Machine learning-based predictive analysis of ENSO response to temperature and precipitation in southwestern mountains

Muchen Yao [1], Zhenxin Fan [2, *], Yu Chen [1], Hongmei Gou [1]

[1] Zhaotong Agrometeorological Experimental Station, Zhaotong, China, 657000

[2] Weixin County Meteorological Bureau, Zhaotong, China, 657000

* Corresponding Author Email: fanzxx@163.com

Abstract. Accurately predicting the response of ENSO events to temperature and precipitation is of great scientific significance and application value. This study takes Zhaotong City as an example, based on the month-by-month temperature and precipitation observation data of Zhaotong City as well as the Oceanic Nino Index (ONI) from 1961-2023, constructs lagged features, and applies the Random Forest (RF) and XGBoost machine learning models to analyze the correlation between ENSO and climate, and carry out prediction studies. The results show that the RF model performs better in temperature prediction, with a mean absolute error (MAE) of 1.67°C, compared with 1.70°C for the XGBoost model; the precipitation prediction error is significantly higher than that of temperature, and the model's prediction values of extreme high temperature and heavy precipitation events are generally low, and there is a phenomenon of lagging or advancing of the rainy season phase. The study reveals the impact of ENSO on the climate of southwest mountains, indicating the potential of machine learning methods in long-term climate prediction, which can provide important technical support for regional prediction and disaster prevention and mitigation, but the precipitation prediction needs to be further combined with local environmental factors to improve the accuracy.

Keywords: ENSO, Temperature Prediction, Precipitation Prediction, Random Forest, XGBoost.

1. Introduction

The complex topography of Southwest China has resulted in diverse climate systems and ecological patterns [1]. Zhaotong City is located in the northeastern part of Yunnan Province, situated in the transition zone from the Sichuan Basin to the Yunnan-Guizhou Plateau, and the terrain is dominated by plateaus and mountains, with mountains and large differences in elevation within the city, forming a complex geomorphological structure. Due to the high mountains and deep valleys, gullies and rivers in Zhaotong's territory, the climate is diverse and complex, not only with the monsoon climate characteristics of distinct dry and rainy seasons, rain and heat in the same season, dry and cold in the same season, but also with a variety of climatic types such as sub-tropical, mid-sub-tropical, north-sub-tropical, south temperate, mid-temperate and north-temperate, etc., and this climatic diversity has provided a rich source of resources for the development of Zhaotong's agriculture. In the southwestern mountainous areas, the impact of ENSO shows significant regional variability, which is also associated with the current frequent occurrence of extreme weather. Studies have shown that El Niño events may lead to less precipitation in southwest China, while La Niña events may bring more precipitation [2]. In the long term, precipitation in southwest China shows a fluctuating upward trend, which is related to the superimposed effects of global climate change and ENSO events [3]. In the context of global warming, extreme warm events in Southwest China increased significantly from 1961 to 2017, while extreme cold events decreased significantly [4]. In addition, ENSO events have exacerbated the drought sensitivity in Southwest China [5]. However, due to the complexity of topography and circulation, the specific influence mechanisms and prediction methods of ENSO events on temperature and precipitation in Zhaotong City still need to be explored in depth.

Compared with machine learning methods, traditional statistical methods usually have a lower demand for computational resources, and in short-term prediction, the performance is more stable, but the adaptability to data changes is weaker, and its performance in long-term prediction is usually

inferior to that of machine learning methods, and it is difficult to capture complex nonlinear patterns [6-8]. Yang et al. [9] used correlation test, wavelet analysis and other methods to study the change characteristics of precipitation and temperature and their relationship with ENSO in Qinghai Province, and Li et al. [10] used lag response and other methods to study the correlation between climate change in the middle and lower reaches of the Yangtze River and ENSO, but there are still deficiencies in the use of machine learning to study the impact of ENSO events on Zhaotong and its prediction methods for extreme weather events.

Therefore, this paper selects ONI index, month-by-month temperature and precipitation data of Zhaotong national meteorological station from 1961 to 2023, and applies two modeling algorithms of Random Forest and XGBoost to carry out experiments to explore the correlation of ENSO on temperature and precipitation in Zhaotong City, and to predict temperature and precipitation in Zhaotong City, with a view to providing technical references to the climate prediction in Southwest China.

2. Materials and methods

2.1. Data acquisition and processing

2.1.1. Data sources and processing

The data used in this paper includ: (1) The month-by-month average temperature and cumulative precipitation meteorological observations of Zhaotong National Station from 1961 to 2023 provided by Zhaotong Meteorological Bureau; (2) The Oceanic Niño Index (ONI) published by the National Oceanic and Atmospheric Administration (NOAA) of the United States was selected as a quantitative indicator of ENSO events, which is publicly available through the official website of CPC (Climate Prediction Center) (https://www.cpc.ncep.noaa.gov). ONI is defined as the 3-month sliding average of sea surface temperature (SST) in the Nino 3.4 region (5°N-5°S, 120°-170°W) of the equatorial Pacific Ocean.

To ensure the continuity and reliability of the temperature and precipitation data, the linear regression method of neighboring stations was used to interpolate individual missing data, and the Z-Score Method was used to monitor and process the outliers in the dataset, and the threshold was adjusted to 3 to minimize the misclassification.

2.1.2. Data set segmentation

The dataset of 63 years totaling 756 months was divided into training and test sets with September 2017 as the split point. Lagged features were constructed from feature engineering to predict future temperature and precipitation using ONI index 1-3 months in advance, so the training set started from April 1961, i.e., the month-by-month data from April 1961-September 2017 was used as the training set, and the month-by-month data from October 2017-December 2023 was used as the test set, and a total of 678 sets of data were constructed as the training set and 75 sets of data for the test set. With temperature and precipitation as dependent variables and ONI index as independent variable, the temperature and precipitation prediction model was built, and the training and test data sets were input to train and test the model.

2.2. Research methodology

2.2.1. Random forest

Random Forest Algorithm is a classification tree-based algorithm with high speed and accuracy in dealing with classification and regression problems. The self-help method (Bootstrap) resampling strategy under the integrated learning framework is used to construct a collaborative multi-decision tree prediction model based on the Bagging algorithm. The method uses bootstrap resampling technique to randomly select n samples from the original samples, construct n decision trees, and arrive at the final prediction result by combining multiple decision tree predictions. The steps to generate a random forest are as follows:

(1) A new self-help sample set is randomly drawn from the original training data set by applying the bootstrap method, and a classification regression tree is constructed from it, with the unsampled samples forming an out-of-bag (OOB) each time.

(2) Given n features, mtry features (mtry ≤ n) are randomly selected at each node of each tree, and by calculating the amount of information embedded in each feature, one of the mtry features with the most classification ability is selected for node splitting.

(3) Each tree is maximized with no clipping.

(4) The generated multiple trees are formed into a random forest, which is used to classify the new data, and the classification result is based on the number of votes of the tree classifiers [11].

2.2.2. XGBoost model

XGBoost is a Boosting algorithm, the core idea of this algorithm is to set multiple weak classifiers into one strong classifier. It provides parallel tree boosting to solve the problem of categorizing data quickly and accurately. The binary classification XGBoost model function expression mainly consists of tree model, predicted probability, loss function, and regularization term [12].

(1) The model function expression is:

$$\hat{y}_i = \Sigma f_k(x_i) \tag{1}$$

Where $\hat{y}_i$ is the predicted value of sample x_i and f_k is the weight of the output leaf node of the k th decision tree.

(2) The expression for the predicted probability ($\hat{p}_i$) is:

$$\hat{p}_i = \frac{1}{1+\mathrm{e}^{-\hat{y}_i}} \tag{2}$$

(3) The expression of the loss function ($L(\phi)$) is:

$$L(\phi) = \Sigma l(y_i, \hat{y}_i) + \Sigma \Omega(f_k) \tag{3}$$

Where $l(y_i, \hat{y}_i)$ is the loss function for sample i, which measures the difference between the model prediction and the actual label. $\hat{y}_i$ is the predicted value of the model for sample i. $\Omega(f_k)$ is the regularization term of the tree model, which is used to control the model complexity and prevent overfitting.

(4) The expression for the regularization term ($\Omega(f)$) is:

$$\Omega(f) = \gamma T + \frac{1}{2n}\lambda |w|^2 \tag{4}$$

Where γ and λ are hyperparameters that regulate regularization and are the number of trees. The regularization term consists of two parts: γT is used to control the complexity of the tree, and $\frac{1}{2n}\lambda |w|^2$ is used to penalize the weights of leaf nodes to prevent overfitting. By minimizing the loss term and controlling the regularization term in the objective function, the XGBoost model can achieve the dual goals of fitting the training data and controlling the model complexity simultaneously.

2.3. Model evaluation

This study uses mean absolute error (MAE) as a model evaluation metric. The mean absolute error is the mean of the absolute value of the difference between the actual value and the predicted value, and indicates the average magnitude of error in the predicted value.

The calculation formula is:

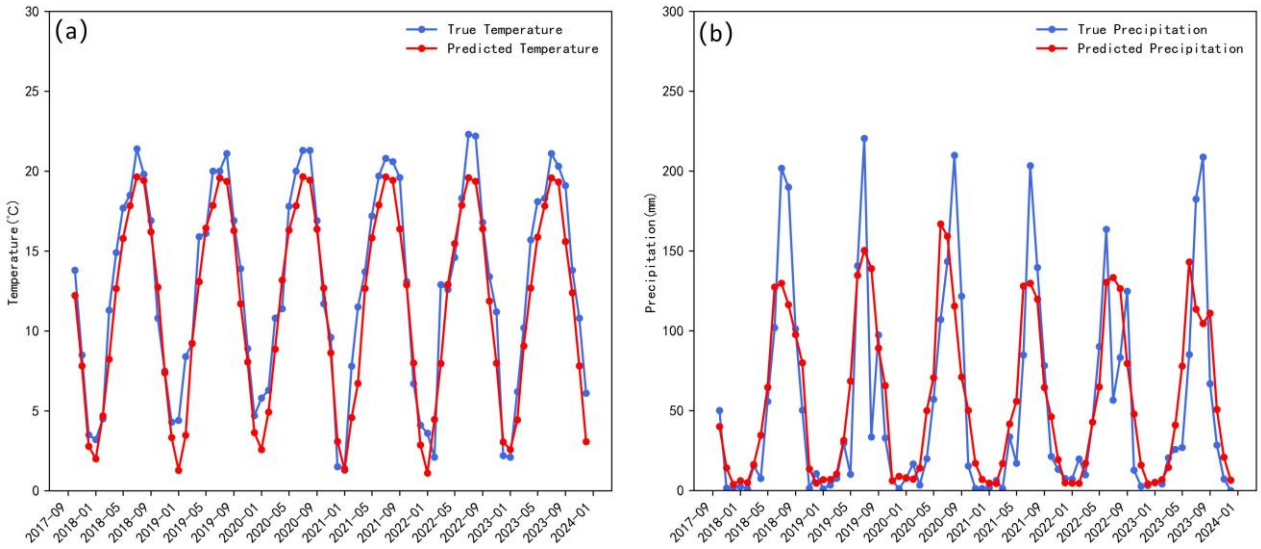
$$MAE = \frac{\Sigma_{i=1}^{n} | Y_i - \tilde{Y}_i |}{n} \quad (5)$$

Where $\tilde{Y}_i$ is the measured value, Y_i is the predicted value, and n is the number of samples. The mean absolute error does not need to consider the direction of the error. Compared to the mean square error, the mean absolute error is less sensitive to outliers, and has a stable gradient for any size of difference, no matter for what kind of input value, which does not lead to the gradient explosion problem and has a more stable solution.

3. Results and analysis

3.1. Construction and analysis of RF-based temperature and precipitation forecasting models

The above dataset division results are used as inputs to the RF model to predict the monthly average temperature and cumulative precipitation from October 2017 to December 2023 (Figure 1). From the analysis of Figure 1, it can be seen that in the performance of temperature prediction, both predicted and actual temperatures show stable seasonal fluctuations, such as summer peaks and winter valleys recurring periodically; the fluctuations of actual temperatures are significantly larger than those predicted, especially in the extreme high temperature period, the predicted values deviate from the actual values up to 1-3°C; and most of the predicted time periods are persistently low. In terms of the performance of precipitation prediction, the model reflects the alternation of rainy and dry seasons, but the predicted peaks are lagged or advanced by 1-2 months; the predicted values of heavy precipitation events are only 65-80% of the actual values.

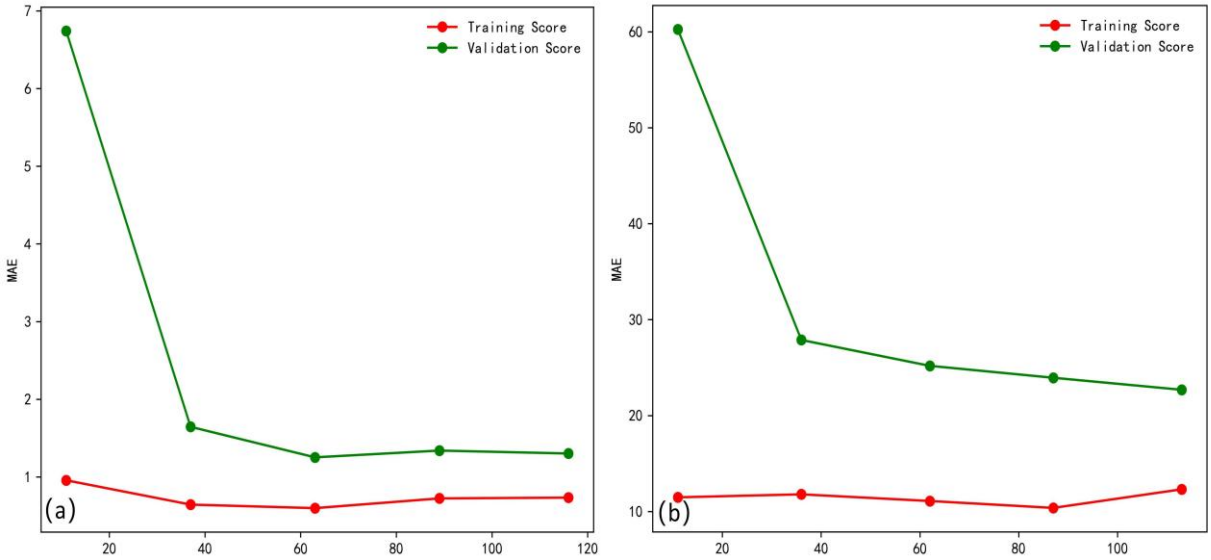


(a)Temperature curve predicted by RF model. (b)Precipitation curve predicted by RF model.

Figure 1. Comparison curves of air temperature and precipitation at Zhaotong station predicted by RF model.

Learning curves are plotted to assess the generalization performance of the model (Figure 2). From the analysis of Figure 2, it can be seen that in temperature prediction, when the sample size is <40, the MAE of the validation set decreases from 6.7 to 1.6, and the MAE of the training set decreases from 1 to 0.6, which indicates that the overfitting is serious at the initial stage, and the model is unable to be generalized; when the sample size is 40-60, the MAE of the validation set decreases from 1.6 to 1.3, and the MAE of the training set stabilizes at 0.6-0.7, and the gap between the two narrows down to about 0.5, and the model starts to capture climate patterns; after sample size > 60, both validation set and training set MAE stabilize, and the model tends to be balanced, but there is still a residual error of about 0.5, which may be affected by local factors other than ONI. In precipitation prediction, the MAE of the validation set decreases from 60 to 28 with a sample size <40, with a

decrease of 53%, reflecting that the precipitation prediction is extremely sensitive to the sample size; the MAE of the validation set slowly decreases from 28 to 25 with a sample size of 40-100, and the MAE of the training set stays around 10, indicating that the model has difficulty in learning the mechanism of heavy precipitation; after a sample size of >100, the MAE of the validation set is still higher than the MAE of the training set by 10 above.

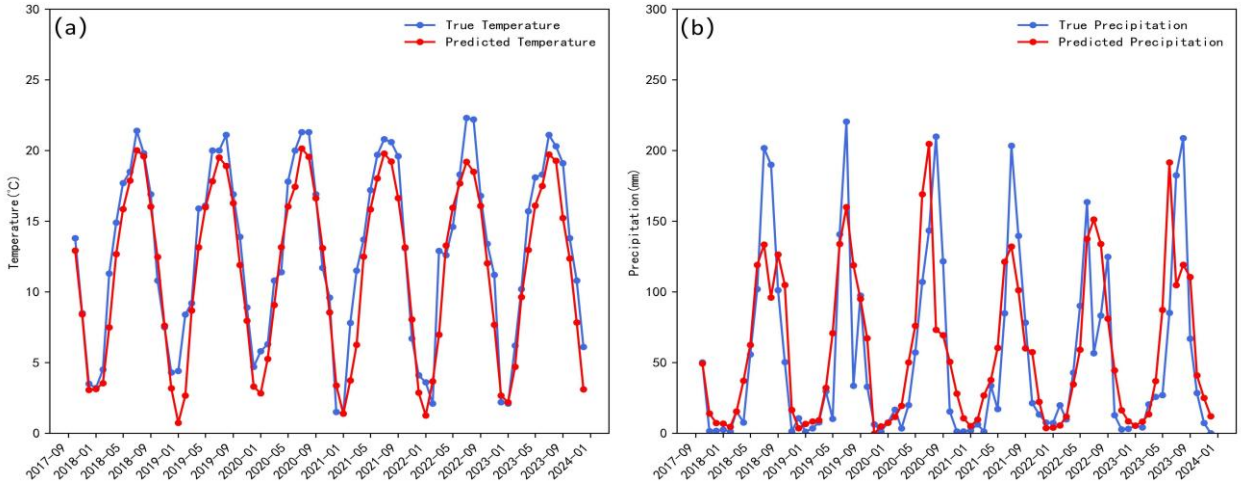


(a)Learning curve of RF model for predicting temperature.
(b)Learning curve of RF model for predicting precipitation.

Figure 2. Learning curve of RF model predicting air temperature and precipitation at Zhaotong station.

3.2. XGBoost model results

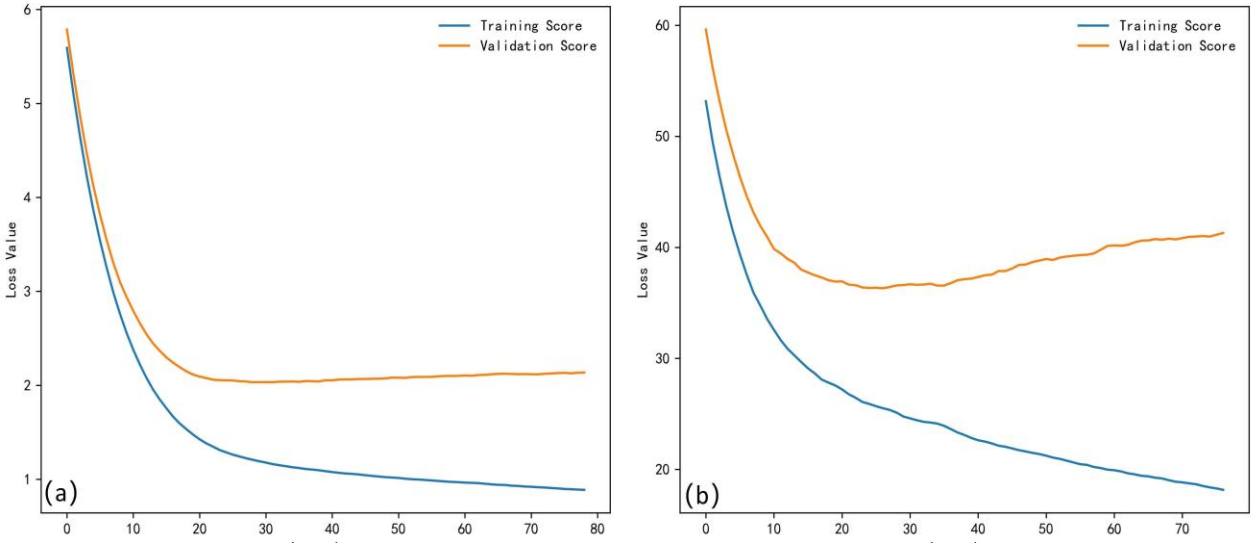
The results of the dataset division were input into the XGBoost model to predict the monthly average temperature and cumulative precipitation from October 2017-December 2023 (Figure 3). From the analysis in Figure 3, it can be seen that in the performance of temperature prediction, the predicted and actual temperature curves show a stable seasonal cycle; the model's prediction of extreme high temperatures is generally lower than the actual value of 1-3°C. In the performance of precipitation prediction, there is a phase shift of the rainy season, with the phenomenon of advancement or lagging; the model underestimates heavy precipitation, and the predicted value of most of the rainstorms with a value of more than 200 mm is only 65-75% of the actual value.



(a)Temperature curve predicted by XGBoost model. (b)Precipitation curve predicted by XGBoost model.

Figure 3. Comparison curves of air temperature and precipitation predicted by XGBoost model at Zhaotong station.

The change curves of the loss values of the training set and test set for temperature and precipitation under 80 iterations are plotted (Figure 4). From the analysis of Figure 4, it can be seen that in the process of temperature prediction, the loss value of the training set and test set decreases rapidly in the early stage, from 5.6 to 1.4 for the training set and from 5.8 to 2.1 for the test set in about 20 iterations, which indicates that the XGB captures the core features of the temperature rapidly; the loss curve tends to be flattened after more than 20 iterations, which indicates that the model has sufficiently learned the existing features, and it is difficult to improve the accuracy with continued training; the final loss of the test set is about twice as much as that of the training set under 80 iterations, reflecting moderate overfitting. The final loss in the test set is about twice that of the training set, reflecting moderate overfitting. In the precipitation prediction, the absolute loss value is high, and the final loss of the training set reaches 41, which is much higher than that of the temperature prediction, indicating that the precipitation is greatly affected by small-scale processes, and it is difficult to model adequately by ONI alone; the loss of the training set still oscillates between 37-41 after 20 iterations, which may be due to the existence of random noise in the precipitation event; the loss of the test set is about 2.3 times of the loss of the training set, and the degree of overfitting is higher than that of the temperature prediction.



(a)The loss curve of XGBoost model predicting temperature.
(b)The loss curve of XGBoost model predicting precipitation

Figure 4. XGBoost model predicts the loss curves of air temperature and precipitation at Zhaotong station.

3.3. Model Evaluation and Analysis of Predictive Results

The errors of RF and XGBoost models applied to precipitation and air temperature prediction are shown in terms of the mean absolute error (Figure 5), and the smaller the mean absolute error is, the closer the predicted values are to the actual observed values. From Figure 5, it can be seen that: (1) the RF model shows a slight advantage in the prediction of both types of elements; (2) there is a predictive element heterogeneity in the model performance, and the accuracy of RF and XGBoost for temperature prediction is better than that of precipitation prediction in general; (3) the temperature prediction is more inaccurate in February and March, and the precipitation prediction is more inaccurate in June and August, which is probably due to the fact that the amount of features is too small, and therefore the prediction error of temperature and precipitation is relatively larger in the case of climatic anomalies. precipitation prediction errors are relatively large.

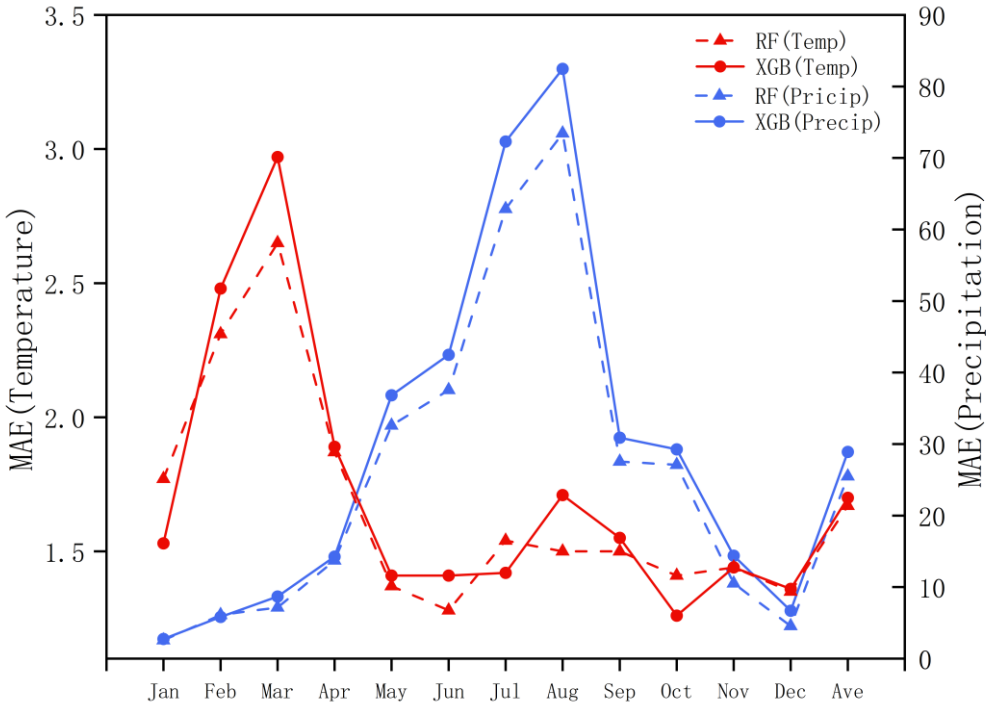


Figure 5. Results of the average absolute error of the two models in predicting temperature and precipitation.

4. Conclusions

In this study, Random Forest and XGBoost models were used to analyze the response relationship between temperature, precipitation and ENSO events in Zhaotong City, and the following conclusions were drawn:

(1) Random forests are slightly better than XGBoost in both temperature and precipitation prediction, with an average error of 1.67°C in temperature prediction and a higher error in precipitation prediction, mainly due to the significant influence of precipitation by small-scale climate processes and random noise.

(2) The limited ability of the two types of models to capture extreme high temperatures and heavy precipitation may be related to the single model feature (relying only on the ONI index) and the insufficient amount of training data.

(3) The weak performance of the model during the seasonal transition and the rainy season reflects the complexity of climate anomalies and the limitations of the model's fit to nonlinear relationships.

This study confirms the potential of machine learning methods in climate prediction in the mountainous regions of Southwest China, providing technical support for regional disaster prevention and mitigation, agro-meteorological services and water resource management. However, the prediction accuracy of the current model still needs to be improved, especially in the prediction of precipitation anomalies and extreme events. The following directions can be considered in future research: firstly, integrating multi-source data to enhance the model's ability to characterize the complex climate system; secondly, exploring hybrid modeling approaches, combining physical climate models and machine learning to improve the analysis of ENSO dynamics; and lastly, enhancing the attribution analysis of extreme climate events, and combining explanatory tools (e.g., SHAP values) to reveal the key driving factors. Through multidisciplinary crossover and technological integration, it is expected to promote the leap from statistical correlation to mechanism-driven climate prediction in Southwest mountains, and provide a more accurate scientific basis for coping with climate change.

References

[1] Zhao Y, Zhang Y, Yan Y, et al. Geographic distribution and impacts of climate change on the suitable habitats of two alpine Rhododendron in Southwest China[J]. Global Ecology and Conservation, 2024, 54: e03176.

[2] Wang L, Ma S. Extreme winter-spring drought in Southwest China in 2023: response to the phase transition from La Niña to El Niño[J]. Environmental Research Letters, 2024, 19(8): 084042.

[3] Qi Z, Cui C, Jiang Y, et al. Changes in the spatial and temporal characteristics of China's arid region in the background of ENSO[J]. Scientific Reports, 2022, 12(1): 17826.

[4] Wang C, Chen C, Zhang S, et al. Variation characteristics of extreme climate events in Southwest China from 1961 to 2017[J]. Heliyon, 2023, 9(9).

[5] Lv A, Fan L, Zhang W. Impact of ENSO events on droughts in China[J]. Atmosphere, 2022, 13(11): 1764.

[6] Latif S D, Hazrin N A B, Koo C H, et al. Assessing rainfall prediction models: Exploring the advantages of machine learning and remote sensing approaches[J]. Alexandria Engineering Journal, 2023, 82: 16-25.

[7] Ma C, Yao J, Mo Y, et al. Prediction of summer precipitation via machine learning with key climate variables: A case study in Xinjiang, China[J]. Journal of Hydrology: Regional Studies, 2024, 56: 101964.

[8] Wani O A, Mahdi S S, Yeasin M, et al. Predicting rainfall using machine learning, deep learning, and time series models across an altitudinal gradient in the North-Western Himalayas[J]. Scientific Reports, 2024, 14(1): 27876.

[9] Yang Dong, Wang Hui, Cheng Junqi, et al. Climate change in Qinghai and its relationship with ENSO in the recent 50 years [J]. Ecology and Environmental Sciences, 2013, 22(4): 547-553.

[10] Li Yu, Chen Min, Luo Jianfeng, et al. Climate change in the middle and lower reaches of the Yangtze river and Its relation to El Niño/La Niña events during 1951-2016[J]. Journal of China Three Gorges University (Natural Sciences), 2018,40(06):16-21.

[11] Yao Dengju, Yang Jing, Zhan Xiaojuan. Feature selection algorithm based onrandom forest[J]. Journal of Jilin University (Engineering and Technology Edition), 2014, 44(01): 137-141.

[12] Ma Yong, Kong Danli, Ye Xiangyang, et al. Application and comparison of type 2 diabetes with comorbid hypertension classfication prediction models based on random forest and XGBoost algorithms[J]. Journal of Guangdong Medical University, 2024,42(05):523-534.