Proceedings of the 4th International Conference on Computing Innovation and Applied Physics
DOI: 10.54254/2753-8818/84/2025.21209

Machine Learning-based Precipitation Prediction for Chengdu City

Enyu Hu[1,a,*]

[1]*School of Mathematics Swufe, Southwestern University of Finance and Economics, Chengdu, Sichuan Province, 611130, China*
a. 42257015@smail.swufe.edu.cn
**corresponding author*

Abstract: Chengdu is located in the western part of the Sichuan Basin, with a humid climate and fertile soil, making it an important agricultural production base. In recent years, seasonal precipitation distribution in Chengdu has been uneven, and extreme precipitation events are frequent. This paper uses a machine learning method to predict precipitation based on seven types of meteorological data including precipitation from January 1990 to December 2023 in Chengdu. Autoregressive Moving Average Model (ARIMA), eXtreme Gradient Boosting (XGBoost), and Random Forest Regression (RFR) are established, support vector Regression (SVR) four machine learning models. The results show that the XGBoost model performs the best among these methods, and it can accurately capture the trend of precipitation, thus providing a reliable means for precipitation prediction and rational planning of water resources in Chengdu. This study contributes by offering a practical application of machine learning techniques to precipitation forecasting, which could inform future strategies for water resource management in the region.

Keywords: Chengdu precipitation, XGBoost ARIMA, Support Vector Machine, Random Forests.

1. Introduction

Chengdu is a unique climate resource in the transition zone from the Northwest Sichuan Plateau to the Sichuan Basin [1]. Precipitation is an important component of the Earth's water cycle, and the average annual precipitation in Chengdu is relatively abundant. Still, the distribution of precipitation in time and space is not uniform, with obvious seasonal and regional differences. This uneven precipitation distribution pattern poses certain challenges to local water resources management and agricultural production. With the influence of global climate change, the precipitation pattern in Chengdu City has also undergone some changes. Over the past 43 years, precipitation has increased in the northwestern part of Chengdu while decreasing in the southeastern part, with an overall decreasing trend [2]. At the same time, the frequency and intensity of extreme precipitation events have increased, which poses a potential threat to the city's flood control and drainage system and agricultural production safety. In addition, with the rapid economic development and population growth in Chengdu City, water usage has been continuously increasing. Therefore, this study is important for rational planning of water resource utilisation, improving the city's flood control capacity, and ensuring the safety of agricultural production.

© 2025 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2753-8818/84/2025.21209

Precipitation prediction using machine learning algorithms has become more sophisticated in recent years, presenting a high degree of accuracy [3]. Zhang used the ARIMA model to predict the 2014-2018 precipitation in Weinan City and found that except for 2018, the relative errors of the four years were all within 10%, and the ARIMA model can better implement the precipitation prediction [4]. Xu established a VMD-QPSO-SVR model to predict the month-by-month precipitation from 1961 to 2022 at four stations in Jiujiang City, and the goodness-of-fit was found to be above 0.7, which verified the validity and superiority of the model [5]. Dong et al. carried out the daily precipitation prediction work for up to 8 days at 689 stations in China by using the XGBoost model. The results were found that the RMSE was 1.819-13.608 mm, and the average gain, coverage, and frequency values were 0.55, 0.45, and 0.29, respectively, which showed good model prediction performance [6].

In terms of precipitation prediction, most of the existing studies have only selected precipitation as the only independent variable. This paper envisages constructing a machine learning model with multiple independent variables and meteorological factors for local short-term precipitation prediction in Chengdu city. To analyze and contrast the prediction outcomes, eXtreme Gradient Boosting (XGBoost), Random Forest Regression (RFR), Autoregressive Moving Average Model (ARIMA), and Support Vector Regression (SVR) were used, aiming to explore the effectiveness of machine learning methods in predicting precipitation.

2. Methods: introduction to theory

2.1. ARIMA

The Autoregressive Sliding Average Model (ARIMA) was introduced in the 1970s by G.P.E. Box and Jenkins G.M. of the University of Wisconsin, USA, which was designed to be used for analysing time-series information, which in turn has been applied in the area of forecasting the future as well as in controlling it [7]. To identify patterns and periodicity in time series data, ARIMA integrates the ideas of autoregression (AR), differencing (I), and moving average (MA). Based on the BIC criterion, the optimal value of (p, q) in the ARIMA precipitation regression model is selected and the regression equation is:

$$y_j = \alpha_1 y_{j-1} + \alpha_2 y_{j-2} + \cdots + \alpha_m y_{j-m} - \beta_1 \varepsilon_{j-1} - \beta_2 \varepsilon_{j-2} - \cdots - \beta_n \varepsilon_{j-n} + \varepsilon_j \quad (1)$$

Where represents the precipitation that the model predicts: α_1, α_2, ..., α_m denote the regression coefficients of the model; β_1, β_2, ..., β_n denote the number of moving auto orders of the model; ε_j denotes the calculation error of the model.

2.2. XGBoost model

Extreme Gradient Boosting Model (XGBoost), as an improved algorithm of Gradient Boosting Decision Tree (GBDT), has a decision tree structure, which directly allows the model to be visualised compared to other machine learning algorithms, as well as having the advantages of preventing overfitting, high computational accuracy, and fast computing speed [8]. The objective function is the sum of the loss function and complexity, which can be written as:

$$Obj_t = \sum_{k=1}^{m} l(q_k, G_{t-1}(\vec{p}_k) + f_t(\vec{p}_k)) + \Omega(f) \quad (2)$$

Where: $\Omega(f_t)$ denotes the regularisation term of the XGBoost model and $l(q_k, G_{t-1}(\vec{p}_k) + f_t(\vec{p}_k))$ is the value of the loss function between the true value and the model's predicted value. γ, λ Denote the specification parameters for the number of control lobes and the regularisation strength, respectively.

DOI: 10.54254/2753-8818/84/2025.21209

To achieve the optimal value of the objective function, it is obtained by Taylor expansion:

$$Obj_t \approx \sum_{k=1}^{m}[l(q_k, G_{t-1}(\vec{p}_k)) + g(\vec{p}_k, q_k)f_t(\vec{p}_k) + \frac{1}{2}h(\vec{p}_k, q_k)f_t^2(\vec{p}_k)] + \Omega(f) \quad (3)$$

Included among these:

$$g(\vec{p}_k, q_k) = \frac{\partial l\left(q, G_{t-1}(\vec{p})\right)}{\partial G_{t-1}(\vec{p})}, h(\vec{p}_k, q_k) = \frac{\partial^2 l(q, G_{t-1}(\vec{p}))}{\partial G_{t-1}^2(\vec{p})} \quad (4)$$

Treat$l(q_k, G_{t-1}(\vec{p}_k))$ as a constant term and continue to simplify the objective function, then take the partial derivatives of all leaf nodesω_j, make the partial derivatives 0, and carry out back generation to get the final objective function:

$$Obj_t^* = \gamma J - \frac{1}{2}\sum_{j=1}^{J}\frac{M_j^2}{N_j+\lambda} \quad (5)$$

Where $M_j = \sum_k g(\vec{p}_k, q_k), N_j = \sum_k h(\vec{p}_k, q_k)$, $\omega_\xi(\vec{p})$ Denotes the value taken by the leaf node.

2.3. Random forest model

Random Forest Regression (RFR) is an approach that uses decision trees. Random Forest is essentially trained by combining multiple decision trees to improve the accuracy of model prediction [9]. Random forest is widely used in financial analysis, environmental science prediction, network security, image recognition, medical diagnosis, and other fields. Random forest has the advantages of a significant quantity of processable information and high precision due to its randomness and the unique characteristics of integrated learning.

2.4. Support vector machine models

Support Vector Regression (SVR) model is a regression algorithm based on Support Vector Machines that can explore the nonlinear relationship between the dependent variable and the response values affected by the independent variables [10]. Due to the use of kernel functions, SVR has the advantage of dealing with strong nonlinear relationships. At the same time, the SVR model is relatively insensitive to possible outliers in the data.

3. Data sources and description

3.1. Data sources

In this paper, seven data items, including daily precipitation (PRCP), temperature (TEMP), mean visibility (VISIB), wind speed (WDSP), dew point (DEWP), maximum sustained wind speed (MXSPD), and barometric pressure (STP), are collected from meteorological observation stations in Shuangliu District, Chengdu City, China, from 1 January 1990 to 31 December 2023. The National Oceanic and Atmospheric Administration (NOAA) of the United States of America's National Center for Environmental Information (NCEI) is the source of the dataset. Strict quality control is applied to every data indicator to guarantee the data's dependability and accuracy.

3.2. Brief description

The multiple interpolation method in SPSS software analysis was used to interpolate the missing data values. The variables with more missing values were directly eliminated, and seven variables, including precipitation, were finally screened out.

DOI: 10.54254/2753-8818/84/2025.21209

Detection of outliers using box-and-line plot method: upper and lower limits are calculated, outliers are identified, and for variables with fewer outliers, such as temperature, precipitation, etc., the outliers are directly eliminated. The outliers are replaced with mean values for variables with more outliers, such as pressure. The formula is as follows, taking the calculation of the lower limit as an example:

$$\text{outlier} < Q_1 - 1.5 \times IQR \tag{6}$$

Where Q_1 is the first quartile and IQR is the interquartile spacing?

Finally, the data set was normalised by equation (6) in order to address the difference in magnitude.

$$X_* = \frac{X - X_{min}}{Xmin_{max}} \tag{7}$$

4. Analysis of statistical results

4.1. Basic statistical analyses (graphs, charts)

Before normalising the data, descriptive statistics were carried out using SPSS software to obtain a mean daily precipitation value of about 2.928 mm, a minimum value of 0 mm, and a maximum value of about 126.9 mm.

Secondly, it was tested whether each set of variables obeyed normal distribution. The JB test was conducted using MATLAB, and the results showed that the H-matrix were all 1, indicating that the original hypothesis was rejected at the 95% confidence level. All the variables did not obey normal distribution, consequently, the analysis employed Spearman's correlation coefficient, which is a statistical indicator used to measure the correlation between two variables. It is calculated principally based on the variables' rank order (position after sorting) rather than the original values of the variables. The formula is as follows:

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} \tag{8}$$

Where d_i is the rank difference between the two sets of data.

The correlation coefficients between precipitation and the remaining six variables were calculated. As shown in Table 1, precipitation has a favorable correlation with temperature, dew point, visibility, wind speed, and maximum sustained wind speed, and it is negatively correlated with pressure. A significance test was conducted, and all p-values were less than 0.01, which is considered to be a significant correlation between the amount of precipitation and the six indexes at a 99 percent confidence level, which demonstrates the reasonableness of the prediction of the amount of precipitation through these six variables.

Table 1: Spearman's correlation coefficients between two of the variables

	Temp	Dew point	Pressure	Increased visibility	Air velocity	The highest sustained wind speed
Measured quantity of rain	0.380**	0.0502**	-0.199**	0.245**	0.410**	0.362**

** The correlation is significant at the 0.01 level (two-tailed).

DOI: 10.54254/2753-8818/84/2025.21209

4.2. ARIMA model and analysis

Firstly, the daily precipitation from 1990 to 2023 was processed by data accumulation to obtain monthly precipitation, and due to the completeness of the data, the time series graph can be plotted directly using SPSS. As shown in Figure 1, the horizontal coordinate is the time, and the vertical coordinate is the corresponding monthly precipitation value. From Figure 1, it can be initially judged that the time series has a smooth trend in precipitation over time. Further ADF unit root test, the time series data ADF test t-statistic is -4.007, 1%, 5%, 10% critical value of -3.447, -2.869, -2.571. A P-value of 0.001 is less than or equal to 0.01, therefore, the initial hypothesis can be rejected with greater than 99% certainty; this time, the sequence is smooth, with no unit root.

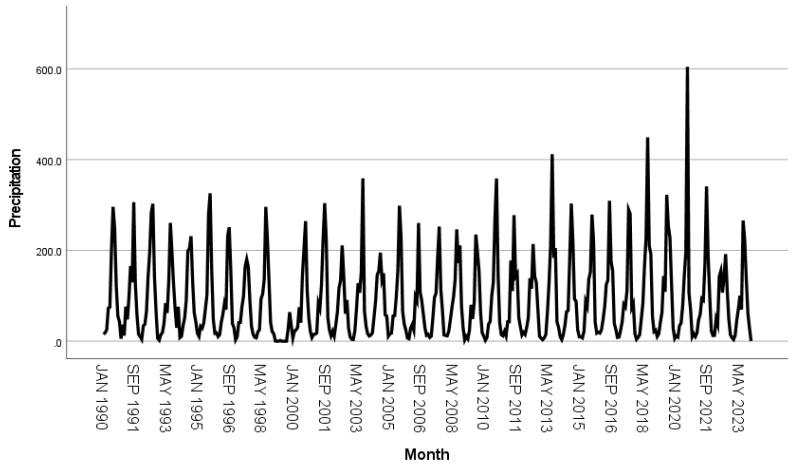


Figure 1: Precipitation time series plot (Photo/Picture credit: Original).

Secondly, the modelling analysis was carried out using the expert modeller of SPSS software, which resulted in the optimal model being the ARIMA(0,1,1) model. Then, the significance of the model, i.e., the white noise of the residual series, was tested, as shown in Table 2. The statistic Q=21.916, the p-value is 0. 146, which is significantly greater than the test level of 0.05, and it can be considered that the residuals are white noise. It is initially determined that the model can identify the precipitation data better.

Table 2: Q test results

Model fit statistics	Jan Boks Q (18)		
Smooth R-square	statisticians	DF	significance
0.753	21.916	16	0.146

Finally, the model was used to forecast the period from January 2024 to December 2025, as shown in Figure 2. The horizontal coordinate is the time variable, and the vertical coordinate is the monthly precipitation. The real measurements are displayed by the red line, the yellow line shows the model-fitted values, and the expected values are displayed by the blue line. The smoothed R-squared is 0.753, and then based on the standard BIC criterion, the normalised BIC is 7.152, and the model prediction is not as expected.

DOI: 10.54254/2753-8818/84/2025.21209

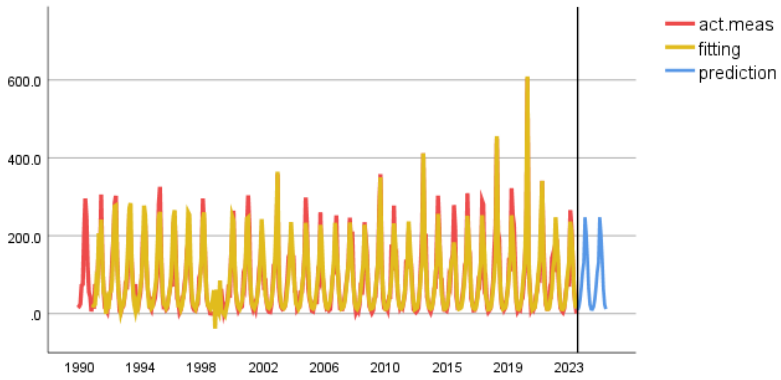


Figure 2: Fitting and forecasting of monthly precipitation in Chengdu city (Photo/Picture credit: Original).

4.3. Predicting rainfall based on machine learning models

GridSearch is a common approach to parameter tuning. It systematically tries different combinations of parameters in a given parameter space to find the optimal parameter configuration [11].

The optimal solution between the number of models and the number of decision trees is selected by using a five-fold cross-validation method of grid search, followed by a grid search for the tree's maximum depth, the minimum number of leaf node samples, the random sampling ratio, the feature random sampling ratio, the regularisation parameter, and the learning rate. In the grid search process, these parameters need to be searched several times, and the search interval is continuously narrowed to obtain their optimal values. As shown in Table 3, the ideal values for the parameters of the three models are obtained, in which the maximum depth of XGBoost is 12.

Table 3: Optimal parameter combinations of XGBoost, Random Forest, SVR

XGBoost	Optimal parameter values	random forest	Optimal parameter values	SVR	Optimal parameter values
Number of base learners	350	Number decision trees	235	Kernel	RBF
Max depth	12	Max depth	10	Highest degree	3
L_1	0	Max samples split	2	C	0.8
Rate of learning	0.1	Leaf min samples	1	epsilon	0.03

Plotting the histogram of the characteristic contribution rate of each variable in the XGBoost model, as shown in Figure 3, shows the characteristic contribution rate of each indicator to the precipitation amount. Among them, the indicator dew point has the largest characteristic contribution rate of about 26%, indicating that it has the largest role in predicting precipitation. The other five indicators also have a role in the prediction.

DOI: 10.54254/2753-8818/84/2025.21209

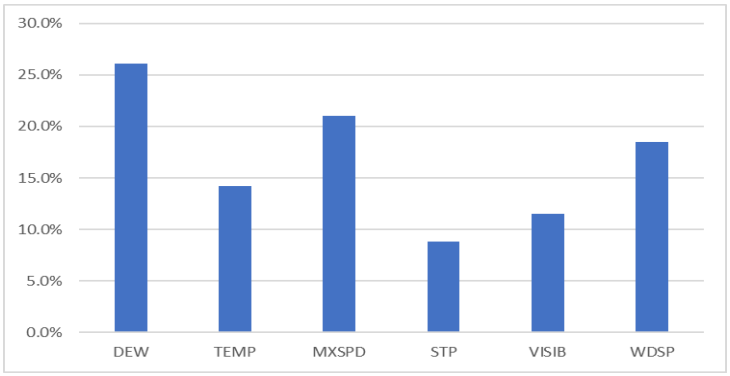


Figure 3: Histogram of feature contribution (Photo/Picture credit: Original).

Finally, the outcomes of the machine learning model in this paper are evaluated using three parameters: mean square error (MSE), mean absolute error (MAE), and decidability coefficient R-squared. The smaller MSE and MAE are, the smaller the prediction error is, and the better the model performance is; R^2 measures the proportion of variability explained by the model to the total variability, and the closer the value is to 1, the stronger the model's explanatory ability is [12]. As shown in Table 4, the evaluation results of the three models are analysed, and XGBoost has an R-squared of 0.947 and the best prediction.

Table 4: Results of the assessment of indicators

	MAP	MAE	R^2
XGBoost model	0.02343	0.01658	0.947
random forest	0.04697	0.05072	0.612
SVR	0.05486	0.03422	0.739

5. Conclusion

In this paper, based on four models of machine learning, ARIMA, XGBoost, Random Forest and SVR, the precipitation prediction of Chengdu city was carried out with the dataset of day-by-day precipitation from January 1990 to December 2023, and the precipitation prediction of Chengdu city was carried out.

Comparing the prediction results of several models, it is found that XGBoost reflects better performance and superiority in predicting precipitation, where the indicator dew point has the largest feature contribution among all indicators, reaching 26%. It is followed by maximum sustained wind speed, wind speed, and temperature. This indicates that when predicting precipitation, one can focus on the dew point temperature of the day; when the dew point temperature rises to a certain level, the water vapour in the air condenses more easily, thus forming precipitation. Meanwhile, temperature and wind speed also play an important role in precipitation formation.

This paper provides quantitative indicators and a theoretical basis for future water conservancy work, the future work of precipitation prediction, machine learning models can be used for precipitation prediction, the establishment of a perfect climate prediction and indicator monitoring system, focusing on daily meteorological indicator changes, comprehensive analysis, rational planning of water resources, and prevention of potential disasters.

The meteorological factors used for prediction in this paper are limited to six. More meteorological data can be added to forecast precipitation and update the model in real time. Machine learning models are not analysed in more detail in this paper. In addition, it fails to consider the effects of geomorphology and extreme weather on precipitation in Chengdu. On the model, limitations such as

DOI: 10.54254/2753-8818/84/2025.21209

overfitting, insufficient generalisation ability, and more complex operations may occur to be improved in the future.

References

[1] *Wang, R., Lin, W., Wang, Y., et al. (2015). Analysis of annual changes in precipitation in Chengdu from 1951 to 2010. Anhui Agricultural Bulletin, 21(18), 146-147.*

[2] *Xie, H., Yang, P., & Zhang, Y. (2022). Spatial and temporal distribution characteristics of annual precipitation in Chengdu. Journal of Chengdu Information Engineering University, 37(3), 356-360.*

[3] *Ma, S., Wang, W., Pei, L., et al. (2023). Analysis of drought trend in Xianyang's "dry belt" area based on LSTM precipitation prediction. Anhui Agricultural Science, 51(13), 192-197.*

[4] *Zhang, C. (2019). Analysis and prediction of precipitation trend in Weinan City based on ARIMA model. Value Engineering, 38(34), 197-199.*

[5] *Xu, C. (2024). Characterisation of precipitation change in Jiujiang City and precipitation prediction based on VMD-QPSO-SVR (Doctoral dissertation, Nanjing University of Information Engineering).*

[6] *Dong, J., Zeng, W., Wu, L., et al. (2023). Enhancing short-term forecasting of daily precipitation using numerical weather prediction bias correction with XGBoost in different regions of China. Engineering Applications of Artificial Intelligence, 117, 105579.*

[7] *Box, G. P. E., & Jenkins, G. M. (1976). Time series analysis: Forecasting and control. Holden-Day.*

[8] *Fu, Z., Zou, H., & Yuan, J. (2024). Optimisation of soot blowing scheme for small circulating fluidised bed boilers based on XGBoost. Control Engineering, 1-5. https://doi.org/10.xxxx/xxxxx*

[9] *He, X., He, H., Yang, L., et al. (2024). Analysis of influencing factors of tobacco sales based on random forest model. Information Technology, (11), 147-153.*

[10] *Duan, X., Yu, G., Jiao, Y., et al. (2024). Prediction model and impact analysis of cutting force of boring titanium alloy based on SVR model. Modern Industrial Economy and Information Technology, 14(11), 150-152, 157.*

[11] *Yan, J., Li, S., Wei, Z., et al. (2024). Shale petrographic logging prediction method based on XGBoost algorithm. Journal of Palaeogeography, 1-14.*

[12] *Pan, L., Chen, S., Wang, H., et al. (2024). Temperature prediction of high-voltage rotor hazardous point based on machine learning. Journal of Southeast University (Natural Science Edition), 1-10.*