International Journal of Dynamics and Control (2024) 12:1180–1196
https://doi.org/10.1007/s40435-023-01250-1

Precipitation prediction in several Chinese regions using machine learning methods

Yuyao Wang[1] · Lijun Pei[1] · Jiachen Wang[2]

Received: 10 April 2023 / Revised: 15 June 2023 / Accepted: 16 June 2023 / Published online: 12 July 2023

Abstract
The severity of global climate change is exemplified by the significant increase in extreme precipitation events, leading to an urgent need for accurate rainfall prediction models to mitigate flood disasters that adversely affect economic and social development. With the rapid progress of machine learning in the big data era, novel solutions to regression problems are being proposed. In this paper, we try to construct and evaluate different rainfall prediction models based on specific humidity, relative humidity, horizontal and vertical water vapor flux, and lifting index as variables, using four classic machine learning algorithms: linear regression, random forest regression, support vector regression, and Bayesian ridge regression. The grid search method is employed for hyperparameter tuning, significantly improving the models' prediction accuracy and generalization ability. Evaluation of the predictive performance of the models on nine typical regions in China, including Zhengzhou, Beijing, and Chengdu, demonstrates that the random forest regression model has the highest predictive accuracy, with an average fitting degree of 0.8 or above, followed by support vector regression and Bayesian ridge regression models. Conversely, the linear regression model may have the poorest predictive performance. Therefore, the random forest regression model is recommended for future precipitation prediction, providing a valuable solution to various regression problems. The appropriate selection of variables for prediction and grid search for hyperparameter tuning are possibly the highlights of this paper.

Keywords Rainfall prediction · Machine learning · Linear regression · Random forest · Support vector regression · Bayesian ridge regression

1 Introduction

Rainfall is a critical factor for human life and the national economy, with changes in rainfall patterns leading to extreme situations such as floods and droughts, greatly affecting agriculture, water resources, and the ecological environment [1]. The increase in extreme rainfall events in recent years, driven by global climate warming, highlights the need for accurate rainfall information to prevent water disasters and manage water resources more effectively. There have been many studies on rainfall prediction problems both domestically and internationally, and the commonly used rainfall prediction methods are mainly divided into two categories: probability statistical methods and time series analysis [2]. However, these methods have limitations. Probability Statistical Methods, including the gray GM(1,1) model, exponential smoothing, and Markov models, can only predict data with large random fluctuations. Time series prediction methods, including autoregressive model, auto regressive moving average models, and autoregressive integrating moving average (ARIMA), tend to predict values close to the average and are inaccurate in predicting extremes [3]. Thus, traditional rainfall prediction models face challenges in precisely forecasting rainfall due to their inherent limitations.

With the advent of the big data era, machine learning methods have become increasingly mature, offering significant advantages in prediction problems due to their high flexibility and data driven learning ability. Machine learning is widely applied in water conservancy engineering for agricultural irrigation, water quality testing, and reservoir scheduling

✉ Lijun Pei
peiilijun@zzu.edu.cn

[1] School of Mathematics and Statistics, Zhengzhou University, Zhengzhou 450001, Henan, China

[2] School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, Sichuan, China

[4]. In the stock market, it predicts future stock prices or returns [5], while in social sciences, it is used in intelligent homes, smart transportation, medical inspection technology, and other fields [6]. These applications underscore the critical role of machine learning in driving progress in these areas. With the potential to discover hidden laws in meteorological data, machine learning can establish appropriate regression prediction models, leading to accurate predictions of future rainfall. Therefore, the use of machine learning for rainfall prediction is a meaningful topic worth exploring [7].

In recent years, researchers from both domestic and foreign institutions have conducted studies on the use of machine learning methods for rainfall prediction. Gocic and Shamshirband utilized the linear regression method to predict the rainfall trend based on monthly data from 1946 to 2012 in 29 regions of Serbia. They divided the predicted 66 year data into three time series combinations to investigate changes in rainfall from 1986 to 2012 [8]. Similarly, Sulaiman predicted heavy rainfall occurrences in the same region using an artificial neural network (ANN) model. They sorted precipitation data from local weather stations between 1965 and 2015 and divided the monthly precipitation values from the past 50 years into different combinations before using them as prediction input data. The predictive performance of the ANN model is evaluated using the mean square error and correlation coefficient [9]. However, there have been relatively fewer studies conducted in this field in China. The constructed rainfall prediction models are mostly based on specific regions, making it difficult to verify their universality and generalization ability. There is also a lack of comparisons between different machine learning models. The scientific standard for model selection has not been established, and selecting factors affecting rainfall is complicated, leading to difficulties in collecting meteorological data [10].

In light of the existing literatures and outstanding issues in rainfall prediction, in this paper we propose a practical solution. Firstly, this study selects five variables, including specific humidity, relative humidity, horizontal water vapor flux, vertical water vapor flux, and lifting index, as variables to combine the causes of the flood at Zhengzhou on July 20, 2021 and the necessary conditions for heavy rainfall. This approach achieves accurate rainfall prediction with fewer variables, reducing the difficulty of data collection. Secondly, the grid search method is used for hyperparameter tuning on the initial machine learning regression models to find the parameter combination that maximizes the prediction accuracy within a certain range, thus improving the model's prediction accuracy and generalization ability. Finally, the adjusted goodness of fit and mean square error are used as evaluation indicators to measure the model's performance. By comparing the rainfall prediction results of several typical regions in China, such as Zhengzhou, Beijing, and Chengdu, the optimal model applicable to the rainfall prediction problem is scientifically selected. These findings offer new ideas and methods for resolving other practical regression problems and serve as a valuable reference for further exploration of regression analysis algorithms.

The remainder of this article is organized as follows. Section 2 elaborates on the theoretical basis of machine learning prediction theory, including linear regression (LR), random forest regression (RFR), and support vector regression (SVR) and Bayesian ridge regression. In Sect. 3, the application of machine learning methods in rainfall prediction is discussed, including the selection of variables and data acquisition and processing. Then we employ the four aforementioned models to predict rainfall results in different regions. Section 4 compares the prediction results of the different methods. Finally, limitations of this study and its future research are also discussed.

2 Prediction theory and methods of machine learning

In this section, theory and methods of machine learning for prediction are introduced.

2.1 Linear regression model

Linear regression algorithm (LR) is a common machine learning algorithm with features such as easy understanding, convenient execution, and wide application [11]. The basic principle of the linear regression model is to analyze the relationship between the dependent variable and one or more independent variables and use this analysis to train a linear equation that approximates the sample data through training data. The goal is to accurately predict the target value as much as possible based on the input variables [12]. When there is only one independent variable in the regression model, it is called a simple linear regression model. When the regression model contains two or more independent variables, it is called a multiple linear regression model. The specific steps for constructing a linear regression model are as follows:

1. Determine the number of variables and select appropriate independent variables based on the scenario.
2. Determine the error measurement standard and select the loss function.
3. Find the optimal model performance by changing the optimizer for the scenario.

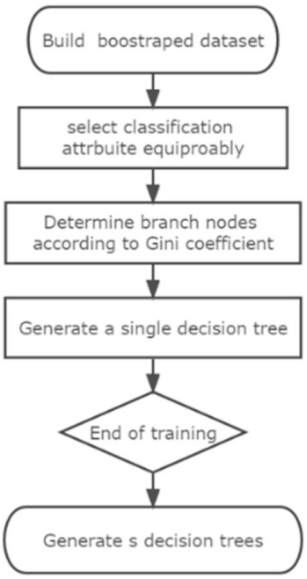

Fig. 1 The process of constructing Random Forest model

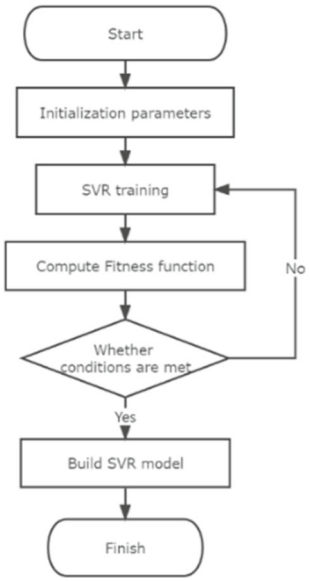

Fig. 2 The process of constructing SVR model

2.2 Random forest regression model

Ensemble learning algorithm is a method of combining multiple classifiers to improve prediction performance [13]. According to the ensemble methods, they can be roughly divided into two categories: Bagging (parallel) and Boosting (serial). The random forest algorithm, known for its simplicity and efficiency, is one of the representative algorithms of ensemble learning algorithms. It is a machine learning algorithm proposed by Breiman in 2001 by combining the Bagging ensemble learning theory with the random subspace method [14]. In classification problems, the final classification result of the random forest model is determined by voting based on the output results of individual decision trees, following the principle of majority rule [15]. In prediction problems, the random forest model integrates the prediction results of many decision trees and outputs them in the form of average [16]. Due to its good performance, the random forest model has achieved great success in many application fields. The training process of the random forest model is shown in Fig. 1 [17].

2.3 Support vector regression model

Support vector regression (SVR) is a machine learning method that uses statistical learning theory to perform regression calculations based on the idea of support vector machine (SVM). It is widely used to solve nonlinear problems and is suitable for finite sample studies, with the ability to theoretically obtain globally optimal solutions [18].

This model transforms the actual problem into a high dimensional feature space through nonlinear transformation and constructs a linear decision function in the high dimensional space to achieve the nonlinear decision function in the original space, effectively solving the dimensionality problem [19]. Additionally, this model has the advantages of simple structure, strong generalization ability, and high prediction accuracy [20]. It has a wide range of applications in function approximation, regression prediction, and other areas. The training process of SVR is displayed in Fig. 2 [21].

2.4 Bayesian ridge regression model

Ridge Regression is a model tuning method specialized in collinearity data analysis [22]. It is essentially an improved least squares estimation method that sacrifices some information and reduces accuracy to obtain more realistic and reliable regression coefficients, and it fits ill-conditioned data better than the least squares method [23].

Bayesian linear regression is a linear regression model solved using Bayesian inference methods in statistical learning [24]. Bayesian linear regression views the model's stochastic parameters as random variables, calculates their posterior probabilities based on the prior probabilities of the model's parameters (weight coefficients), and uses the maximum likelihood method to estimate unknown parameters.

The Bayesian ridge regression model combines the advantages of the ridge regression model and Bayesian linear regression model and introduces an L_2 regularization term on the basis of Bayesian linear regression estimation [25]. It is not only suitable for predicting normal data but also has a good fitting effect when there is multicollinearity among independent variables. Due to the model's adaptive ability to

Table 1 RPD Degree of Stability

RPD	Degree of Stability
RPD $\leq$ 1.4	Unstable
1.4 $\leq$ RPD $\leq$ 2.0	Relatively Stable
2.0 $\leq$ RPD	Highly Stable

Next we will use Zhengzhou to illustrate how the prediction models work. The results are similar in other regions

Table 2 Intercept and Coefficients

β_1	β_2	β_3	β_4	β_5	β_0
2.7806	2.7125	0.8325	0.8845	0.1598	6.5965

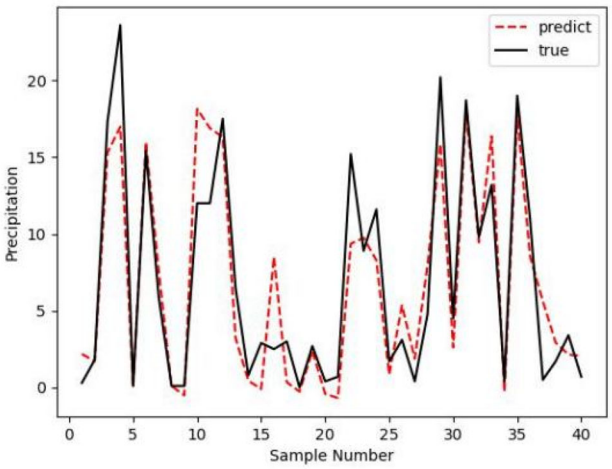

Fig. 3 Forecast comparison results of the LR model

Table 3 Parameters of Random Forest

Parameter	The effect of parameters on model	Level
n_estimators	Number of base classifiers, the more the number, the more stable the model	Level1
max_depth	The smaller the maximum depth, the simpler the model is	Level2
mini_samples_leaf	The smaller the value, the more complex the model	Level3
mini_samples_split	The smaller the value, the more complex the model	Level3
max_leaf_nodes	Maximum number of leaf nodes, used to prevent overfitting by limiting the maximum number of leaf nodes	Level3
max_features	Number of features to be considered	Level4
criterion	Evaluation criterion for decision tree feature splitting, default is Gini index	Level5

the data, it can reuse data, which largely solves the problem of overfitting in maximum likelihood estimation [26].

3 Machine learning based rainfall prediction

3.1 Data acquisition and processing

3.1.1 Data description

In this paper, we use a dataset acquired from the Google Earth Engine platform, including monthly rainfall data of nine regions in China: Zhengzhou, Beijing, Hohhot, Kunming, Lhasa, Lanzhou, Chengdu, Jinan, and Xi'an, from January 1, 1980 to April 1, 2013.

3.1.2 Feature selection

The essence of rainfall is the condensation of water vapor in the air, and the amount of rainfall depends on the water vapor content, condensation efficiency, and duration. Therefore, analyzing the water vapor conditions in the atmosphere mainly involves the distribution of water vapor content and water vapor flux [27].

There are various ways to express water vapor content: specific humidity (q) represents the mass of water vapor per unit mass of moist air, while water vapor pressure (e) represents the partial pressure of water vapor in moist air. Both specific humidity and water vapor pressure represent the absolute water vapor content in the atmosphere and are therefore referred to as absolute humidity. Relative humidity (RH) is also a fundamental physical parameter that characterizes water vapor content and is widely used in rainfall prediction[28].

Water vapor flux, also known as water vapor transport, is a highly important parameter that indicates both the intensity and direction of water vapor transport [29]. It can be divided into horizontal water vapor flux and vertical water vapor flux. The horizontal or vertical transport of water vapor is an essential component of the atmospheric water cycle and is closely related to the formation of precipitation. Therefore, long-term observations of water vapor flux are of significant research significance for predicting rainfall.

The lifting index is a measure of convective instability. It represents the temperature difference between an air parcel, starting from the observed surface and rising along the dry adiabatic process to the lifting condensation level, and then

Table 4 Optimal Hyperparameter Combination

Parameter	Optimal Value	Parameter	Optimal Value
n_estimators	10	max_leaf_nodes	None
max_depth	6	max_features	sqrt
mini_samples_leaf	7	criterion	gini
mini_samples_split	2		

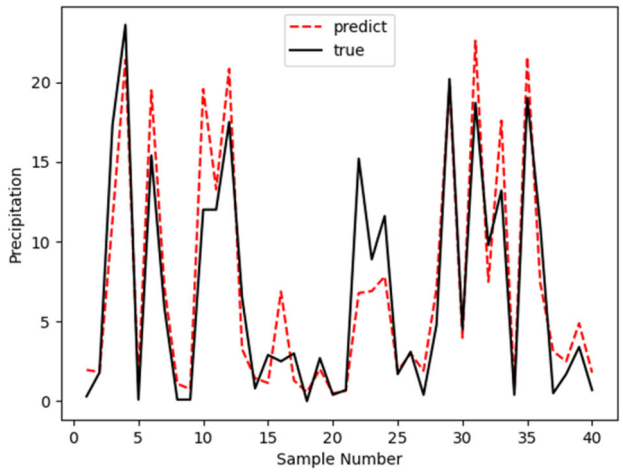

Fig. 4 Forecast comparison results of the Random Forest model

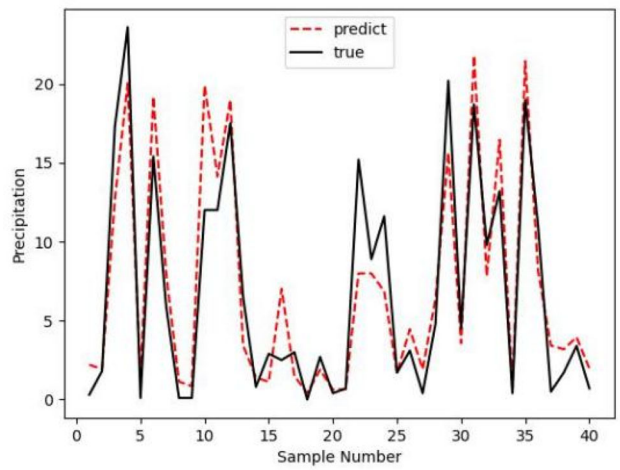

Fig. 5 Forecast comparison results of the SVR model

ascending along the moist adiabatic process to 500 hPa, compared to the actual atmospheric temperature at that level[30]. When the lifting index is negative, it indicates atmospheric instability, and the larger the negative value, the greater the degree of instability, making it more prone to precipitation. Conversely, positive values indicate atmospheric stability and a lower likelihood of precipitation.

Therefore, taking into consideration the main factors influencing rainfall and the necessary conditions for heavy

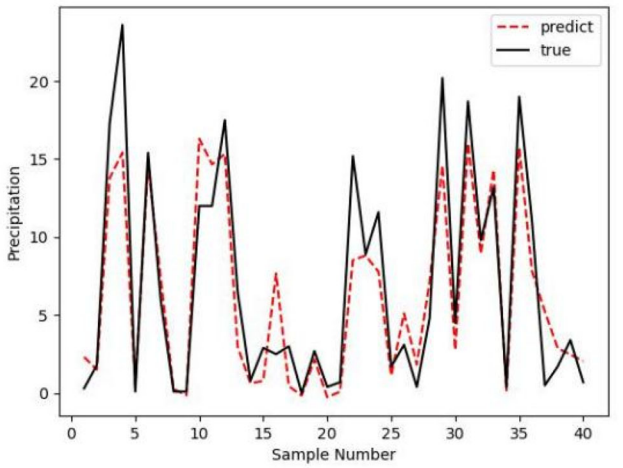

Fig. 6 Forecast comparison results of the Bayesian Ridge Regression model

rainfall, which include abundant water vapor, atmospheric dynamics, and atmospheric stability, we ultimately select five physical quantities: specific humidity (q), relative humidity (RH), horizontal water vapor flux $|F_H|$, vertical water vapor flux F_z, and lifting index (LI) as feature indicators for predicting rainfall. They can be calculated as follows:

$$\text{Specifichumidity} : q = \frac{0.6220e}{P - 0.3780e}, \tag{1}$$

$$\text{Relativehumidity} : RH = \frac{e}{e_w}, \tag{2}$$

$$\text{Horizontalwatervaporflux} : |F_H| = \frac{|V|q}{g}, \tag{3}$$

$$\text{Verticalwatervaporflux} : F_z = -\frac{wq}{g}, \tag{4}$$

$$\text{Liftingindex} : LI = T_{500} - T'_{\text{suf}}. \tag{5}$$

In the formula above, e represents water vapor pressure, e_w represents saturated water vapor pressure, u is the longitudinal wind, v is the latitudinal wind, w is the vertical wind, V is the total wind speed, T_{500} is the temperature of an air parcel that rises along the moist adiabat from the convective condensation level to 500 Pa, and T'_{suf} is the actual temperature at 500 hPa.

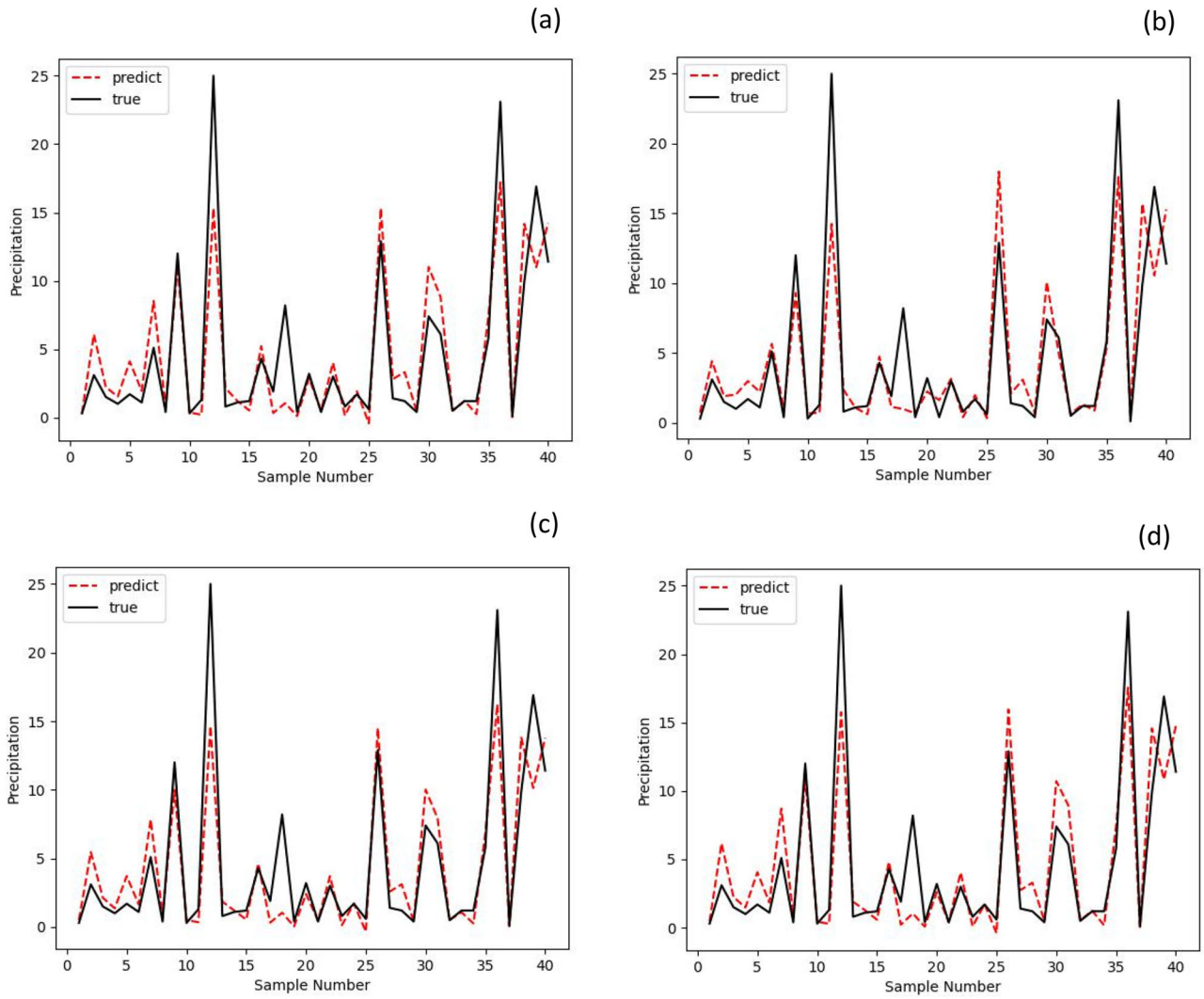

Fig. 7 Forecast results comparison of the LR, RF, SVR, BRR models for Beijing in (a), (b), (c), (d) respectively

3.1.3 Data processing

Python software is used to process and analyze the retrieved dataset. Based on the variable parameters in the dataset, five variables are calculated as input data for the model, with the daily average of monthly rainfall as the output data. A total of 400 sets of sample data are obtained after data processing. The train_test_split function is called to divide the sample data into 90% training samples and 10% testing samples before the data is standardized. The standardization formula is as follows,$x^* = \frac{x-\mu}{\sigma}$. (6)

The training sample is used to build the model and the testing sample is used to evaluate the model. To effectively compare the predictive performance of each model and ensure that the prediction results are not affected by the randomness of the data set division, a random seed is selected for each region, and the same sample sequence is used each time the model is constructed.

3.2 Model construction

The model construction in this study is based on the Python learning library. Four models, namely linear regression, random forest, SVR, and Bayesian ridge regression, are selected for regression prediction [31]. To compare the predictive performance and stability of each model, the adjusted R squared, mean squared error (MSE), and RPD (relative percent deviation) are chosen as evaluation indicators, and their calculation formulas are as follows:

$$Adjusted - R^2 = 1 - \frac{(1-R^2)(n-1)}{(n-k-1)}, \tag{7}$$

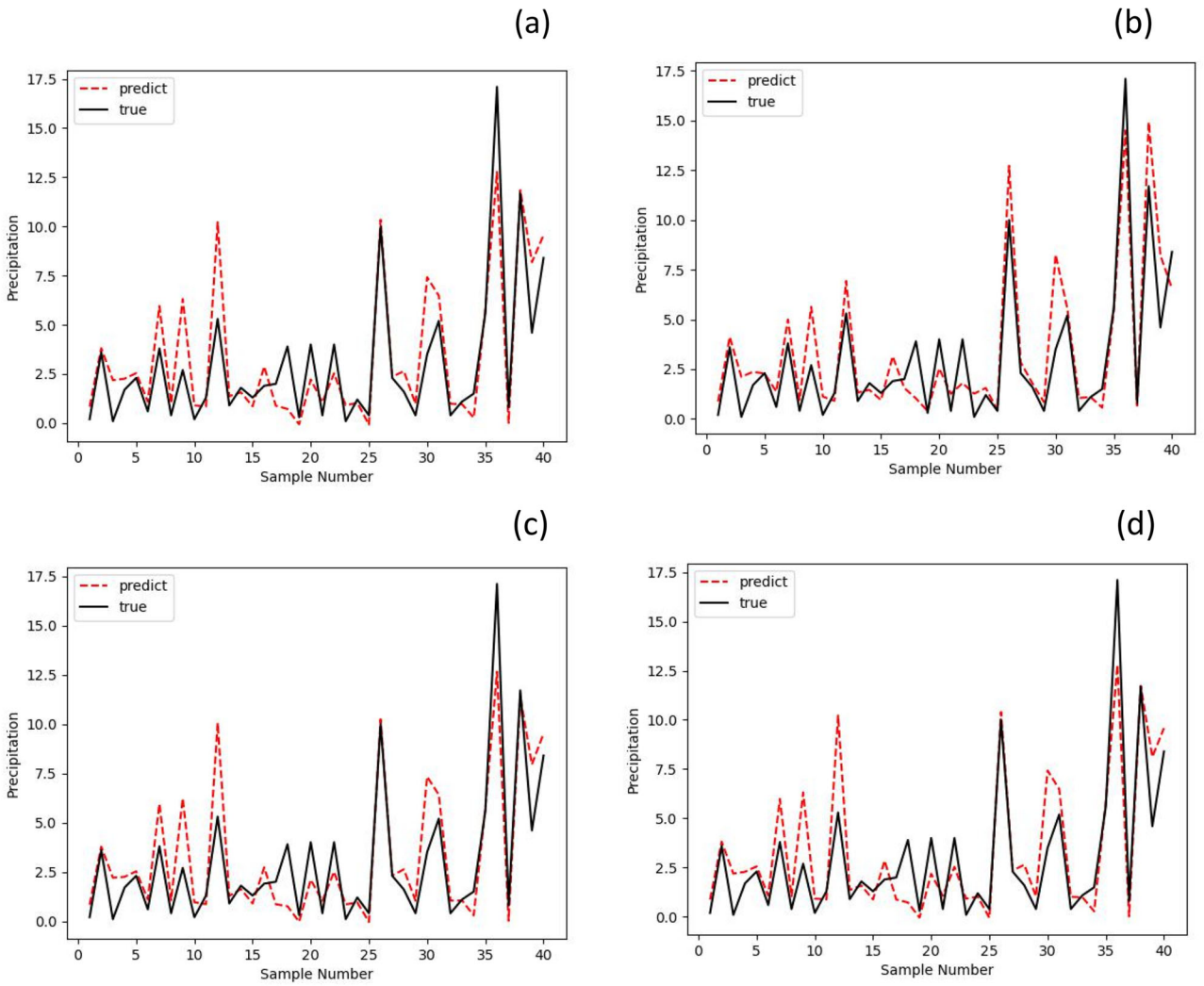

Fig. 8 Forecast results comparison of the LR, RF, SVR, BRR models for Hohhot in (a), (b), (c), (d) respectively

$$MSE = \frac{\sum_{i=1}^{n} (\widehat{y_i} - y_i)^2}{n}, \tag{8}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (\widehat{y_i} - y_i)^2}{n}}, \tag{9}$$

$$SD = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \overline{y})^2}{n}}, \tag{10}$$

$$RPD = \frac{SD}{RMSE}. \tag{11}$$

The adjusted R-squared is a measurement of the goodness of fit of the model, and the larger the value, the better the model fits the data. Mean Squared Error (MSE) is a common measure of the average squared difference between the predicted and actual values. It is commonly used in statistical modeling and machine learning to evaluate the performance of regression models. Mathematically, MSE is calculated by taking the average of the squared differences between the predicted and actual values. RPD (relative percent deviation) is an indicator to measure the stability of the model [32], and the stability level corresponding to the RPD value is shown in Table 1.

3.2.1 Linear regression model

The linear regression model can be written in the following matrix form [33].

$$QX = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_5 x_5 + \beta_0, \tag{12}$$

$$QX = \sum_{j=0}^{5} \beta_j x_j = X'\beta, \tag{13}$$

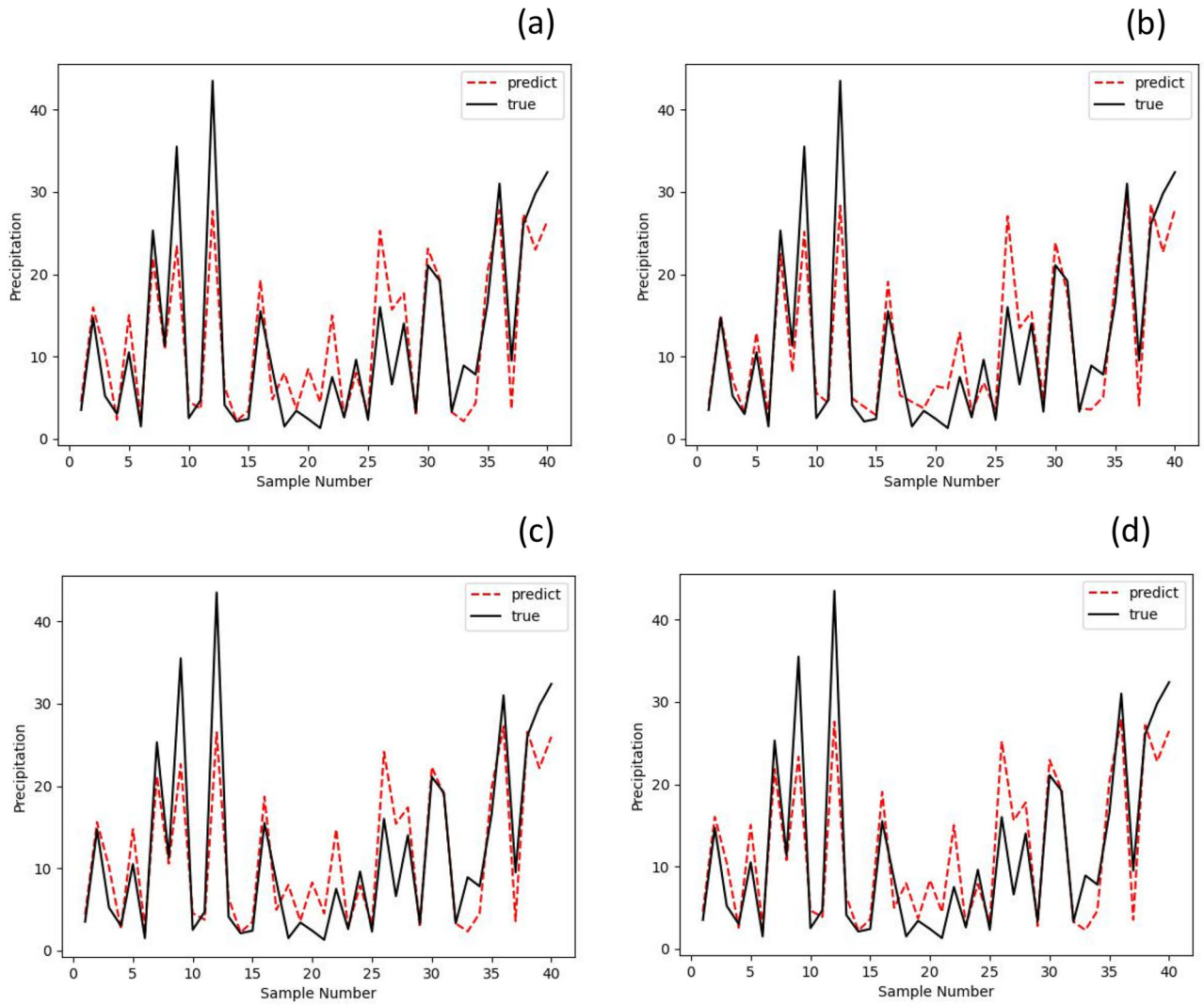

Fig. 9 Forecast results comparison of the LR, RF, SVR, BRR models for Kunming in (a), (b), (c), (d) respectively

$$X^{'} = [x_1, x_2, \cdots, x_5, 1],\ \beta \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_5 \\ \beta_0 \end{pmatrix} \tag{14}$$

where $Q(X)$ is the predicted value, x_j are the variables, and β_j are the weights corresponding to the features. Here x_1= specific humidity (q), x_2= relative humidity(RH), x_3 = horizontal water vapor flux $(|F_H|)$, x_4 = vertical water vapor flux(F_z), and x_5 = lifting index(LI).

To evaluate the model's predictive performance, the following loss function is utilized,

$$J(\beta) = \sum_{i=1}^{n} (QX_i - q_i)^2. \tag{15}$$

$J(\beta)$ is the loss value, n is the evaluation capacity, $Q(X)_i$ is the predicted value of the ith training sample, and q_i is the true value of the training sample.

The normal equation method [34] and gradient descent method [35] are used to find the optimal solution for the loss function, $minJ(\beta)$.

Comparison of the prediction results indicates that the gradient descent method has a better optimization effect. The final model's adjusted R-squared value on the test set is 0.836, and the mean squared error is 7.971. The Intercept and Coefficients are shown in Table 2.

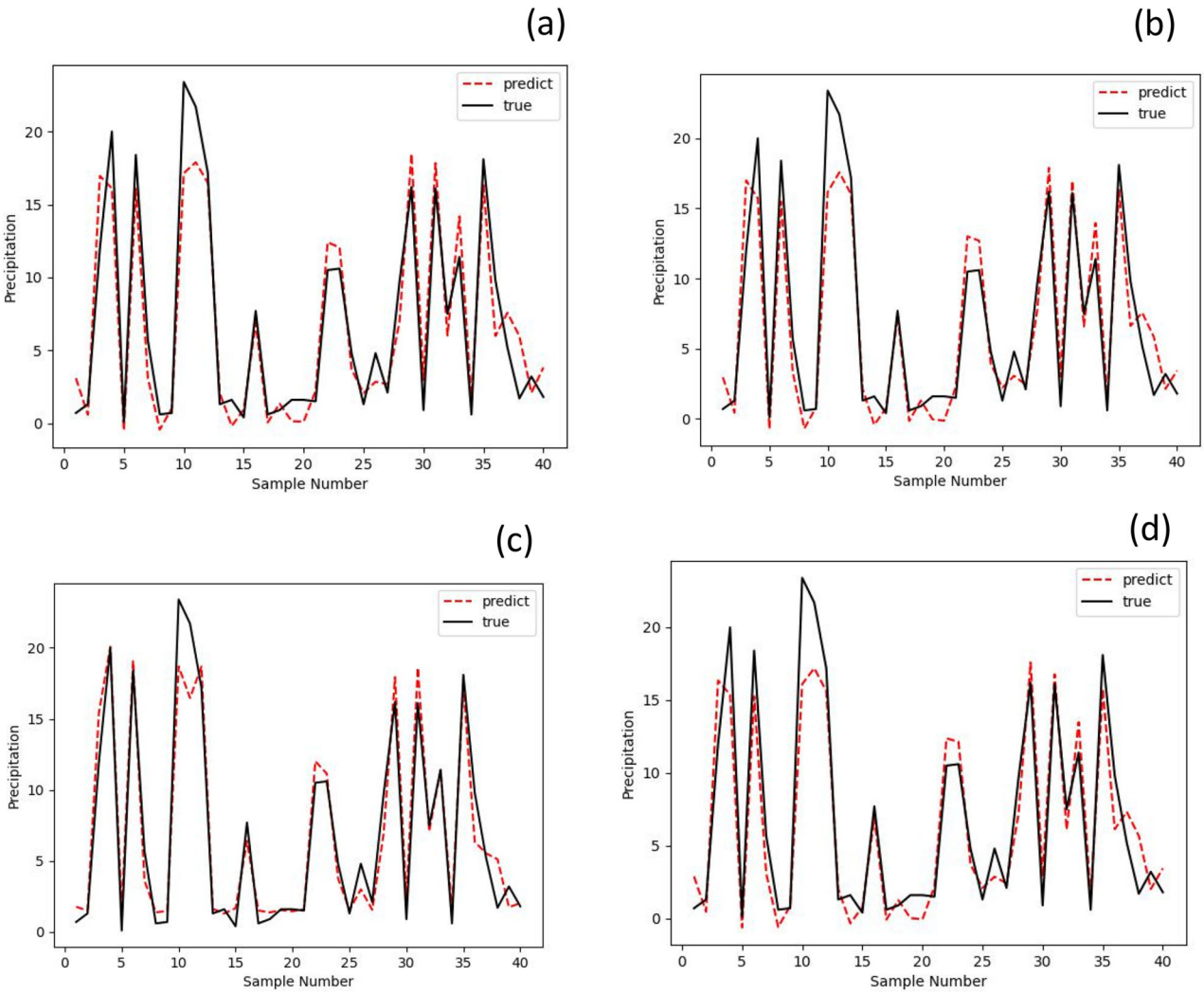

Fig. 10 Forecast results comparison of the LR, RF, SVR, BRR models for Lhasa in (a), (b), (c), (d) respectively

We also tested the significance of the regression model and regression coefficients for the 5 independent variables. In the linear regression model for the Zhengzhou region, an F-test was conducted on the regression equation. The results, evaluated at a significance level of 0.05, indicate a statistically significant linear relationship between $Q(X)$ and the independent variables x_1, x_2, x_3, x_4 and x_5 as a whole, confirming the significance of the regression equation. Furthermore, a t-test was performed on the regression coefficients, revealing that, under the significance level of 0.05, the independent variables x_1, x_2, x_3, and x_4 passed the significance test, while x_5 did not, suggesting that x_5 may not have a significant impact on $Q(X)$ for the Zhengzhou region. However, upon analyzing the linear regression model results in the remaining 8 regions, it was observed that x_5 was statistically significant in some of the eight regions. To ensure a consistency across the regions so that the results can be directly compared, x_5 was kept in the linear regression model for all nine regions.

The comparison of predicted values and true values based on the test set is described in Fig. 3. The vertical axis is rainfall, and the horizontal axis is serial number.

The linear regression model is implemented using the Linear Regression and SGD Regressor packages in Python.

3.2.2 Random forest model

Import the Random forest regressor package, use the preprocessed data set to train the initial model, and obtain a goodness of fit of 0.746 and mean squared error of 12.375 on the test set. To improve the predictive performance of the model and reduce generalization error, we use grid search to tune the model [36].

Grid search is an exhaustive search method that involves specifying parameter values. It optimizes the estimation

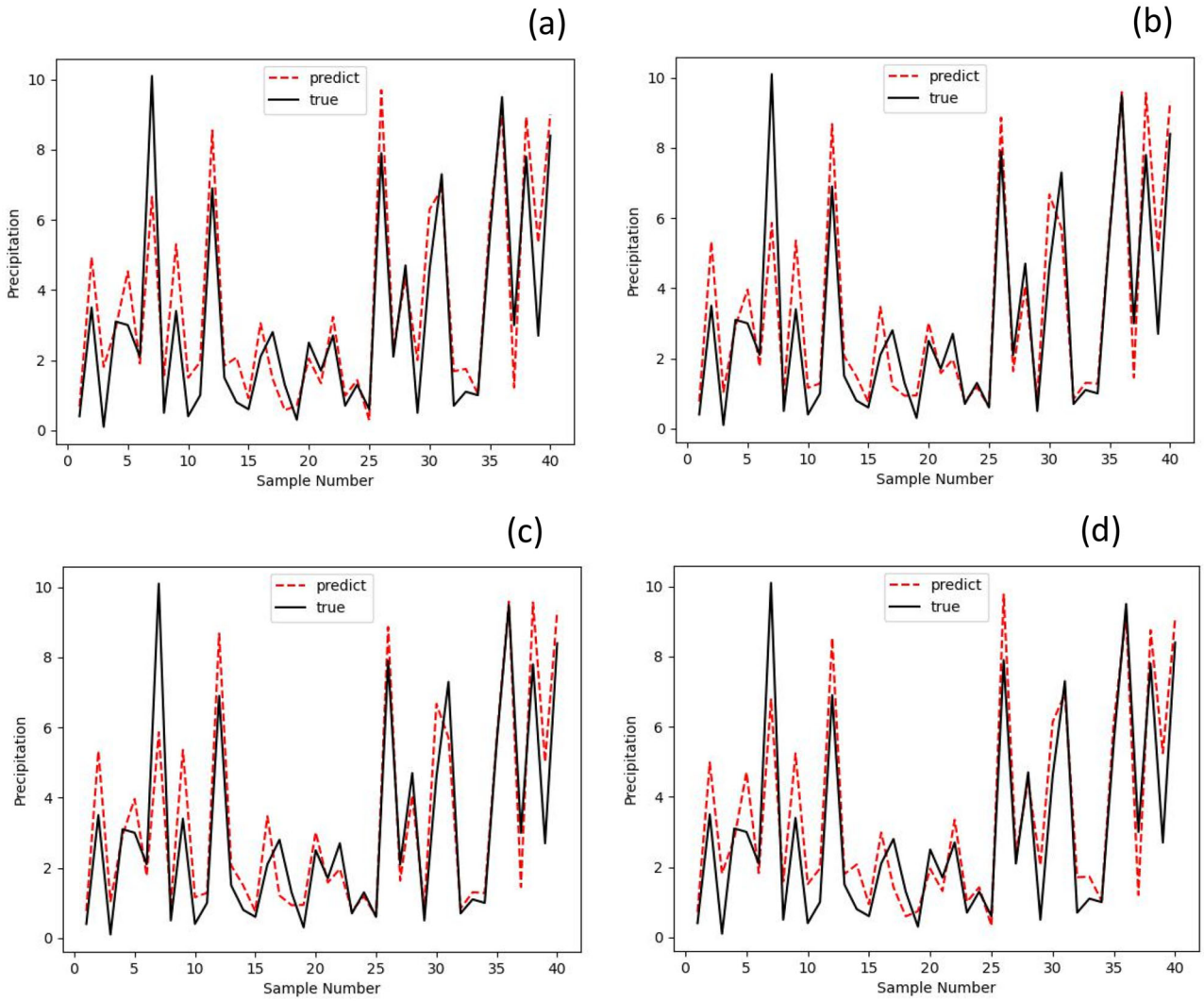

Fig. 11 Forecast results comparison of the LR, RF, SVR, BRR models for Lanzhou in (a), (b), (c), (d) respectively

function's parameters through cross-validation to obtain the optimal learning algorithm [37]. It involves creating a "grid" by listing all possible combinations of parameter values. After each training iteration, the best estimators are used to obtain the optimal combination [38]. This process is repeated, gradually narrowing down the parameter range, to obtain the best parameter combination.

Compared to traditional machine learning algorithms, random forest models have more complex hyperparameters, which can be roughly divided into three categories. The first category is the number of weak learners. When the number of weak learners increases, the complexity of the model also increases, and the generalization ability of the model initially increases but then decreases or remains unchanged, which means the learning ability of the model becomes stronger with the increase of n_estimatores, but at the same time, the risk of overfitting also increases. The second category is the structure of weak learners. The weak learner in the random forest model is decision tree, mainly including two parts: branching criterion and tree structure parameters. Generally, the more complex the structure of a single tree, the higher the overall complexity of the ensemble algorithm, the model is more prone to overfitting. Therefore, it is necessary to adjust the parameters appropriately to prune the decision tree and find the optimal structure for the decision tree. At the same time, when abnormal data appears in the training samples, the anti-jamming ability of the decision tree is poor, and the random forest algorithm to some extent enhances the anti-jamming ability of the model. The third category is the data used to train the weak learners. By controlling the random selection of data samples and feature variables, the risk of overfitting of the model is reduced, thereby improving the generalization ability of the model.

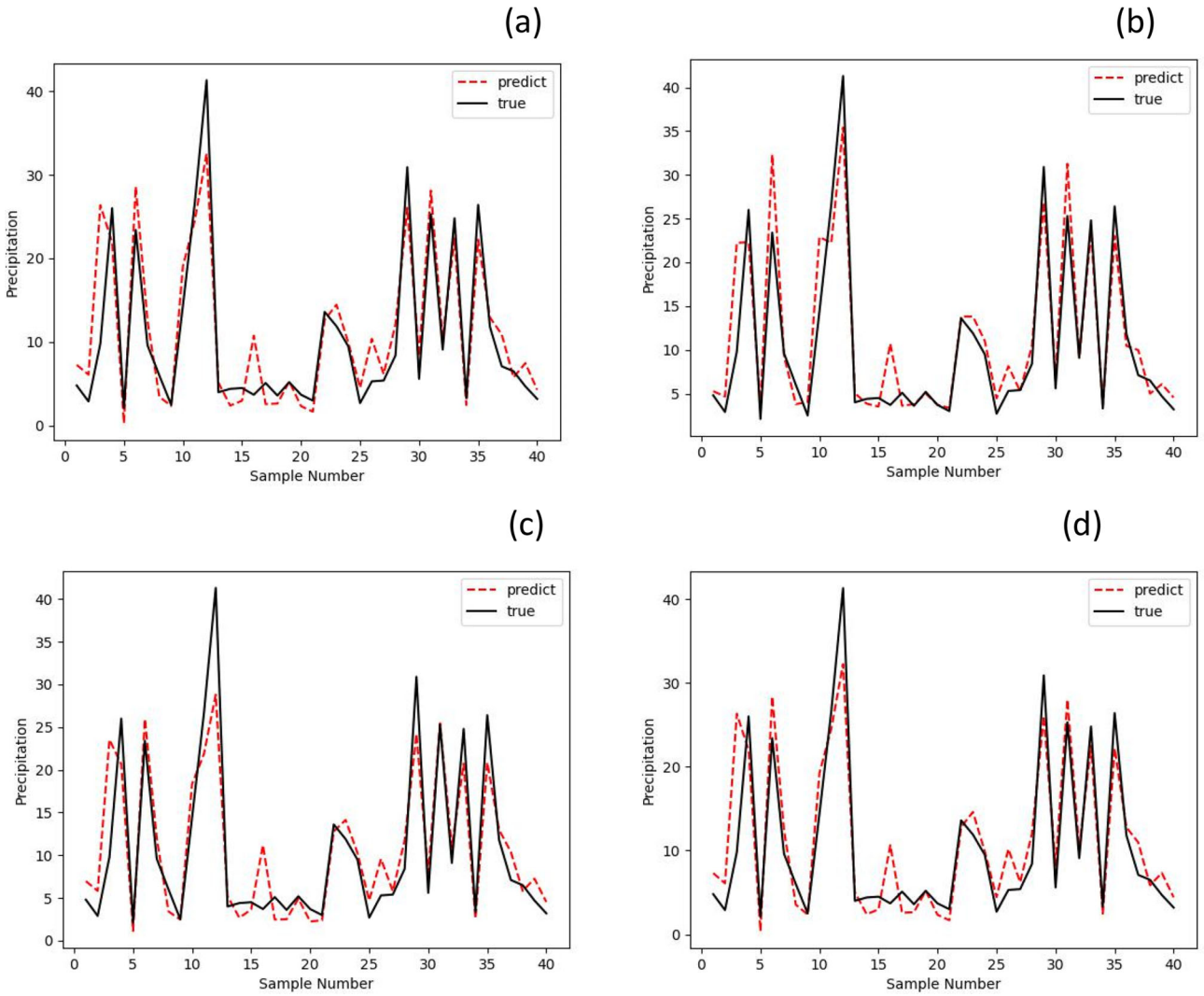

Fig. 12 Forecast results comparison of the LR, RF, SVR, BRR models for Chengdu in (a), (b), (c), (d) respectively

The adjustable parameters and parameter importance of random forest are presented in Table 3.

Three rounds of parameter optimization are conducted for this model. For the 7 adjustable parameters, three values are selected for each parameter in each round of model training, resulting in 2187 possible combinations for evaluation.

After each training, the optimal hyperparameter combination for the random forest model based on the Zhengzhou region is presented in Table 4.

The model's goodness of fit on the test set is 0.827, with a mean squared error of 8.389. Compared to the initial model, there is an improvement in the goodness of fit by 0.081 and a decrease in mean squared error by 3.986, indicating an improved fitting performance.

The predicted values and actual values are compared in Fig. 4.

3.2.3 Support vector regression model

Import the Linear SVR package, use the preprocessed data set to train the model, and test all kernel functions, which are 'linear', 'poly', 'rbf', 'sigmoid', and 'precomputed' in turn. By comparing the predicted results, the Linear SVR model has the best prediction performance, with a goodness of fit of 0.830 and mean squared error of 8.246 on the test set. The predicted values and actual values are displayed in Fig. 5. The vertical axis is rainfall, and the horizontal axis is sample number [39].

3.2.4 Bayesian ridge regression model

Import the linear_model package, call the linear_model. Bayesian Ridge function to train the model using the preprocessed data set, and tune the regularization strength

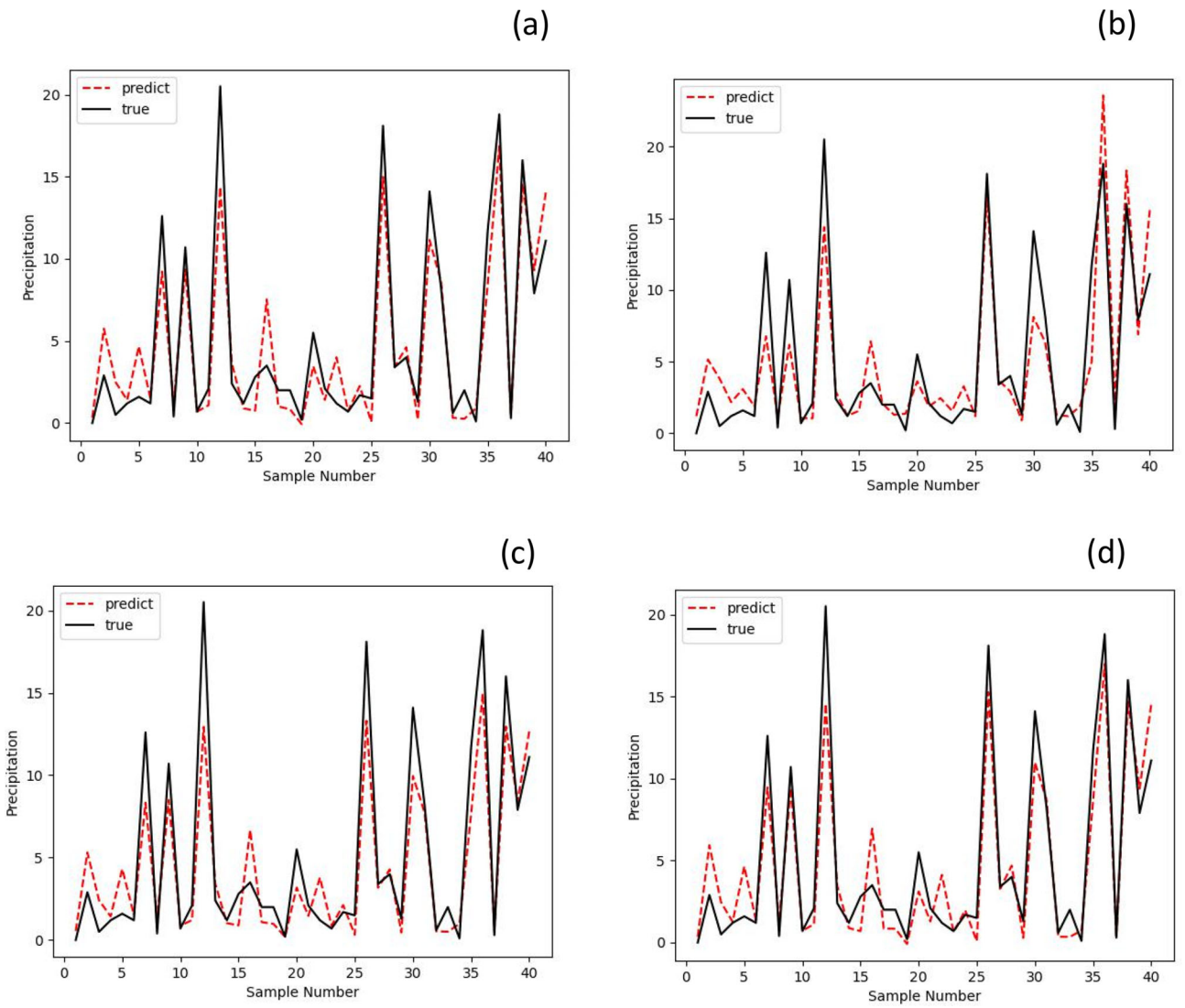

Fig. 13 Forecast results comparison of the LR, RF, SVR, BRR models for Jinan in (a), (b), (c), (d) respectively

parameter [40]. The final model had a goodness of fit of 0.841 and mean squared error of 7.703 on the test set. The predicted values and actual values are compared in Fig. 6.

Following the above steps, linear regression (LR), random forest regression (RFR), support vector regression (SVR), and Bayesian ridge regression are used to predict the rainfall in eight other regions, including Beijing. The comparison of predicted and actual results are presented in Figs. 7, 8, 9, 10, 11, 12, 13, 14.

3.3 Comparison of Predicted Results

Four machine learning models are used to predict rainfall in nine regions of China, and the results are summarized in Tables 5. Table 3 shows the adjusted R-squared values, and Table 6 describes the mean squared error (MSE). Table 7 presents the Relative Percentage Difference (RPD). The optimal model for each region based on both fitting performance, stability is shown in Table 8.

From Tables 2 and 3, it can be seen that based on the predicted results of rainfall in the nine regions, the random forest model has the best prediction performance, with an average adjusted R-squared (Adjusted R^2) of 0.801 for the nine regions. The SVR model is second, with an average adjusted R-squared of 0.785 for the nine regions, followed by the Bayesian ridge regression model and the linear regression model with an average adjusted R-squared 0.783 and 0.781, respectively. In addition, the mean MSE values of the four regression models are 11.319, 9.893, 11.191, and 11.245, respectively, which again demonstrates that random forest model outperforms other models.

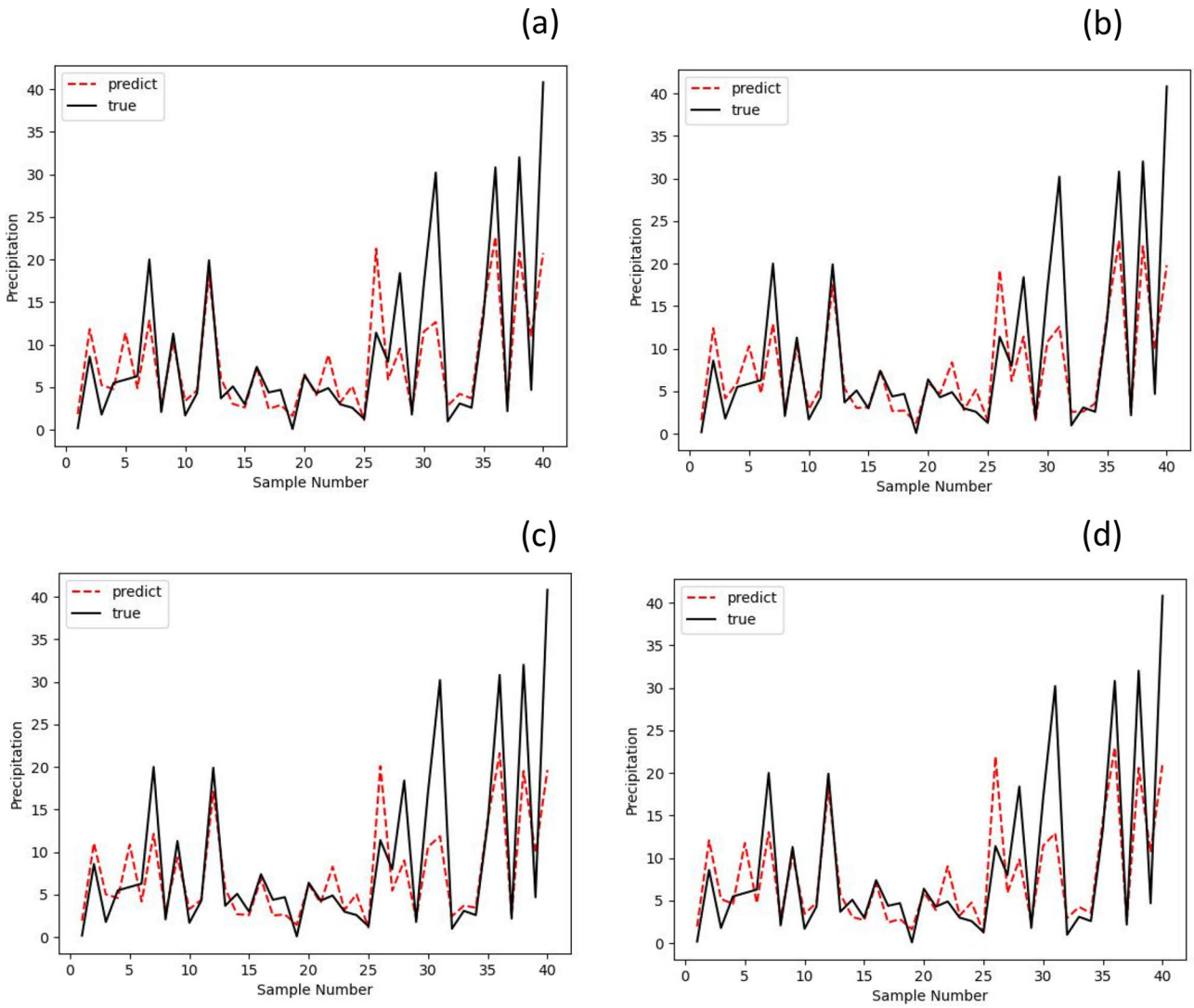

Fig. 14 Forecast results comparison of the LR, RF, SVR, BRR models for Xi'An in (a), (b), (c), (d) respectively

Table 5 Comparison of fitting performance of LR, RF, SVR, BRR models

No	Region	Regression Prediction Model of Rainfall(Adjusted R^2)			
		Linear Regression	Random Forest	Support Vector Regression	Bayesian Ridge Regression
1	Zhengzhou	0.836	0.827	0.830	0.841
2	Beijing	0.785	0.814	0.780	0.785
3	Hohhot	0.751	0.817	0.758	0.750
4	Kunming	0.782	0.837	0.776	0.782
5	Lhasa	0.888	0.914	0.884	0.891
6	Lanzhou	0.815	0.807	0.820	0.817
7	Chengdu	0.818	0.880	0.815	0.820
8	Jinan	0.706	0.701	0.693	0.709
9	Xi'an	0.650	0.615	0.706	0.653

Table 6 comparison of MSE of LR, RF, SVR, BRR models

No	Region	Regression Prediction Model of Rainfall(MSE)			
		Linear Regression	Random Forest	Support Vector Regression	Bayesian Ridge Regression
1	Zhengzhou	7.971	8.389	8.246	7.730
2	Beijing	7.873	6.828	8.059	7.875
3	Hohhot	3.040	2.236	2.954	3.047
4	Kunming	25.869	19.309	26.577	25.812
5	Lhasa	5.580	4.18	5.651	5.330
6	Lanzhou	1.421	1.485	1.385	1.410
7	Chengdu	16.702	10.994	16.980	16.517
8	Jinan	13.750	13.995	14.364	13.981
9	Xi'an	19.665	21.62	16.506	19.502

Table 7 Comparison of RPD of LR, RF, SVR, BRR models

No	Region	Regression Prediction Model of Rainfall(RPD)			
		Linear Regression	Random Forest	Support Vector Regression	Bayesian Ridge Regression
1	Zhengzhou	2.466	2.247	2.428	2.508
2	Beijing	2.161	2.193	2.133	2.158
3	Hohhot	2.001	2.379	2.032	2.002
4	Kunming	2.142	2.399	2.112	2.143
5	Lhasa	2.958	3.859	2.941	3.028
6	Lanzhou	2.320	2.298	2.357	2.337
7	Chengdu	2.346	2.594	2.322	2.354
8	Jinan	2.876	2.208	2.501	2.923
9	Xi'an	1.702	1.754	1.632	1.709

Table 8 Optimal rainfall prediction model for each region

No	Region	Model with Optimal Fitting Effect	Model with Optimal Degree of Change	Model with Optimal Stability
1	Zhengzhou	Bayesian Ridge Regression	Bayesian Ridge Regression	Bayesian Ridge Regression
2	Beijing	Random Forest	Random Forest	Random Forest
3	Hohhot	Random Forest	Random Forest	Random Forest
4	Kunming	Random Forest	Random Forest	Random Forest
5	Lhasa	Random Forest	Random Forest	Random Forest
6	Lanzhou	Support Vector Regression	Support Vector Regression	Support Vector Regression
7	Chengdu	Random Forest	Random Forest	Random Forest
8	Jinan	Bayesian Ridge Regression	Linear Regression	Bayesian Ridge Regression
9	Xi'an	Support Vector Regression	Support Vector Regression	Random Forest

As described in Table 5, all the constructed models have good stability, except Xi 'an, the RPD value of the other eight regions reaches 2.0, which is a high stability level.

Out of the nine regions analyzed, the random forest model exhibits superior performance in five regions. However, in Xi'an, the random forest model significantly deviates from the actual values, indicating that it is not suitable for predicting rainfall in that region [41]. The SVR model, on the other hand, demonstrates higher accuracy in predicting rainfall for that region. In Zhengzhou and Jinan, the Bayesian ridge regression model provided predictions closer to actual values [42]. Similarly, in Lanzhou, the SVR model achieves better accuracy. Although difference in adjusted R-squared values across the four models are relatively small for these three regions, the random forest model showed a degree of universality. The random forest model provides the optimal stability model for six regions based on the stability index (RPD), indicating its ability to obtain table predictions [43].

3.4 Prediction model analysis

Based on the prediction results, the random forest model demonstrates higher accuracy compared to other models in rainfall prediction. That is probably due to the following characteristics of the model:

1. The random forest model utilizes ensemble learning, employing bagging methods during the sampling process to extract training subsets. This ensures independence among the individual decision trees and effectively avoids overfitting.
2. When each decision tree is constructed in the random forest model, not all attributes are involved in the node splitting process of every tree. Instead, a random selection of a few attributes is used as feature indicators for attribute evaluation. The introduction of randomness guarantees the diversity among the sub-models. The greater the differences between the sub-models, the better the effect of model fusion, effectively enhancing the model's tolerance to noise and outliers. Therefore, compared to other algorithms, the random forest algorithm demonstrates higher predictive accuracy and stronger model generalization ability in forecasting future rainfall. Moreover, the robust predictive performance of the random forest model also holds value for exploring other regression problems in real-world applications.

4 Conclusions

Through the analysis of rainfall data from nine typical regions in China, four regression models are constructed to predict rainfall, with the grid search technique used to optimize their performance. Based on the experimental results, the following conclusions can be drawn.

1. The four machine-based regression models yield good results overall. Among these models, the random forest model stand out with its high accuracy in predicting rainfall, achieving a goodness of fit of 0.8 or higher for seven regions and demonstrating good model stability with a small mean square error value. However, for some cities in China North regions with significant fluctuations in rainfall, such as Lanzhou and Xi'an, the support vector regression model exhibits better fitting and demonstrates unique advantages in solving small sample and nonlinear problems.
2. (2) The prediction accuracy of the random forest model is heavily influenced by its parameters, and the use of grid search method to finetune these parameters results in significant improvement in the model's performance. However, it should be noted that this approach requires considerable training time to explore all points on the grid. Further refinement of the grid search method is needed to optimize the model and improve its training speed and prediction efficiency.

This study successfully utilizes a combination of causes of the July 20, 2021 Zhengzhou flood and necessary conditions for heavy rainfall to accurately select five variables, resulting in precise rainfall prediction with minimal variables. It significantly reduces the data collection difficulty. Furthermore, the study employs the grid search method to optimize the model parameters of the initial four regression models, which substantially enhance the prediction accuracy and generalization ability of the models. By comparing the rainfall prediction outcomes of nine typical regions in China, the study evaluates the fitting effect and stability performance of the four machine learning regression models. The results confirm that the random forest model is more appropriate for predicting rainfall than other forecasting models. This extensive research is valuable for machine learning researchers in developing better methods for predicting rainfall with higher accuracy in the future.

In recent years, the China Meteorological Administration has adopted artificial intelligence for rainfall forecasting. As a core field of artificial intelligence, machine learning has demonstrated its importance in predicting extreme weather and other fields. However, some limitations have been identified in these methods, such as issues with data quality, feature selection, and model selection. Improvements in monitoring meteorological data and refining models can lead to more accurate rainfall predictions, enabling more rational management and integrating utilization of water resources and

providing scientific guidance for the prevention of flood disasters.

Acknowledgements The author would like to acknowledge the National Natural Science Foundation of China (NO. 11972327) for the financial support for this paper. They also would like to thank the anonymous reviewers for their helpful and kind suggestions.

Author contributions LP has made the substantial contributions to the conception and design of the work, and interpretation of the predication results, revised critically the important content, approved the final version to be published; YW has carried out the total analysis and predication, and drafted the manuscript; JW has collected the data for analysis and prediction, checked and polished the draft.

Funding This work was supported by the National Natural Science Foundation of China (NO. 11972327).

Availability of data and materials The data sets used or analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Conflict of Interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. The authors declare the following Financial interests/personal relationships which may be considered as potential competing interests: No.

References

1. Chen X, Chen Y, Shi J (2020) Modeling and prediction of rainfall radar echo data based on Machine learning. J Nanjing Univ Inf Sci Technol,20,12(4):483494
2. Bouaziz M, Medhioub E, Csaplovisc E.(2021) A machine learning model for drought tracking and forecasting using remote precipitation data and a standardized precipitation index from arid regions. J Arid Environ
3. (2008) Based on least squares support vector machine (SVM) rainfall prediction [J]. The people of the Yangtze River, 9 (1): 2931. https://doi.org/10.16232/j.carolcarrollnki.10014179.2008.19.001
4. Lange H, Sippel S (2020) Machine learning applications in hydrology. Forestw Interact, 233257
5. Leung CKS, MacKinnon RK, Wang Y (2014) A machine learning approach for stock price prediction. In: Proceedings of the 18th international database engineering & applications symposium, pp 274277
6. Li S, Bai Y (2022) Book review: text as data: a new framework for machine learning and the social sciences
7. Ahmed K, Sachindra DA, Shahid S, Iqbal Z, Nawaz N, Khan N (2020) Multimodel ensemble predictions of precipitation and temperature using machine learning algorithms. Atmos Res 236:104806
8. Gocic M, Shamshirband S, Razak Z, Petković D, Ch S, Trajkovic S (2016) Longterm precipitation analysis and estimation of precipitation concentration index using three support vector machine methods. Adv Meteorol, 2016.
9. Haiden T, Kann A, Wittmann C, Pistotnik G, Bica B, Gruber C (2011) The integrated nowcasting through comprehensive analysis (INCA) system and its validation over the Eastern Alpine region. Weather Forecast 26(2):166183
10. Alizamir M, Kim S, Kisi O, ZounematKermani M (2020) A comparative study of several machine learning based nonlinear regression methods in estimating solar radiation: case studies of the USA and Turkey regions. Energy 197:117239
11. Xu L, Yu J. (2020) Different optimizer under Gaussian noise on the study of the influence of the LR performance. Comput Technol Dev, 712
12. Yao W, Li L (2014) A new regression model: modal linear regression. Scand J Stat 41(3):656671
13. Liu Y, Wang Y, Zhang J (2012) New machine learning algorithm: random forest. In: Inf Comput Appl Third Int Conf, ICICA 2012, Chengde, China, September 1416, 2012. Proceedings 3. Springer, pp 246252
14. Yisen W, Shutao X (2018) Review of stochastic forest algorithm for ensemble learning. Inf Commun Technol:4955. (in Chinese). https://doi.org/10.3969/j.issn.16741285.2018.01.009.
15. Jain N, Jana PK (2023) LRF: A logically randomized forest algorithm for classification and regression problems. Expert Syst Appl 213:119225
16. Zhang W, Wu C, Li Y, Wang L, Samui P (2021) Assessment of pile drivability using random forest regression and multivariate adaptive regression splines. Georisk: Assess Manage Risk Eng Syst Geohazards, 15(1), 2740.
17. Xue L, Liu Y, Xiong Y, Liu Y, Cui X, Lei G (2021) A datadriven shale gas production forecasting method based on the multiobjective random forest regression. J Petrol Sci Eng 196:107801
18. Liang C, Jinhong W, Tao H, et al (2018) Regional transportation carbon based on SVR prediction research. J Transp Syst Eng Inf Technol 19(2):13 to 19. https://doi.org/10.16097/j.carolcarrollnki.10096744. 2018.02.003
19. Xu Weiya Xu, Wei YL (2021) Deformation prediction of toppling deformed slope based on LMBP and SVR. J Hohai Univ (Nat Sci) 49(1):6469
20. Wang YG, Wu J, Hu ZH, McLachlan GJ (2023) A new algorithm for support vector regression with automatic selection of hyperparameters. Pattern Recogn 133:108989
21. Kurani A, Doshi P, Vakharia A, Shah M (2023) A comprehensive comparative study of artificial neural network (ANN) and support vector machines (SVM) on stock forecasting. Ann Data Sci 10(1):183208
22. Michimae H, Emura T (2022) Bayesian ridge estimators based on copulabased joint prior distributions for regression coefficients. Comput Statistics 37(5):27412769
23. Imane M, Aoula ES, Achouyab EH (2022) Using Bayesian ridge regression to predict the overall equipment effectiveness performance. In: 2022 2nd international conference on innovative research in applied science, engineering and technology (IRASET). IEEE, pp 14
24. Na MH, Cho WH, Kim SK, Na IS (2022) Automatic weight prediction system for Korean cattle using Bayesian ridge algorithm on RGBD image. Electronics 11(10):1663
25. Degener A (2022) Prediction of appropriate L2 regularization strengths through Bayesian formalism
26. Cheng K, Lu Z (2021) Adaptive Bayesian support vector regression model for structural reliability analysis. Reliab Eng Syst Saf 206:107286
27. Gupta S, McFarquhar GM, O'Brien JR et al (2022) Factors affecting precipitation formation and precipitation susceptibility of marine stratocumulus with variable above-and below-cloud aerosol concentrations over the Southeast Atlantic. Atmos Chem Phys 22(4):2769–2793
28. Bailey A, Aemisegger F, Villiger L et al (2023) Isotopic measurements in water vapor, precipitation, and seawater during EUREC 4 A. Earth Syst Sci Data 15(1):465–495
29. Ricciotti JA, Cordeira JM (2022) Summarizing relationships among landfalling atmospheric rivers, integrated water vapor

transport, and California watershed precipitation 1982–2019[J]. J Hydrometeorol 23(9):1439–1454
30. Czajka B, Barthlott C, Kohler M et al (2023) Analysis of the impact of selected sources of uncertainty on precipitation simultaions of summer convection over Central Europe[R]. Copernicus Meet
31. Sun D, Xu J, Wen H, Wang D (2021) Assessment of landslide susceptibility mapping based on Bayesian hyperparameter optimization: a comparison between logistic regression and random forest. Eng Geol 281:105972
32. Torbeck L (2010) When to use percent relative standard deviation—and how to do so correctly. Pharm Technol 34(1):263
33. Battey HS, Reid N (2021) Inference in highdimensional linear regression. arXiv preprint arXiv:2106.12001
34. Hongzhi Y, Baorong Z (2018) Normal equations based on machine learning linear regression ana*lysis*. J Geek, https://doi.org/10.3969/j.issn.1672528X.2018.07.171
35. Arora S, Li Z, Panigrahi A (2022) Understanding gradient descent on the edge of stability in deep learning. In: International conference on machine learning. PMLR, pp 9481024
36. Belete DM, Huchaiah MD (2022) Grid search in hyperparameter optimization of machine learning models for prediction of HIV/AIDS test results. Int J Comput Appl 44(9):875886
37. Belete DM, Huchaiah MD (2022) Grid search in hyperparameter optimization of machine learning models for prediction of HIV/AIDS test results[J]. Int J Comput Appl 44(9):875–886
38. Afzal A, Aabid A, Khan A, Khan SA, Rajak U, Verma TN, Kumar R (2020) Response surface analysis, clustering, and random forest regression of pressure in suddenly expanded highspeed aerodynamic flows. Aerosp Sci Technol 107:106318
39. Pisner DA, Schnyer DM (2020) Support vector machine. In *Machine learning*. Academic Press, pp 101121
40. Santos CFGD, Papa JP (2022) Avoiding overfitting: a survey on regularization methods for convolutional neural networks. ACM Comput Surv (CSUR) 54(10s):125
41. Cervantes J, GarciaLamont F, RodríguezMazahua L, Lopez A (2020) A comprehensive survey on support vector machine classification: applications, challenges and trends. Neurocomputing 408:189215
42. Sheykhmousa M, Mahdianpari M, Ghanbari H, Mohammadimanesh F, Ghamisi P, Homayouni S (2020) Support vector machine versus random forest for remote sensing image classification: a metaanalysis and systematic review. IEEE J Sel Top Appl Earth Obs Remote Sens 13:63086325
43. Pirone D, Cimorelli L, Del Giudice G, Pianese D (2023) Short-term rainfall forecasting using cumulative precipitation fields from station data: a probabilistic machine learning approach. J Hydrol 617:128949