DOI: 10.54254/2755-2721/94/2024MELB0066

Research on National Short-Term Precipitation Forecast Method Based on Machine Learning

YiTing Wang

School of International Business, Tianjin Foreign Studies University, Tianjin, 300000, China

wangyiting@ldy.edu.rs

Abstract. Climate change is closely related to human lives. With the development of information technology, meteorological data gradually show the characteristics of big data. The purpose of this study is to compare the application effects of different machine learning models in short-term precipitation prediction, to improve the accuracy and efficiency of prediction. Based on the annual average precipitation data of prefecture-level cities in China from 1990 to 2022, this paper uses Random Forest, eXtreme Gradient Boosting and Neural Networks to construct prediction models, and comprehensively evaluate them. In this study, the data is preprocessed to ensure the data quality of the input model. Next, the data is fed into three machine learning algorithm models, Random Forest, eXtreme Gradient Boosting and Neural Networks. Finally, the prediction performance of each model is evaluated by various indexes. In this paper, it is found that the Random Forest model has the best performance, and its prediction accuracy is higher than the other two models, which has great application potential in the field of short-term precipitation prediction. This study shows that a reasonable selection of machine learning methods and optimization of model parameters can effectively improve the accuracy of short-term precipitation prediction. This paper provides some empirical evidence for precipitation prediction, which will help to make more effective decisions in dealing with extreme weather events and climate change challenges in the future.

Keywords: Precipitation forecast, random forest, eXtreme gradient boosting, neural networks.

1. Introduction

Meteorological variations exert influences on all facets of human society and are intimately associated with people's daily existence. Climate change constitutes one of the events that receive the utmost attention from people [1]. Water resources are the basis for all living things to survive and reproduce in the natural ecological environment, and at the same time, those are the foundation for all human production activities and the sustainable development of the regional social economy [2]. Precipitation constitutes one of the most prevalent weather phenomena within the spectrum of complex meteorological variations, and the prediction of precipitation is one of the most fundamental meteorological forecasting services [3]. An accurate prediction of precipitation has significant implications for the daily lives of both urban and rural populations, human health as well as for aspects such as public health, and industrial and agricultural production [3].

Short-term precipitation forecasting has consistently remained a focal point within the domain of meteorological research. The primary challenge inherent in this pursuit emanates from the precipitation processes' inherently nonlinear and unpredictable characteristics, coupled with the constraints imposed by the limited spatial resolution and comprehensive observational data availability [4]. While traditional statistical methodologies and dynamical models may offer beneficial insights into precipitation forecasting to a certain degree, but inherently possess notable inadequacies in managing extensive datasets, high-dimensional attributes, and intricate nonlinear correlations. Machine learning algorithms are grounded in statistical learning theory, utilizing these methods to distillate salient features from multidimensional datasets, thereby accommodating intricate linear or nonlinear correlations. These algorithms have been instrumental in addressing and emulating numerous processes and critical elements within the hydrological cycle [5]. Due to the powerful capabilities of machine learning techniques in data approximation and feature extraction, these have emerged as a potent instrument for precipitation forecasting challenges.

In recent years, the development of information technology and data science has been rapid, and machine learning has seen rapid growth as a data analysis method with high efficiency, accuracy, and automation, which has been widely applied in the field of weather forecasting [6]. In an era of rapid technological advancement, the amount of data in human production and life has increased geometrically and gradually developed from bytes to gigabytes, terabytes, petabytes, and even yottabytes [7]. Big data technology came into being and gradually became the focus of scientific research [7]. By introducing machine learning techniques, the complex dependencies and patterns between precipitation and related climate variables can be more fully explored. Moreover, machine learning models are capable of integrating data from disparate sources and types, which contributes to enhancing the spatial resolution and precision of precipitation predictions. From an applied perspective, improving the performance of short-term precipitation forecasting will have profound implications for disaster prevention and reduction, resource management, and policy-making. The present study endeavors to delve into the technological pathway and implementation approaches for short-term precipitation forecasting via machine learning algorithms, with the ultimate goal of offering more accurate and time-sensitive predictive methodologies for relevant disciplines.

This paper involves a variety of data preprocessing methods, including missing value processing, outlier detection, data normalization and feature selection, aimed at enhancing the precision and dependability of predictions. Initially, the data preprocessing phase encompasses the selection of features. This involves the transformation of raw meteorological data, the handling of missing values, the implementation of encoding schemes for categorical variables, and the subsequent partitioning of the dataset into subsets for training and testing purposes. Furthermore, variables displaying a significant correlation with precipitation are identified and selected as input parameters for the model [8]. Concerning method selection, traditional machine learning algorithms such as Random Forest (RF), eXtreme Gradient Boosting (XGBoost) and Neural Networks possess substantial advantages in the range of regression prediction [3]. One of the methods used is RF, which belongs to the category of ensemble learning, which is a process of generating multiple simple models and analyzing their results [9]. Neural Networks are widely used machine learning tools and can be found with different architectures: simple ones and even very complex configurations [10]. Its efficacy is highly contingent on the nature of the data under analysis [10]. This study employs a comprehensive methodology encompassing experimental analysis and precipitation prediction models, utilizing annual average precipitation data of prefecture-level cities in China from the period of 1990 to 2022 to conduct research.

This article substantially enhances the precision of short-term precipitation forecasts through the application of machine learning methodologies. The anticipated results will yield a highly efficient and dependable tool for meteorological agencies and associated decision-makers, while simultaneously delving into novel avenues for the utilization of machine learning technology within the realm of meteorology.

2. Methodology

2.1. Data source and description

This study uses the annual average precipitation data of prefecture-level cities in China from 1990 to 2022, which was downloaded from the Model Whale Community platform. The average annual precipitation data for all provinces, cities, and counties in China from 1990 to 2022 provides a detailed record of the long-term precipitation patterns in different regions of China. These data cover a wide geographical area of China, including different climate zones and terrains. Average precipitation is the average amount of precipitation in a year and is crucial for understanding the water resources situation, agricultural irrigation needs, and the development of flood control measures in different regions. The precipitation in different provinces, cities, and counties may vary significantly due to the geographical location, terrain features, and climate conditions.

2.2. Index selection and description

The data set selected in this paper has 369 longitudinal data and 41 horizontal indicators. This is a longitudinal analysis of the data set after preprocessing, involving provinces and the average precipitation of each province from 1990 to 2022, as shown in Table 1, through which can conduct precipitation forecasts by comprehensively analyzing these indicators.

Table 1. The average precipitation of all provinces in China from 1990 to 2022.

Province	Average Precipitation(10-3)	Province	Average Precipitation(10-3)
Anhui	0.003612	Inner Mongolia	0.000983
Beijing	0.001743	Ningxia	0.001155
Fujian	0.004830	Qinghai	0.001437
Gansu	0.001188	Shandong	0.002039
Guangdong	0.005037	Shanxi	0.001737
Guangxi	0.005343	Shaanxi	0.002399
Guizhou	0.004577	Shanghai	0.003751
Hainan	0.004382	Sichuan	0.003650
Hebei	0.001636	Taiwan	0.007706
Henan	0.002356	Tianjin	0.001739
Heilongjiang	0.001827	Xizang	0.002435
Hubei	0.003827	Hong Kong	0.004925
Hunan	0.004928	Xinjiang	0.000701
Jilin	0.002137	Yunnan	0.004331
Jiangsu	0.003146	Zhejiang	0.004772
Jiangxi	0.005143	Chongqing	0.004289
Liaoning	0.002066	-	-

2.3. Method introduction

This paper mainly uses the following three methods to predict precipitation, and selects the annual precipitation of each province as a multivariate analysis, to obtain the model analysis results.

2.3.1. Random forest (RF). The Random Forest algorithm used in this study is a highly flexible machine learning method and an ensemble learning algorithm used for classification and regression. RF predicts by constructing multiple decision trees and determines the final output by aggregating the results of each tree through voting or averaging. The advantage of this algorithm is its robustness to noisy data and overfitting. RF algorithm is a Bagging algorithm that uses decision trees as estimators, with the basic unit being a decision tree [2]. The higher the number of trees in the model, the higher the accuracy, without overfitting the model, thus improving the accuracy of the prediction results. The RF algorithm implements parallel computing during model training, with fast running speed, high prediction accuracy, strong generalization ability, and insensitivity to missing features in the split [2]. The working principle of RF is to train each decision tree using bootstrap samples, which are samples drawn with replacement from the original data, and only consider a subset of features at each split point to increase the diversity of the model.

2.3.2. eXtreme gradient boosting (XGBoost). In machine learning applications, XGBoost is an advanced ensemble algorithm based on the Boosting framework. The essence is based on the gradient boosting algorithm, using decision trees as weak classifiers, and is the implementation and extension of the Gradient Boosting Decision Tree (GBDT) algorithm [11, 12]. This method has shown excellent performance in parallel computing efficiency, handling missing values, and prediction accuracy, and is widely used in data science competitions and various machine learning tasks. XGBoost optimizes the second-order Taylor expansion of the objective function and uses an additive model to only optimize the submodel for the current step in each iteration, effectively improving the accuracy and computational efficiency of the model. XGBoost is currently widely used in various fields, similar to other machine learning algorithms, and is typically used for classification and prediction data, and can also be used for feature selection.

2.3.3. Neural networks. The Neural Network constitutes a computational model comprised of an extensive array of interconnected neurons, which are deployed to elucidate intricate patterns within datasets. The architecture of a Neural Network encompasses multiple tiers, such as an input layer, a constellation of hidden layers, and an output layer. Each layer is composed of numerous nodes that are linked using synaptic weights, which facilitate the conveyance of information between nodes. Through the iterative application of the backpropagation algorithm, the Neural Networks can internalize and decipher the intrinsic attributes and correlations resident in the data.

3. Results and discussion

3.1. Data preprocessing

Before training the model, the annual average precipitation data collected from each prefecture-level city in China from 1990 to 2022 was thoroughly preprocessed, as shown in Figure 1.

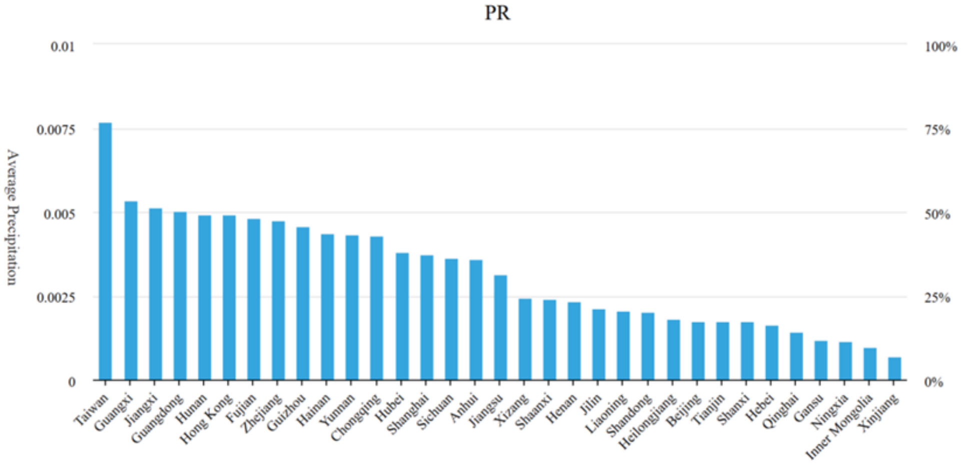

Figure 1. Histogram of average precipitation in each province from 1990 to 2022.

The data preprocessing included handling missing values, detecting outliers, normalizing the data, and feature selection. The missing values were filled in using interpolation methods, obvious outliers were removed, and the data was standardized to conform to the standard normal distribution. The data of the same province is integrated to simplify the data and unify the data magnitude of precipitation. The characteristics that were highly correlated with precipitation were selected through correlation analysis and feature importance scoring.

3.2. Random forest model results

Before using the RF model, set its parameters to default values, as shown in the following Table 2.

Table 2. RF model parameter setting.

Parameter name	Parameter value
Data preprocessing	norm
Training set ratio	0.8
Decision tree quantity	100
Node splitting criterion	gini
Minimum sample number of node splitting	2
Minimum sample number of leaf nodes	1
Maximum tree depth	No limit
Limit the maximum number of features	auto
Whether there is put back sampling	Yes
Whether to perform out-of-pocket data testing	Yes

As can be seen from Table 3, the RF model demonstrates commendable stability and precision during cross-validation. Through the amalgamation of numerous decision trees, the RF model possesses the capability to manage high-dimensional data and mitigate the risk of overfitting. Inference from the test set indicates that the RF model yielded a Mean Absolute Error (MAE) of 4.213 millimeters and a Root Mean Square Error (RMSE) of 5.631 millimeters, suggesting robust predictive capabilities of the model.

Table 3. Results of RF model evaluation.

Index	Training set time	Test set time
R-squared	0.940	0.638
Mean Absolute Error	1.692	4.213
Mean-square Error	5.424	31.713
Root Mean Square Error	2.329	5.631
Median Absolute Deviation	1.090	3.305
Mean Absolute Percentage Error	0.833	0.504
Explained Variance Score	0.940	0.640
Mean Squared Logarithmic Error	0.075	0.225

3.3. XGBoost model results

Before using this method for research, first set its parameters as the Table 4. The maximum depth value of the tree was adjusted from the default value of 6 to 4, and the remaining parameters were set as default so that the model could better predict the data set.

Table 4. XGBoost model parameter setting.

Parameter name	Parameter value
Data preprocessing	norm
Training set ratio	0.8
Elevator type	gbtree
Learner number	100
Learning rate	0.1
Maximum tree depth	4
Sample rate	1.0
Characteristic sampling rate	1.0
Minimum child node weight	1.0
Split income threshold	0.0
L1 regularization	0.0
L2 regularization	1.0

As an integrated algorithm, the XGBoost model has high accuracy and interpretability. This method can show good generalization ability on some complex data sets. In this study, by adjusting the parameter of maximum tree depth, the XGBoost model achieved an MAE of 4.212 mm and an RMSE of 5.825 mm on the test set, as shown in Table 5. According to the fitting effect analysis, the performance of this method in this dataset is inferior to that of random forest.

Table 5. Results of XGBoost model evaluation.

Index	Training set time	Test set time
R-squared	0.990	0.613
Mean Absolute Error	0.675	4.212
Mean-square Error	0.926	33.927
Root Mean Square Error	0.962	5.825
Median Absolute Deviation	0.450	2.887
Mean Absolute Percentage Error	0.275	0.480
Explained Variance Score	0.990	0.616
Mean Squared Logarithmic Error	0.014	0.223

3.4. Neural networks model results

When using the Neural Networks model to analyze and predict data, its parameters should be set and modified first, as shown in Table 6. For the data set used in this study, change the Batch Size from auto to custom and customize it to 60. To make the model more complex, the hidden layer of neurons is set to 25 layers with 100 neurons each. Retain the default values for other parameters.

The deployment of deep learning Neural Network models has emerged as a predominant trend in the domain of precipitation forecasting research within the past few years. The model, which encompasses several hidden layers, incorporates the ReLU activation function to enhance nonlinearity. Subsequent to rigorous training and meticulous parameter optimization, as can be seen from Table 7 the Neural Networks model yielded an MAE of 6.061 millimeters and an RMSE of 7.418 millimeters on the validation dataset.

Table 6. Neural Networks model parameter setting.

Parameter name	Parameter value
Data preprocessing	norm
Training set ratio	0.8
Hidden layer neuron setup	(25,100)
Activation function	relu
Weight optimization method	adam
L2 regularization coefficient	0.0001
Initial learning rate	0.0001
Learning rate optimization method	constant
Customize the Batch Size	60
Maximum iterations	200
Optimize tolerance	0.001

Table 7. Results of Neural Networks model evaluation.

Index	Training set time	Test set time
R-squared	0.494	0.372
Mean Absolute Error	5.240	6.061
Mean-square Error	45.405	55.023
Root Mean Square Error	6.738	7.418
Median Absolute Deviation	3.886	5.723
Mean Absolute Percentage Error	2.470	0.648
Explained Variance Score	0.497	0.372
Mean Squared Logarithmic Error	0.311	0.321

3.5. Comparison results

Comparing the performances of the three models, the author found that RF had the best overall accuracy and stability, with the best predictive effect. For the data set used in this study, among the three methods, the RF model is more effective in predicting precipitation. XGBoost is slightly inferior to the RF model, but it also has better predictive power. Neural Networks model requires larger and more complex data sets to work well.

The preprocessing stage of this paper is an important foundation for establishing an accurate prediction model. Firstly, this study cleans and normalizes the original precipitation data, thereby effectively improving the data quality and reducing the influence of outliers, laying a good foundation for the subsequent training of machine learning algorithms.

Before using three methods to predict precipitation in this study, parameters were set and adjusted for all three models. All three methods set the data preprocessing method as the norm and the proportion of the training set as 0.8, that is, 80% of all data was used for training the model, and the remaining 20% was used to test the performance of the model.

By comparing the MAE and RMSE values of the results of the three models, it can be seen that for this data set, using RF for precipitation prediction has the best effect. This is because both of these two values are better when they are closer to 0, and in this study, the values obtained by the RF result are the smallest, showing a high prediction accuracy. However, the disadvantage of RF is that when the data set is very large, the training time of the model will increase significantly, which is a concern for real-time precipitation prediction systems.

XGBoost, another effective prediction tool, was also included in the study scope. XGBoost usually provides better generalization ability on large sample datasets, with excellent performance and flexibility. However, the performance of XGBoost is highly dependent on the choice of parameters, and inappropriate parameter settings may lead to poor fitting effects, affecting the final prediction results. This study tries to solve this problem by changing the maximum depth of the tree. The maximum depth of the tree was set to 3, 4, 5, 6, and 10 respectively. As shown in Table 8, the fitting effect of the dataset was best when the maximum depth of the tree was set to 4 according to the R-squared value.

Table 8. R-square values for maximum depth values of different trees.

Maximum tree depth	Training set time	Test set time
3	0.955	0.549
4	0.990	0.613
5	0.999	0.589
6	1.000	0.598
10	1.000	0.567

According to the R-square value, the results obtained by the two methods are between 0 and 1, although the difference is not large, the R-square value of RF is closer to 1. Thus, it can be seen that the fitting effect of RF is better than that of XGBoost.

Table 9. R-square values under the number of layers of different hidden layers of neurons.

Hidden layer neuron setup	Training set time	Test set time
1	-1.015	-1.180
5	0.030	0.017
10	0.133	0.079
15	0.387	0.183
20	0.418	0.276
25	0.494	0.372
30	0.332	0.244

This study attempted to use the Neural Networks model to predict precipitation (Table 9). In this experiment, the L2 regularization coefficient was changed several times to change the degree of fitting, but the final effect is still average according to the R-square value. Since there are a total of 369 data in this data set, which is not much for the Neural Networks model, the batch size value was customarily set to 60, and other parameters were set to default. The data fitting model obtained for the first time was very bad, and the R-square value was less than 0, meaning that this method is not feasible. Then, the important parameter value of the hidden layer neurons was changed. Due to the current small data sample and low feature items, it was decided to make the model more complex, that is, to increase the number of neuron layers. The hidden layer neurons were set to 5 layers, 10 layers, 15 layers, 20 layers, 25 layers, and 30 layers, with 100 neurons in each layer. As can be seen from Table 9 before setting to 25 layers, the effect of the data fitting model is getting better and better. When set to 30 layers, the effect of the data fitting model is not as good as before, and the experiment is stopped. Although Neural Networks have strong data analysis capabilities, the training process requires much more and more complex data resources, and this model is more suitable for analyzing large data sets. As the model becomes more complex, the time consumed by the model for analyzing data also becomes longer, which will limit its promotion in practical applications to a certain extent.

Based on the performance of the above methods, it can be known that each method has its advantages and is suitable for different prediction scenarios. RF is suitable for small data set analysis and prediction and initial exploration due to its stability and ease of use. XGBoost performs worse than RF on small sample problems, and it is recommended to use RF for such data analysis and prediction. Neural Networks show great potential in processing large data sets and capturing complex relationships and are especially suitable for long-term trend analysis and high-precision complex prediction tasks.

Finally, the findings of this study emphasize the importance and potential of applying machine learning techniques in the field of short-term precipitation prediction. With the continuous advancement of technology, more model fusion strategies and algorithm optimizations can be explored in the future to improve the accuracy and efficiency of predictions. In addition, this study did not analyze the climatic differences in different regions, and future work can also focus on customized machine learning solutions to better serve the precipitation prediction needs of specific regions.

4. Conclusion

This study aims to explore short-term precipitation prediction methods based on machine learning. Through the use of data preprocessing techniques, RF, XGBoost and Neural Networks, a comprehensive analysis was conducted on the annual average precipitation data of various prefecture-level cities across the country from 1990 to 2022. Combining the performances of these three machine learning algorithms in precipitation prediction, an effective method for short-term precipitation prediction is provided.

By comparing the prediction results of different models, it can be found that the RF model can provide more stable and accurate predictions when the dataset is small and the complexity is low, and the effect is better than the XGBoost model. However, the Neural Networks model has obvious advantages in predicting a large number of data items and fitting complex datasets.

These results have significant implications for future research. Firstly, the results confirm the effectiveness and potential of machine learning methods in precipitation prediction. Secondly, the methods and prediction results adopted in this study can provide a reference for the prediction of similar climate variables and promote the development of related fields. In addition, these achievements also suggest that future work can further explore model fusion strategies to integrate the advantages of different models and improve the accuracy of predictions.

The prospects for future research include but are not limited to the following aspects. Firstly, considering the variability of meteorological conditions and regional differences, future research can develop customized prediction models for different regions. Secondly, emerging models and techniques in the field of deep learning, such as Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN), may offer additional advantages in processing spatial and temporal sequence data and are worth exploring in future work.

In conclusion, this study not only provides an effective machine-learning solution for short-term precipitation prediction but also lays a foundation for future research and application of meteorological prediction, demonstrating broad development prospects. With the advancement of technology and the development of new algorithms, it is expected that this field will witness more innovations and breakthroughs.

References

[1] Jie S 2023 Research on Rainfall Prediction Method Based on Machine Learning. Shenyang University of Technology, 519.

[2] Min G 2023 Application of Machine Learning in Precipitation Forecasting. Automation Applications, 64(07), 22-25

[3] Kanghui S, An X and Hou X 2024 Research on the Short-term Temperature Forecast Model of Jiangxi Based on the LightGBM Machine Learning Algorithm. Plateau Meteorology, 1-16.

[4] Xing S, Zhou C, et al. 2015 Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting. Neural Information Processing Systems, 28.

[5] Yuhang Z and Aizhong Y 2022 Machine Learning for Precipitation Forecasts Postprocessing: Multi model Comparison and Experimental Investigation. Hydrometeor, 22. 3065-3085.

[6] Yuan S and Tai Z 2024 The Current Application of Machine Learning in Public Health. Preventive Medicine Forum, 30(01), 77-80.

[7] Tiant T, Dong J, Tao C and Guan G 2022 Medium-and Long-Term Precipitation Forecasting Method Based on Data Augmentation and Machine Learning Algorithms. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 15, 1000-1011.

[8] Liyew C M and Melese H A 2021 Machine learning techniques to predict daily rainfall amount. J Big Data, 8, 153.

[9] Yiwen M and Sorteberg A 2020 Improving Radar-Based Precipitation Nowcasts with Machine Learning Using an Approach Based on Random Forest. Wea. Forecasting, 35, 2461-2478.

[10] Anochi J A, de Almeida V A and de Campos Velho H F 2021 Machine Learning for Climate Precipitation Prediction Modeling over South America. Remote Sensing, 13, 2468.

[11] Ogunleye A and Wang Q G 2019 XGBoost model for chronic kidney disease diagnosis. IEEE/ACM transactions on computational biology and bioinformatics, 17(6), 2131-2140.

[12] Zhuo L, Cheng W and Qiu L, et al. 2019 Prediction model of aluminized layer thickness based on X-ray fluorescence and extreme gradient lifting. Advances in laser Light and Optoelectronics, 59(21), 262-269.