity Index	108
	3	East Atlantic Pattern, EA	105
	4	Northern Hemisphere Polar Vortex Area Index	88
	5	Western Pacific Warm Pool Strength index	78
May	1	Tibet Plateau Region 2 Index	101
	2	Tibet Plateau Region 1 Index	96
	3	East Asian Trough Position Index	93
	4	Northern Hemisphere Subtropical High Northern Boundary Position Index	74
	5	Eurasian Meridional Circulation Index	73
June	1	North Atlantic Triple index	113
	2	Antarctic Oscillation, AAO	102
	3	Northern Hemisphere Polar Vortex Intensity Index	101
	4	Tropical Pacific Outgoing Long Wave Radiation Index	96
	5	Western Pacific Warm Pool Strength index	88
July	1	NINO Z SSTA Index	138
	2	Western Pacific Warm Pool Strength index	88
	3	East Atlantic Pattern, EA	75
	4	North Atlantic Oscillation , NAO	72
	5	Cold Air Activity Index	65
August	1	Mid-Eastern Pacific 200mb Zonal Wind Index	85
	2	Kuroshio Current SST Index	83
	3	NINO B SSTA Index	70
	4	Equatorial Pacific 180-100W Upper 300m temperature Average anomaly index	68
	5	Scandinavia Pattern , SCA	67
September	1	Tibet Plateau Region-1 Index	112
	2	Asia Polar Vortex Area Index	91
	3	Asian Zonal Circulation Index	90
	4	Eurasian Zonal Circulation Index	82
	5	Quasi-Biennial Oscillation Index	80
October	1	NINO B SSTA Index	102
	2	50hPa zonal wind Index	95
	3	East Atlantic Pattern, EA	90
	4	Western Pacific Warm Pool Strength index	86
	5	North American-Atlantic Subtropical High Northern Boundary Position Index	83
November	1	Eastern Pacific Subtropical High Ridge Position Index	157
	2	Quasi-Biennial Oscillation Index	127
	3	Cold-tongue ENSO Index	105
	4	North American Subtropical High Northern Boundary Position Index	105
	5	Solar Flux Index	97
December	1	South China Sea Subtropical High Ridge Position Index	122
	2	Equatorial Pacific 130E-80W Upper 300m temperature averaged anomaly index	119
	3	Atlantic Subtropical High Northern Boundary Position Index	107
	4	India-Burma Trough Intensity Index	103
	5	South Indian Ocean Dipole Index	102

The NATI represents a key mode of interannual variability in North Atlantic sea surface temperatures. The intricate relationships between the variability of NATI and ENSO have been extensively demonstrated in multiple research endeavors [38], [39]. For example, Ham et al. [40] highlighted that the presence of positive sea surface temperature anomalies (SSTAs) in the North Tropical Atlantic during spring could trigger unusual lower level easterly winds over the tropical western Pacific, thereby promoting the occurrence of winter La Niña events. Another mechanism involves the stimulation of midlatitude atmospheric teleconnections (Wang et al., 2011, 2013; Yan et al., 2018) [41], [42]. Gu et al. [43] identified that during early winter, North Atlantic SST anomalies exhibit a tripolar pattern of "+-+" ("-+-"), corresponding to above-normal (below-normal) summer precipitation in the middle and lower reaches of the Yangtze River. Through observations and numerical experiments, Zuo et al. [44] revealed that when the anomalous phase of North Atlantic SST displays a meridional tripole pattern of "+-+", it triggers a remote teleconnection wave train emanating from high latitudes in Eurasia. This induces a positive anomaly in atmospheric potential height over the Ural Mountains, leading to excessive precipitation in the Yangtze River Basin. Conversely, when the phase is "-+-", the Ural Mountains blocking high-pressure system is suppressed, resulting in reduced precipitation in the middle and lower reaches of the Yangtze River.

2) Discussions on the Influence of the West Wind Drift Current SST Index: The results of the mRMR-RFE-RF forecasting factor selection method indicate that the West Wind Drift Current SST Index is the most frequently chosen factor in the January grid precipitation data across 167 grid points. Without considering the lag time between predictors and precipitation, the West Wind Drift Current SST Index was selected 200 times. The index from the previous September was chosen the most, totaling 48 selections. Yu et al. [45] discovered that, for the turning-type tropical cyclones (TCs), the presence of higher-than-average sea surface temperature (SST) anomalies in the West Wind Drift region from May to July leads to the formation of positive vorticity anomalies. These anomalies are observed extending from the northern Philippines to Taiwan over the period from July to September. Bulgin et al. [46] focused on global sea-surface temperature anomaly trends, including the SST trends in the West Wind Drift Current area. It provides a comprehensive analysis of SST anomalies' trends, variability, and persistence over the last four decades. Li et al. [47] mentioned that the correlation of various conducive weather patterns with several SST indices, including the West Wind Drift Current SST Index.

3) Discussions on the Influence of the Cold-Tongue ENSO Index: The study found that without considering the lag effect between forecast factors and precipitation, the Cold Tongue ENSO Index was selected 153 times in January grid precipitation data, and when the lag time was four months, the index was chosen 56 times. In November, without considering the lag effect, it was selected 157 times, and with a two-month lag, it was chosen 97 times.

Kug et al. [48] classified El Niño events based on the maximum warming area's location, naming those with a warming core area in the eastern to central Pacific as "Cold Tongue El Niño." Wang et al. [49] revealed distinct patterns associated with different types of El Niño events. For the eastern type, there was a noticeable OKJ wave train (from the Sea of Okhotsk - east of Japan - west of the International Date Line in the northern subtropical high) in the northern hemisphere's mid- to high-latitudes. This pattern led to excessive precipitation south of the Yangtze River due to an abnormal high-pressure block over the Sea of Okhotsk and a southward and westward shift of the Northwest Pacific subtropical high, facilitating moisture transport. Conversely, the central type El Niño caused an abnormal low pressure in the Sea of Okhotsk and an eastward shift of the Northwest Pacific subtropical high, resulting in no significant precipitation anomalies in China south of the Yangtze River, but increased moisture transport to North China. Karori et al. [50] demonstrated that in the boreal summer subsequent to the peak phase of a Cold Tongue El Niño event, an increased precipitation anomaly is observed in the Yangtze River. Simultaneously, decreased precipitation anomalies are seen in both the southern and northern regions of China. These findings underscore the significant impact of the Cold Tongue ENSO Index on precipitation.

4) Discussions on the Influence of the Antarctic Oscillation: The results of this study's forecast factor selection method show that the Antarctic Oscillation index was chosen 146 times in the February grid precipitation data without considering the length of the lag effect between forecast factors and precipitation. When the lag time was one year, the Antarctic Oscillation index was selected 117 times, indicating its significant impact on precipitation in the study area.

Numerous studies have demonstrated that the Antarctic Oscillation influences summer precipitation in eastern China through atmospheric teleconnection mechanisms [51], [52], [53]. Some researchers discovered that early Antarctic Oscillation influences the strength and position of the Northwest Pacific subtropical high in the later period, causing deviations in the position of the East Asian summer monsoon rain belt and thereby affecting precipitation in the Yangtze River Basin. When the early Antarctic Oscillation is high, the subsequent Western Pacific subtropical high tends to be stronger and more southward, leading to more precipitation in the Yangtze River Basin, and vice versa.

5) Discussions on the Influence of the NINO B and NINO Z SSTA Index: The study's method for selecting forecast factors shows that the NINO Z sea temperature index in July grid precipitation data was chosen 138 times without considering the duration of the lag effect between the forecast factor and precipitation. When considering a one-month lag, the NINO Z index was chosen 55 times. Meanwhile, the West Pacific Subtropical High Ridge Line Index with a three-month lag was selected 33 times. For the NINO B sea surface temperature anomaly index in grid precipitation data, it was chosen 70 times in August and 102 times in October, indicating the significant influence of NINO Z and NINO B sea temperature indices on regional precipitation.

The NINO Z sea temperature index is derived from the area-weighted average of Nino1+2 (0°–10°S, 90°–80°W), Nino3 (5°N–5°S, 150°–190°W), and Nino4 (5°N–5°S, 160°–150°W), representing the central and eastern Pacific Ocean's surface temperature. Influenced by El Niño, the frequency and intensity of precipitation in the Yangtze River Basin increase during summer [54]. Additionally, in the year following a strong El Niño, increased heavy precipitation poses serious risks to economic and property safety [55].

V. CONCLUSION

In the intricate field of medium- and long-term hydrological forecasting, precision and sophisticated modeling are important. This study introduces a refined method predictor selection, showcasing the mRMR-RFE-RF method. This technique elegantly combines MI and RF, offering a sophisticated yet streamlined mechanism for refining the predictor set to its most essential elements. Our validation using established ML models demonstrates not only the robustness of the mRMR-RFE-RF method but also its superiority over conventional single-predictor screening techniques.

The mRMR-RFE-RF method has been shown to significantly bolster the precision and efficiency of precipitation forecasting across a spectrum of ML models, highlighting its broad applicability. For instance, in January, the XGB model, when informed by the mRMR-RFE-RF method, exhibited a MAPE of approximately 0.57, which is notably 11.4% lower than that of the MI method and 3.6% lower than the RF method alone. This trend of enhanced performance is echoed in the month of July, where the MAPE for the RF model, under the guidance of mRMR-RFE-RF, was approximately 0.32, outperforming the MI and RF methods by a significant margin of 20.9% and 19.4%, respectively.

Our models have demonstrated exceptional forecasting proficiency for the months of July and August, a result that may be attributed to the ACC evaluation index's particular sensitivity to varying levels of precipitation. Following the application of the mRMR-RFE-RF selector, we have observed a consistent performance in both RF and XGB models, contrasting with the more erratic performance of the LGB model across interannual and intermonthly dimensions. For example, in August, the ACC value for the mRMR-RFE-RF in the XGB model hovered around 0.3, while other methods only managed to achieve around 0.2.

The efficacy of the mRMR-RFE-RF method is further corroborated through the analysis of a range of indices, including the North Atlantic Triple, West Wind Drift Current SST, Cold-tongue ENSO, Antarctic Oscillation, and NINO B and Z SSTA, which affirms its rationality in the selection of pertinent factors. The NATI, for example, was selected 93 times for January grid precipitation data and 113 times for June grid precipitation data, underscoring its considerable influence on the study area's precipitation.

This article elucidates a pioneering method to predictor-screening in hydrological forecasting, underscored by a case study within the Danjiangkou basin. By augmenting the accuracy of precipitation forecasting, this method paves the way for innovative practices in the selection of predictors for medium- and long-term hydrological forecasting. The comprehensive analysis confirms the method's efficacy and trustworthiness, offering a robust instrument for precise precipitation forecasting across a variety of hydrological scenarios.

REFERENCES

[1] J. Chadalawada, H. Herath, and V. Babovic, "Hydrologically informed machine learning for rainfall-runoff modeling: A genetic programming-based toolkit for automatic model induction," *Water Resour. Res.*, vol. 56, no. 4, pp. 1–23, Mar. 2020.

[2] T. Tang, Z. Liang, Y. Hu, B. Li, and J. Wang, "Research on flood forecasting based on flood hydrograph generalization and random forest in Qiushui river basin," *J. Hydroinformat.*, vol. 22, no. 6, pp. 1588–1602, Aug. 2020.

[3] C. Shen, "A transdisciplinary review of deep learning research and its relevance for water resources scientists," *Water Resour. Res.*, vol. 54, no. 11, pp. 8558–8593, Aug. 2018.

[4] R. Reinoso-Rondinel, M. Rempel, M. Schultze, and S. Trömel, "Nationwide radar-based precipitation nowcasting—A localization filtering approach and its application for Germany," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 1670–1691, 2022.

[5] Q. Shen et al., "Soil moisture retrieval from multipolarization SAR data and potential hydrological application," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 6531–6544, 2023.

[6] H. Qiao et al., "A novel method for snow depth retrieval using improved dense medium RVoG model," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 1874–1884, 2023.

[7] H. Tyralis, G. Papacharalampous, N. Doulamis, and A. Doulamis, "Merging satellite and gauge-measured precipitation using LightGBM with an emphasis on extreme quantiles," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 6969–6979, 2023.

[8] W. Wang, H. Lu, T. Zhao, L. Jiang, and J. Shi, "Evaluation and comparison of daily rainfall from latest GPM and TRMM products over the Mekong River basin," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 6, pp. 2540–2549, Jun. 2017.

[9] C. Bai, F. Sun, J. Zhang, Y. Song, and S. Chen, "Rainformer: Features extraction balanced network for radar-based precipitation nowcasting," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 4023305.

[10] F. Kratzert, D. Klotz, and M. Herrnegger, A. K. Sampson, S. Hochreiter, and G. S. Nearing, "Toward improved predictions in ungauged basins: Exploiting the power of machine learning," *Water Resour. Res.*, vol. 55, no. 12, pp. 11344–11354, Nov. 2019.

[11] H. Chen, V. Chandrasekar, R. Cifelli, and P. Xie, "A machine learning system for precipitation estimation using satellite and ground radar network observations," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 2, pp. 982–994, Feb. 2020.

[12] C. Bai, D. Zhao, M. Zhang, and J. Zhang, "Multimodal information fusion for weather systems and clouds identification from satellite images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 7333–7345, 2022.

[13] F. L. R. Torres, L. M. M. Lima, M. S. Reboita, A. R. de Queiroz, and J. W. M. Lima, "Integrating hydrological and machine learning models for enhanced streamflow forecasting via Bayesian model averaging in a hydro-dominant power system," *Water*, vol. 16, no. 4, Feb. 2024, Art. no. 586.

[14] E. Rozos, "Assessing hydrological simulations with machine learning and statistical models," *Hydrology*, vol. 10, Feb. 2023, Art. no. 49.

[15] L. J. Slater et al., "Hybrid forecasting: Blending climate predictions with AI models," *Hydrol. Earth Syst. Sci.*, vol. 27, no. 9, pp. 1865–1889, May 2023.

[16] E.-I. Koutsovili, O. Tzoraki, N. Theodossiou, and G. E. Tsekouras, "Early flood monitoring and forecasting system using a hybrid machine learning-based approach," *ISPRS Int. J. Geo- Inf.*, vol. 12, no. 11, Nov. 2023, Art. no. 464.

[17] T. Tang, D. L. Jiao, T. Chen, and G. Gui, "Medium and long-term precipitation forecasting method based on data augmentation and machine learning algorithms," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 1000–1011, 2022.

[18] X. He, H. Guan, and J. Qin, "A hybrid wavelet neural network model with mutual information and particle swarm optimization for forecasting monthly rainfall," *J. Hydrol.*, vol. 527, pp. 88–100, Aug. 2015.

[19] J. G. Bowden, G. C. Dandy, and H. R. Maier, "Input determination for neural network models in water resources applications. Part 1-background and methodology," *J. Hydrol.*, vol. 301, no. 1–4, pp. 75–92, Jan. 2005.

[20] L. Schmidt, F. Heße, S. Attinger, and R. Kumar, "Challenges in applying machine learning models for hydrological inference: A case study for flooding events across Germany," *Water Resour. Res.*, vol. 56, no. 5, pp. 1–10, May 2020.

[21] S. S. Karimi, N. Saintilan, L. Wen, and R. Valavi, "Application of machine learning to model wetland inundation patterns across a large semiarid floodplain," *Water Resour. Res.*, vol. 55, no. 11, pp. 8765–8778, Nov. 2019.

[22] M. Sebban and R. Nock, "A hybrid filter/wrapper approach of feature selection using information theory," *Pattern Recognit.*, vol. 35, no. 4, pp. 835–846, Apr. 2002.

[23] K. Yu, H. Li, W. Xie, and L. Wang, "A hybrid feature-selection method based on mRMR and binary differential evolution for gene selection," *Processes*, vol. 12, no. 2, Feb. 2024, Art. no. 313.

[24] P. Yang, H. Huang, and C. Liu, "Feature selection revisited in the single-cell era," *Genome Biol.*, vol. 22, Nov. 2021, Art. no. 321.

[25] N. Pudjihartono, T. Fadason, A. W. Kempa-Liehr, and J. M. O'Sullivan, "A review of feature selection methods for machine learning-based disease risk prediction," *Front. Bioinf.*, vol. 2, Jun. 2022, Art. no. 927312.

[26] A. Abdo, R. Mostafa, and L. Abdel-Hamid, "An optimized hybrid approach for feature selection based on chi-square and particle swarm optimization algorithms," *Data*, vol. 9, no. 2, Jan. 2024, Art. no. 20.

[27] Z. Liu, Q. Li, J. Zhou, W. Jiao, and X. Wang, "Runoff prediction using a novel hybrid ANFIS model based on variable screening," *Water Resour. Manage.*, vol. 35, pp. 2921–2940, Jun. 2021.

[28] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Mach. Learn.*, vol. 46, no. 1-3, pp. 389–422, Jan. 2002.

[29] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001.

[30] T. Tang, D. Jiao, and G. Gui, "Medium- and long-term hydrological forecasting using mutual information and random forest predictors selector," in *Proc. IEEE 23rd Int. Conf. Commun. Technol.*, Wuxi, China, Oct. 2023, pp. 1428–1433.

[31] J. Wu, X. J. Gao, Y. L. Xu, and J. Pan, "Numerical simulation and uncertainty analysis of regional climate change in East Asia and Southeast Asia," *Chin. Acad. Meteorological Sci.*, vol. 8, no. 3, pp. 147–152, Aug. 2015.

[32] X. Ying, X. Gao, Y. Shen, C. Xu, Y. Shi, and F. Giorgi, "A daily temperature dataset over China and its application in validating a RCM simulation," *Adv. Atmospheric Sci.*, vol. 4, no. 26, pp. 153–162, Jul. 2009.

[33] T. Ho, "The random subspace method for constructing decision forests," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 8, pp. 832–844, Aug. 1998.

[34] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf.*, 2016, pp. 785–794.

[35] G. Ke et al., "LightGBM: A highly efficient gradient boosting decision tree," in *Proc.: 31st Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 3149–3157.

[36] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1226–1238, Aug. 2005.

[37] R. Kohavi and G. John, "Wrappers for feature subset selection," *Artif. Intell.*, vol. 97, no. 1–2, pp. 273–324, Dec. 1997.

[38] I. Richter, H. Tokinaga, Y. Kosaka, T. Doi, and T. Kataoka, "Revisiting the tropical atlantic influence on El Niño-southern oscillation," *J. Climate*, vol. 34, no. 21, pp. 8533–8548, Nov. 2021.

[39] F. Jiang and W. J. Zhang, "Understanding the complicated relationship between ENSO and wintertime north tropical atlantic SST variability," *Geophysical Res. Lett.*, vol. 49, no. 5, Feb. 2022, Art. no. e2022GL097889.

[40] Y. G. Ham, J. S. Kug, J. Y. Park, and F. F. Jin, "Sea surface temperature in the north tropical atlantic as a trigger for El Niño/southern oscillation events," *Nature Geosci.*, vol. 6, no. 2, pp. 112–116, Jan. 2013.

[41] X. Wang, C. Z. Wang, W. Zhou, D. X. Wang, and J. Song, "Teleconnected influence of north Atlantic sea surface temperature on the El Niño onset," *Climate Dyn.*, vol. 37, no. 3–4, pp. 663–676, May 2011.

[42] X. Yan, J. Z. Ren, J. H. Ju, and S. Yang, "Influence of springtime Atlantic SST on ENSO: Role of the Madden-Julian oscillation," *J. Meteorological Res.*, vol. 32, no. 3, pp. 380–393, Jul. 2018.

[43] W. Gu, C. Li, X. Wang, W. Zhou, and W. Li, "Linkage between Mei-Yu precipitation and north Atlantic SST on the decadal timescale," *Adv. Atmospheric Sci.*, vol. 26, no. 1, pp. 101–108, Jan. 2009.

[44] J. Zuo, W. Li, C. Sun, L. Xu, and H. Ren, "Impact of the north Atlantic sea surface temperature tripole on the East Asian summer monsoon," *Adv. Atmospheric Sci.*, vol. 30, no. 4, pp. 1173–1186, Jun. 2013.

[45] Y. Yu et al., "Modulation of Pacific sea surface temperature on two types of tropical cyclone tracks affecting Northeast China," *Front. Earth Sci.*, vol. 10, May 2022, Art. no. 810943. [Online]. Available: https://www.frontiersin.org/articles/10.3389/feart.2022.810943/full

[46] C. E. Bulgin, C. J. Merchant, and D. Ferreira, "Tendencies, variability and persistence of sea surface temperature anomalies," *Sci. Rep.*, vol. 10, 2020, Art. no. 7986, doi: 10.1038/s41598-020-64785-9.

[47] J. Li et al., "Winter particulate pollution severity in North China driven by atmospheric teleconnections," *Nature Geosci.*, vol. 15, pp. 349–355, Apr. 2022. [Online]. Available: https://www.nature.com/articles/s41561-022-00933-2

[48] J. S. Kug, F.-F. Jin, and S.-I. An, "Two types of El Niño events: Cold tongue El Niño and warm pool El Niño," *J. Climate*, vol. 22, no. 6, pp. 1499–1515, Mar. 2009.

[49] Y. Wang, A. R. Lupo, and J. Qin, "A response in the ENSO cycle to an extratropical forcing mechanism during the El Niño to La Niña transition," *Tellus A*, vol. 65, no. 1, Dec. 2013, Art. no. 22431.

[50] M. A. Karori, J. Li, and F. Jin, "The asymmetric influence of the two types of El Niño and La Niña on summer rainfall over Southeast China," *J. Climate*, vol. 26, no. 13, pp. 4567–4582, Jul. 2013.

[51] H. J. Wang and K. Fan, "Central-north China precipitation as reconstructed from the Qing dynasty: Signal of the antarctic atmospheric oscillation," *Geophysical Res. Lett.*, vol. 32, Dec. 2005, Art. no. L24705.

[52] K. Fan, "Atmospheric circulation in southern hemisphere and summer rainfall over Yangtze River valley," *Chin. J. Geophys.*, vol. 49, pp. 599–606, May 2006.

[53] K. Fan and H. J. Wang, "Antarctic oscillation and prediction on summer rainfall over eastern China in 2006 (in Chinese)," *J. Appl. Meteorological Sci.*, vol. 17, pp. 383–384, Jan. 2006.

[54] Y. Zhao and H. Zhang, "Impacts of SST warming in tropical Indian Ocean on CMIP5 model-projected summer rainfall changes over Central Asia," *Climate Dyn.*, vol. 46, pp. 3223–3238, Aug. 2015.

[55] X. Yang et al., "Key to ENSO phase-locking simulation: Effects of sea surface temperature diurnal amplitude," *NPJ Climate Atmospheric Sci.*, vol. 6, Sep. 2023, Art. no. 159.

Tiantian Tang (Member, IEEE) received the Dr. Eng. degree in hydrology and water resources from Hohai University, Nanjing, China, in 2021.

Since 2021, she has been a Postdoctoral Fellow with the School of Geographic and Biologic Information, Nanjing University of Posts and Telecommunications, Nanjing. She has published more than ten peer-reviewed international papers and led a special foundation of The National Key Laboratory of Water Disaster Prevention. Her recent research interests include medium- and long-term hydrological forecasting, machine learning, deep learning, and flood forecasting.

Dr. Tang was granted the Jiangsu Provincial Science and Technology Deputy Project in 2022.

Tao Chen received the Dr. Eng. degree in hydrology and water resources from Hohai University, Nanjing, China, in 2020.

Since 2023, he has been a Senior Engineer with the Hydrology and Water Resources Department, Nanjing Hydraulic Research Institute, Nanjing. His recent research interests include multisource precipitation fusion and hydrology simulation.

Guan Gui (Fellow, IEEE) received the Ph.D. degree in information and communication engineering from the University of Electronic Science and Technology of China, Chengdu, China, in 2012.

From 2009 to 2014, he was with Tohoku University as a Research Assistant and a Postdoctoral Research Fellow. From 2014 to 2015, he was an Assistant Professor with the Akita Prefectural University, Akita, Japan. Since 2015, he has been a Professor with Nanjing University of Posts and Telecommunications, Nanjing, China. His recent research interests include intelligence sensing and recognition, intelligent signal processing, and physical layer security.

Dr. Gui has published over 200 IEEE Journal/Conference papers and won several best paper awards, e.g., ICC 2017, ICC 2014, and VTC 2014-Spring. His contributions to intelligent signal analysis and wireless resource optimization have earned him the title of Fellow of the IEEE, IET, and AAIA. He received the IEEE Communications Society Heinrich Hertz Award in 2021, top 2% scientists of the World by Stanford University in 2021–2023, the Clarivate Analytics Highly Cited Researcher in Cross-Field in 2021–2023, the Highly Cited Chinese Researchers by Elsevier in 2020–2023, the Member and Global Activities Contributions Award in 2018, the Top Editor Award of IEEE Transactions on Vehicular Technology in 2019, the Outstanding Journal Service Award of KSII Transactions on Internet and Information System in 2020, the Exemplary Reviewer Award of IEEE Communications Letters in 2017, the 2012 Japan Society for Promotion of Science (JSPS) Postdoctoral Fellowships for Foreign Researchers, and the 2018 Japan Society for Promotion of Science (JSPS) International Fellowships for Overseas Researchers. He was also selected as the Jiangsu Specially Appointed Professor in 2016, the Jiangsu High-level Innovation and Entrepreneurial Talent in 2016, and the Jiangsu Six Top Talent in 2018. Since 2022, he has been a Distinguished Lecturer of the IEEE Vehicular Technology Society. He is serving or served on the editorial boards of several journals, including IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, *IEICE Transactions on Communications*, *Physical Communication*, *Wireless Networks*, *IEEE Access*, *Journal of Circuits Systems and Computers, Security and Communication Networks*, *IEICE Communications Express,* and *KSII Transactions on Internet and Information Systems*, *Journal on Communications*. In addition, he served as the IEEE VTS Ad Hoc Committee Member in AI Wireless, General Cochair of EAI MOBIMEDIA 2024, General Cochair of ICN2024, General Cochair of GAIIS 2024, TPC Chair of IoTML 2024, TPC Chair of FAIML 2024, Executive Chair of IEEE ICCT 2023, Workshop Chair of LANTINCOM2023, TPC Chair of PRAI 2022, TPC Chair of ICGIP 2022, Executive Chair of VTC 2021-Fall, Vice Chair of WCNC 2021, TPC Chair of PHM 2021, Symposium Chair of WCSP 2021, General Cochair of Mobimedia 2020, TPC Chair of WiMob 2020, Track Chairs of EuCNC 2021 and 2022, VTC 2020 Spring, Award Chair of PIMRC 2019, and TPC member of many IEEE international conferences, including GLOBECOM, ICC, WCNC, PIRMC, VTC, and SPAWC.