Medium- and Long-Term Precipitation Forecasting Method Based on Data Augmentation and Machine Learning Algorithms

Tiantian Tang, *Member, IEEE*, Donglai Jiao, Tao Chen, and Guan Gui, *Senior Member, IEEE*

***Abstract*—Accurate medium and long-term precipitation forecasting plays a vital role in disaster prevention and mitigation and rational allocation of water resources. In recent years, there are various methods for medium- and long-term precipitation forecasting based on machine learning algorithms. However, machine learning has a high demand for the size of sample data. Therefore, this article proposes a data augmentation algorithm based on the K-means clustering algorithm and synthetic minority oversampling technique (SMOTE), which can effectively enhance sample information. Besides, through constructing random forest (RF), extreme gradient boosting (XGB), recurrent neural network (RNN), and long short-term memory (LSTM) are, respectively, constructed as the models to forecast monthly grid precipitation of the Danjiangkou River Basin. This study aims to improve the accuracy of medium- and long-term precipitation forecasting. The main results are the following two aspects: 1) in most years, the anomaly correlation coefficient and Pg score of SMOTE-km-XGB and SMOTE-km-RF exceed that of XGB and RF; furthermore, compared with the other three methods, SMOTE-km-XGB method is more suitable for precipitation forecasting in the studied basin in this article; and 2) the forecasting results of two deep learning methods (RNN and LSTM) show that the sample data processed by the K-means clustering algorithm and SMOTE data augmentation algorithm have not achieved considerable results in deep learning. This study improves the accuracy of precipitation forecast by expanding and balancing the information of sample data, and provides a new research idea for improving the accuracy of medium- and long-term hydrological forecasting.**

***Index Terms*—Extreme gradient boosting (XGB), K-means, long short-term memory (LSTM), machine learning (ML), medium- and long-term precipitation forecasting, random forest (RF), recurrent neural network (RNN), synthetic minority oversampling.**

I. Introduction

MEDIUM- and long-term precipitation forecasting is an important part of hydrological science, and always plays a key role in flood control, disaster reduction, and the comprehensive utilization of water resources. However, with the growth of the forecast period, the influencing factors of medium- and long-term precipitation forecasting increasingly lead to more uncertainties in forecasting and cause a decrease in the forecasting accuracy. This has always been a difficult point in the field of precipitation forecasting. Therefore, the in-depth study of the medium- and long-term forecasting theory and methods not only has important scientific value for enriching and developing the precipitation forecasting theory but also has important practical significance for disaster reduction and prevention and social and economic sustainable development [1], [2].

However, medium- and long-term precipitation forecasting, in terms of providing the total amount of precipitation in a certain period of time in the future, is considered to be one of the most difficult challenges in global climate models because its forecast accuracy is influenced by many factors, such as precipitation location, duration, frequency and intensity, orography, and land use [3], [4]. Traditional medium- and long-term forecasts mainly use statistical methods, dynamic methods, and a combination of statistics and dynamics to produce forecasts. In recent years, with the rapid development of the global satellite remote sensing, cloud computing, and cloud storage technology, the possibility and stability of the operation of the general circulation models (GCMs) have been further improved, and the GCMs have gradually replaced the classical statistical model and become the main tool to release real-time monthly seasonal scale forecast information for major meteorological–hydrological forecasting centers around the world [5], [6]. At the same time, with the rapid development of computer technology, the machine learning (ML) method based on Big Data mining technology has been gradually applied to medium- and long-term precipitation forecasts because of its high generalization ability and strong robustness. The medium- and long-term precipitation forecasting methods based on ML mainly build correlations between precipitation and predictors. There are many factors affecting

Manuscript received September 1, 2021; revised December 4, 2021 and December 30, 2021; accepted December 31, 2021. Date of publication January 5, 2022; date of current version January 20, 2022. This work was supported in part by the Natural Science Foundation of Jiangsu Province under Grant BK20191384, in part by the China Postdoctoral Science Foundation under Grant 2019M661896, in part by the Summit of the Six Top Talents Program of Jiangsu under Grant XYDXX-010, in part by the Program for High-Level Entrepreneurial and Innovative Team under Grant CZ002SC19001, in part by the Foundation of Nanjing Vocational College of Information Technology under Grant YK20210501, and in part by Natural Science Research Start-up Foundation of Recruiting Talents of Nanjing University of Posts and Telecommunications (Grant No.NY221154). *(Corresponding author: Guan Gui.)*

Tiantian Tang and Donglai Jiao are with the School of Geographic and Biologic Information, Nanjing University of Posts and Telecommunications, Nanjing 210023, China (e-mail: tangtt@njupt.edu.cn; jiaodonglai@njupt.edu.cn).

Tao Chen is with the Nanjing Hydraulic Research Institute, Hydrology and water resources department, Nanjing 210029, China (e-mail: tchen@nhri.com).

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

Digital Object Identifier 10.1109/JSTARS.2022.3140442

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

precipitation. In addition to the well-known meteorological-climatic factors, predictor factors such as soundings [7], local-scale effects, such as the mountain valley circulation effect [8], atmospheric environment [9], and satellite imagery information observed in the preconvective environment, vary greatly [10].

In recent years, with the continuous progress of science and technology and the rapid development of information technology, the amount of data in human production and life has increased geometrically, and gradually developed from bytes to gigabytes, terabytes, petabytes, and even yottabytes. Big data technology came into being and gradually became the focus of scientific research. However, the diversity and mass of Big Data is both a blessing and a challenge for the medium- and long-term hydrological forecasting field. With the vigorous promotion of water conservancy informatization, both the observation data of surface meteorological stations and the observation data based on satellite remote sensing have made great progress in recent years in the "quality" and "quantity" of data with strong temporal and spatial attributes and have gradually ushered in the "Big Data era" of hydrology. These new information sources enrich our understanding and enhance our modeling ability. However, in the face of data information from various sources and structures, how to use data mining technology to explore its intrinsic value and connection from massive meteorological and hydrological information is the frontier research field of developing hydrological forecasting [11]. More importantly, without advanced technology, we may not even realize what kind of hidden and abstract information can be extracted or the limitation of the accuracy of this extraction, which leads to the insufficient use of available data [12].

To solve the aforementioned problems, some data-driven methods, such as ML methods, have been proposed and widely used in the face of complex and large variable relations to help us extract useful information from the growing data [13]–[17]. Therefore, combining advanced ML methods with traditional hydrological methods to realize medium- and long-term precipitation forecasting is not only an extension and improvement of the traditional precipitation forecasting but also a great progress in the interdisciplinary development of hydrological work. ML methods can be divided into shallow ML and deep learning according to the depth of the network. Shallow ML methods are widely used in hydrology, such as random forest (RF) [18]–[20], support vector machine (SVM) [21]–[23], extreme gradient boosting (XGB) [24]–[26], light gradient boosting machine (LGB) [27]–[29], etc. Deep learning methods such as recurrent neural network (RNN) [30]–[32] and long short-term memory (LSTM) [33] are not widely used in hydrology. In addition, in the field of artificial intelligence, rather than whether the ML model can show better performance, the ability to debug the model or the model itself, the more important decisive factor is often the data volume used to build the model. In the field of hydrology, the length of hydrological series is limited, and sometimes it is difficult to meet the number of samples needed by ML to build a better model, which greatly affects the forecasting accuracy. Therefore, it is critical to expand the hydrological series data within a reasonable range to meet the basic requirements of the ML model modeling to make the ML model play a better role in medium- and long-term precipitation forecasting. Data augmentation technology is a common technical method in the field of ML, which is used to expand sample data information, and has made outstanding achievements in the field of biomedical image segmentation [34], environmental sound classification [35], text recognition, and image recognition [36]–[38]. According to a large number of studies, the commonly used data augmentation methods include time stretching (TS) [39], synthetic minority oversampling technique (SMOTE) [40], linear prediction cepstral coefficients (LPCC) [41], etc. However, data augmentation is rarely used in hydrology.

Based on the aforementioned background, taking the Danjiangkou River Basin as the study area, this article constructs a data augmentation algorithm based on the K-means clustering algorithm and SMOTE to expand the precipitation series. In the meantime, taking the augmented sample data as input, two shallow ML models (RF, XGB) and two deep learning models (RNN, LSTM) are constructed to compare the prediction results before and after the expansion of precipitation data. In addition, the differences, advantages, and disadvantages of prediction results between shallow ML and deep learning models are discussed in depth.

II. Proposed Method

A. Data Augmentation

ML can be roughly divided into supervised learning and unsupervised learning according to the types of supervision. Supervised learning requires that the training data be marked and that the computer can identify the marked sample data by using specific patterns. Supervised learning can be divided into two categories: classification and regression. Classification consists of training a machine to classify a set of data. For example, in hydrological work, we divide floods into different grades, which can be regarded as categories. By training a computer with this marked (classified) flood data, we build an ML model that can accurately judge which category (grade) the flood belongs to for new flood sample data. Regression is defined as training a machine to predict the future according to the previously marked data. For example, when we perform precipitation forecasting, precipitation data are regarded as the label of sample data, and atmospheric circulation factors and other influencing factors are taken as characteristics. The machine can learn potential laws from the sample data after the machine learning model is trained. This model can forecast precipitation according to the input of new atmospheric circulation factors. Unsupervised learning [42] analyzes the inherent characteristics and structure of data by learning a large number of unlabeled data. The main methods of unsupervised learning are clustering and dimension reduction. Clustering refers to grouping data according to their characteristics. The grouping and classification algorithms mentioned here are different. The groups of classification algorithms are artificially defined, while the groups in clustering algorithms are computer defined. For example, in the aforementioned classification example, first, the floods are artificially divided into several categories to classify the new floods; in clustering, instead of artificially dividing the floods

into several categories, the training samples are automatically divided into different categories by machines. Dimension reduction finds the common points among data to reduce the variables of datasets and reduce the occurrence of redundancy. Supervised learning and unsupervised learning have their own advantages and disadvantages. At present, supervised learning is the most commonly used method for hydrological forecasting and has achieved good results. Using an unsupervised learning method to cluster sample data can avoid the influence of subjective factors on hydrological events and make the machine run more objectively, thus enhancing the credibility of forecast results. Therefore, in this study, hydrological data are clustered first to objectively classify them into different categories, and then, the data expansion method is used to expand the few samples. The expansion leads to a more balanced distribution and makes it easier to achieve better results in model forecasting.

B. K-Means Clustering Algorithm

The K-means algorithm is a typical unsupervised ML algorithm that is used to solve clustering problems. Because of its simple and rapid implementation, it has become one of the top ten classical data mining algorithms. The basic ideas and steps of the K-means algorithm include the following.

1) Step 1: Sample data without labels constitute a sample set $D = \{x^{(1)}, x^{(2)}, \ldots, x^{(n)}\}$. First, the sample set is divided into K classes, randomly select cluster centers to form $U = \{u_1, u_2, \ldots, u_k\}$.

2) Step 2: Traverse the sample set to calculate the distance from $x^{(i)}$ $(i = 1, 2, \ldots, n)$ to $u_1, u_2, \ldots, u_k$, and $x^{(i)}$ divide into this category when the cluster center point u_j $(1 \le j \le k)$ with the shortest distance appears.

3) Step 3: Traverse $u_1, u_2, \ldots, u_k$ and move the new location of the cluster center to the mean value of this category. That is $u'_j = (1/c)\sum_{d=1}^{c} x^{(d)}$, where $j = 1, 2, \ldots, k$, where $x^{(d)}$ represents the samples belonging to u'_j category, and c is the number of training sample points in this category.

4) Step 4: Repeat *S*tep 2 until the cluster center is no longer changed.

5) Step 5: Finally, the distance between the samples in the same class and the center of the samples is the closest, which is, the samples in the same class have high similarity, that is, the sum of squares between the samples in the same class and the center of the cluster is the smallest

$$\min \sum_{i=1}^{n} \sum_{j=1}^{k} \|x^{(i)} - u_{c^{(i)}}\|^2. \tag{1}$$

C. Proposed SMOTE Algorithm

The SMOTE algorithm was proposed by Chawla *et al.* in 2002. By artificially synthesizing sample data, the problem of unbalanced data and too few samples can be solved. The main principle of SMOTE is to linearly interpolate the training samples, generate an appropriate number of samples according to the oversampling rate, expand and augment the datasets of fewer samples, and then, train the learner with the new training sample set, thus improving the accuracy of the ML model. The SMOTE data augmentation algorithm, as an oversampling method that can effectively deal with unbalanced data, has been widely and maturely applied in fraud detection and risk control identification fields [43], [44], but has rarely been applied in hydrology. Therefore, this article takes this data augmentation algorithm as an exploratory preliminary attempt to broaden new development methods and ideas for improving hydrological forecasting accuracy. The principle of the SMOTE algorithm is as follows.

1) Step 1: The number of multiclass samples in the sample set is N_+, and the number of small-class samples is N_-. Calculate the imbalance degree IR and oversampling rate K of the original dataset, which are expressed, respectively, as

$$IR = \frac{N_+}{N_-} \tag{2}$$

$$K = \lfloor IR \rfloor. \tag{3}$$

2) Step 2: For each minority sample x_i, calculate the Euclidean distance with other minority samples, and find k nearest neighbors (the Euclidean distance is the smallest), where k is generally 5.

3) Step 3: According to the oversampling rate k, randomly select k samples with returns from K nearest neighbors, record them as $\bar{x}_i$ $(i = 1, 2, \ldots, K)$, and calculate $(x_i - \bar{x}_i)$.

4) Step 4: Use the following formula to synthesize new sample x^i_{new}:

$$x^i_{\text{new}} = x + \text{rand}(0, 1)(x_i - \bar{x}_i), \tilde{} I = 1, 2, \ldots, K. \tag{4}$$

Circulate the aforementioned steps to synthesize new samples artificially. The diagram of the proposed SMOTE algorithm is shown in Fig. 1.

D. Data Augmentation Algorithm Based on Combination of K-Means and SMOTE

In this article, first, the original sample datasets are clustered into three categories by using the K-means algorithm so that they can be objectively classified into different categories. Then, the SMOTE data augmentation algorithm is used to balance the unbalanced data among sample categories to achieve data balance and augmentation. Fig. 2 shows the process of the data augmentation algorithm.

III. ML Methods

The hierarchical structure of ML algorithms can be divided into shallow ML and deep learning. Shallow ML develops rapidly and maturely, which successfully solves the problems of low artificial efficiency and strong subjectivity. Deep learning is a new technology. It has a deep network structure and can realize multilayer and step-by-step extraction. Because of its powerful network performance, it is widely used in many fields. Through the development of the ML theory in medium- and long-term hydrological forecasting, shallow ML is mainly based on the application of decision tree models, while the application of deep learning is limited, and the forecasting accuracy of the two is indistinguishable. Therefore, this article considers applying

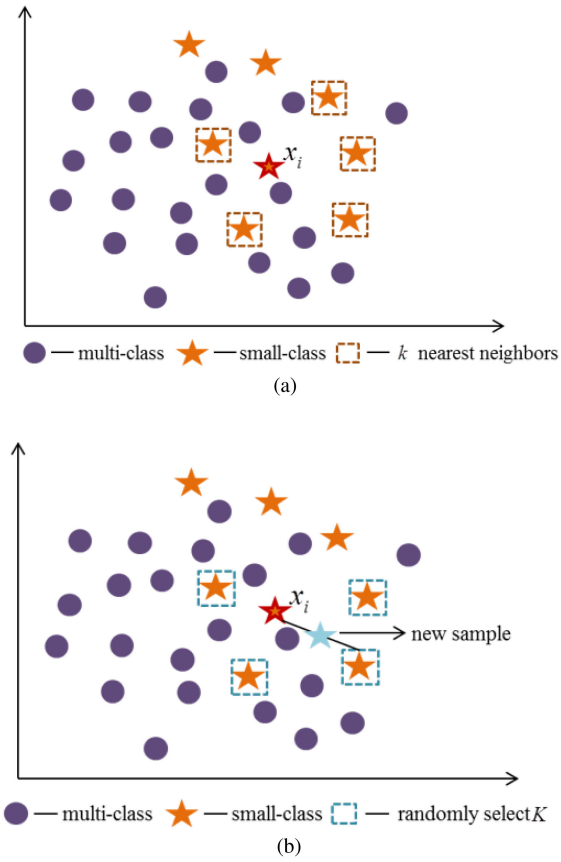

Fig. 1. Diagram of the proposed SMOTE algorithm. (a) k-nearest neighbor sample calculation. (b) New sample synthesis.

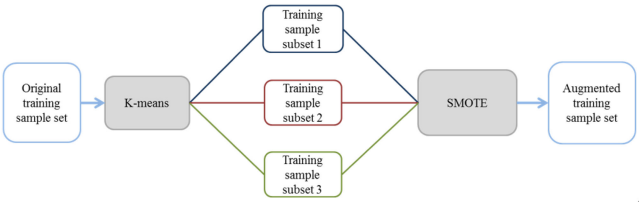

Fig. 2. Diagram of the data augmentation algorithm process.

two shallow ML models (RF, XGB) and two deep learning models (RNN, LSTM) to medium- and long-term precipitation forecasting for depth discussion and comparative analysis. The principles of various models are as follows.

A. Random Forest (RF)

Random Forests (RF) is an ML algorithm combining the Bagging ensemble learning theory [45] and random subspace method [46]. It uses boostrap technology to sample the original samples and generate multiple training samples. Each subset of training samples is randomly selected by the random subspace method to construct a decision tree, and finally, the optimal result is selected by voting or averaging. A large number of studies show that RF can effectively overcome the problems of noise and overfitting, and has higher accuracy in forecasting (O'Neil).

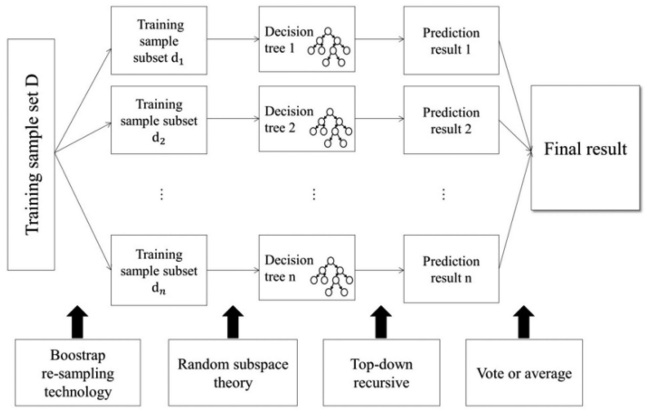

Fig. 3. Main structure of RF.

The main steps of applying the stochastic forest method in forecasting are as follows.

1) Step 1: It is assumed that M predictors are obtained through screening, which together with hydrological series constitute training sample set $D = \{(x_i, y_i), x_i \in X, y_i \in Y, i = 1, 2, \ldots, N\}$, in which X is the explanatory variable of M-dimensional vector composed of predictors, Y is the objective variable of predicants (precipitation or runoff), and the sample capacity is N.

2) Step 2: k training sample subsets d_k are randomly selected from the training sample set D by the boostrap resampling technique, and the capacity of the training sample subsets is N.

3) Step 3: k CART decision trees are constructed for k subsets of training samples. According to the random subspace theory, m indicators (usually $m = \sqrt{M}$) are randomly selected from M indicators as node attribute values of the decision tree.

4) Step 4: Each decision tree grows recursively from top to bottom to finally get a predicted value. Vote (mean) the results of k CART decision trees as the final classification (regression) result, which is the final predicted value. The main structure diagram of the RF model is shown in Fig. 3.

B. Extreme Gradient Boosting (XGB)

Extreme gradient boosting (XGB) is a serial boosting algorithm in ensemble learning proposed by C. Tianqi in 2014. It is essentially a CART decision tree ensemble model, which uses the prediction results of K trees as the final result. XGB is a model integrating K CART decision trees. For sample set $D = \{(x^i, y^i)\}$ ($|D| = N, x^i \in R^m, y_i \in R$), XGB linearly combines K learners as

$$\hat{y^i} = \phi(x^i) = \sum_{k=1}^{K} f_k(x^i). \tag{5}$$

The loss function in XGB is defined as

$$L_t = \sum_{k=1}^{K} L\left(y^i, f_{t-1}(x^i) + h_t(x^i)\right) + \Omega(h_t) \tag{6}$$

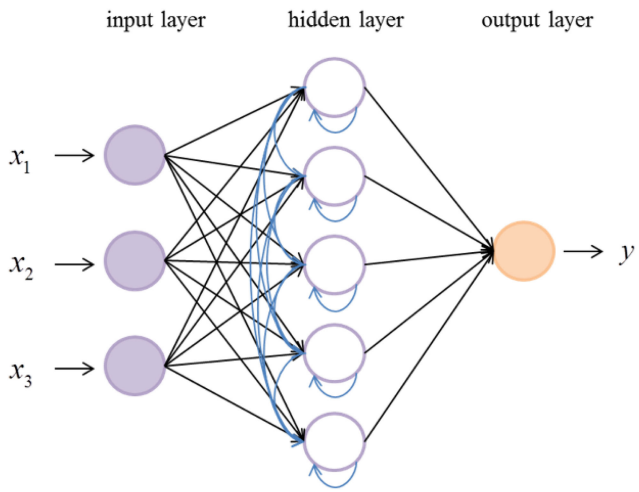

Fig. 4. Structure of the RNN.

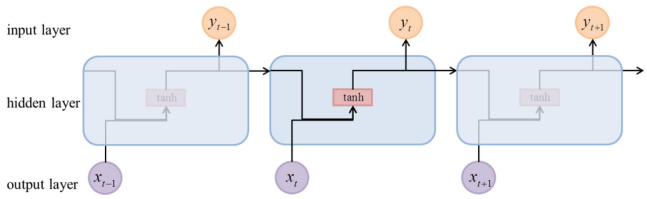

Fig. 5. Expanded structure of the RNN.

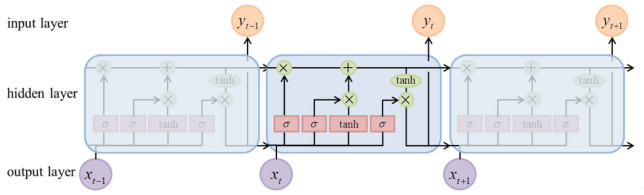

Fig. 6. Structure of the LSTM.

where $f_{t-1}(x^i)$ is a single learner in the previous round, and a tree $h_t(x^i)$ can be found in the iteration of the round to minimize $L(y^i, f_{t-1}(x^i) + h_t(x^i))$. $\Omega(h_t)$ is a regularization term to prevent the model from overfitting

$$\Omega(h_t) = \gamma J + \lambda/2 \sum_{j=1}^{J} \omega_{tj}^2 \tag{7}$$

where J is the number of leaf nodes in the decision tree, and ω_{tj} is the optimal value of the j leaf node in the jth iteration. γ and λ are coefficients, which need to be adjusted in practical application. Our goal is to get the corresponding model when the loss function is minimized.

C. Recurrent Neural Network (RNN)

The recurrent neural network (RNN) originated from the Hopfield network proposed by Hopfield in 1982. It is a special neural network structure that was proposed according to the viewpoint that human cognition is based on past experience and memory. With the development of deep learning and Big Data, researchers have found that the RNN has strong data mining ability; thus, it has gradually become widely used.

The neural network includes an input layer, a hidden layer, and an output layer. The layers are connected by weights, and the value of the output layer is calculated by the activation function. Neurons in each layer are not connected with each other. The biggest difference between the RNN and the feed forward neural network is that the weights are connected between neurons in layers, which is very important in time series prediction. The current output is also related to the previous output. Each neuron in each layer is not independent from the others but has a directional cycle. Therefore, the RNN will memorize previous information and apply it to the calculation of the current output. On the basis of the feed forward neural network diagram, the structure of the RNN is shown in Fig. 4.

D. Long Short-Term Memory (LSTM)

To solve various problems caused by gradient disappearance in the RNN [47], Hochreiter proposed an improved model based on the RNN-long-term and short-term memory model (LSTM) [48] in 1997, which was further improved by Grave [49] in 2008. We expand the standard RNN in the previous section (taking the activation function as tan*h* function as an example) as shown in Fig. 5. The difference between the LSTM and traditional RNN is that there is only one network layer in the unit of the traditional RNN network, while there are four network layers in the LSTM. The structure of the LSTM is shown in Fig. 6.

IV. Study Area and Dataset

A. Study Area

Danjiangkou River Basin is located between $31^\circ \sim 34^\circ$ N latitude and $106^\circ \sim 112^\circ$ E longitude, with a drainage area of about 95 217 km^2, accounting for 60% of the total area of Hanjiang River Basin (see Fig. 7). In the basin, mountains account for 79%, hills account for 18%, and only 3% are plains. The south side is bounded by Micang Mountain and Daba Mountain, and the north side is bounded by Qinling Mountains. In the basin, there are dense rivers, developed water systems, and abundant water resources. The water systems are distributed on both sides of the Han River in pulse shape, and the tributaries are generally short. There are many water conservancy projects in the basin, among which Danjiangkou Reservoir is the water source of the Middle Route Project of South-to-North Water Transfer, which plays a key role in national water resources dispatching. The annual average temperature in the basin is about 15∼17 °C, and the average evaporation in the basin is 900∼1500 mm. The precipitation and water vapor are abundant, but they are unevenly distributed during the year. The rainy season is mostly concentrated in May to October, and the precipitation accounts for more than 80% of the whole year. The average annual precipitation is about 700∼1100 mm.

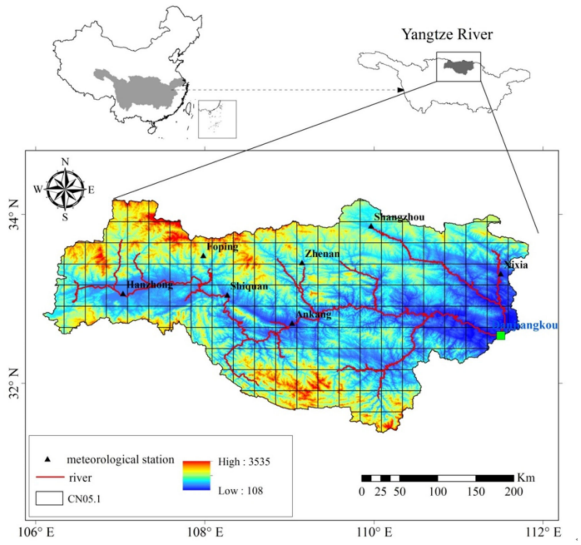

Fig. 7. Map of the Danjiangkou river basin.

B. Dataset

This article adopts two kinds of datasets. The predictands dataset is precipitation dataset, while the predictors used are remote-related climate indexes.

1) Predictands Data: The measured data used in this article are the gridded dataset CN05.1, which was established by W. Jia [50] in 2012 on the basis of the method adopted by dataset CN05 [51], with a spatial resolution of 0.25 × 0.25. CN05.1 based on the daily data of more than 2400 meteorological observation stations (including national reference climate stations, national basic meteorological stations, and national general meteorological stations) distributed all over Chinese, the interpolation calculation is carried out by anomalous approach [52], and the time used in this article is the monthly precipitation data from 1982 to 2015. Fig. 7 shows the distribution of CN05.1 grid in Danjiangkou river basin.

2) Predictors Data: The predictors are based on 130 remote-related climate indexes provided by the National Climate Center in China.[1] Climate indexes contain three parts: 88 atmospheric circulation indexes, 26 sea surface temperature indexes, and 16 other indexes.

V. Results and Discussions

In this article, the K-means clustering algorithm and SMOTE algorithm are combined to augment the sample data to improve the forecasting accuracy. Then, four precipitation forecasting models are constructed, specifically, two shallow ML models (RF, XGB) and two deep learning models (RNN, LSTM), to analyze and observe the stability and universality of the data augmentation algorithm. To verify the practical applicability in the data augmentation algorithm and the difference of forecasting accuracy before and after augmentation, the anomaly correlation coefficient (ACC) and Pg score are used to test and evaluate, respectively, the forecasting accuracy of the precipitation model. The two scores are introduced as follows.

1) *Anomaly correlation coefficient (ACC):* Anomaly correlation coefficient (ACC), also known as spatial similarity coefficient, is an evaluation index determined and recommended by the 11th working conference of the World Meteorological Organization (WMO) held in Italy in 1996, which reflects the spatial similarity between predicted and measured values. The range of ACC is $[-1, 1]$, and the closer it is to 1, the higher the prediction accuracy. The ACC calculation formula is

$$\mathrm{ACC}_n = \frac{\sum_{m=1}^{M}(\Delta o_{m,n} - \overline{\Delta o_n}) \times (\Delta f_{m,n} - \overline{\Delta f_n})}{\sqrt{\sum_{m=1}^{M}(\Delta o_{m,n} - \overline{\Delta o_n})^2 \times \sum_{m=1}^{M}(\Delta f_{m,n} - \overline{\Delta f_n})^2}} \tag{8}$$

where $m = 1, 2, \ldots, M$ is the number of grid points in the region. $n = 1, 2, \ldots, N$ is the sample capacity of time series. $o_{m,n}$ is the observed precipitation. $f_{m,n}$ is the forecasted precipitation. $\overline{o_m}$ is the average of observed precipitation at the grid m, where $\overline{o_m} = (1/N)\sum_{n=1}^{N} o_{m,n}$. $\Delta o_{m,n}$ is the differential value between the observed precipitation and multiyear average precipitation at time n and $\Delta o_{m,n} = o_{m,n} - \overline{o_m}$. $\overline{\Delta o_m}$ is average value of observed precipitation and multiyear average precipitation at time n at all grid points in the region, and $\overline{o_n} = (1/M)\sum_{m=1}^{M} o_{m,n}$. $\overline{f_m}$ is the average of forecasted precipitation at the grid m, and $\overline{f_m} = (1/N)\sum_{n=1}^{N} f_{m,n}$. $\Delta f_{m,n}$ is the differential value between forecasted precipitation and multiyear average precipitation at time n and $\Delta f_{m,n} = f_{m,n} - \overline{f_m}$. Δf_m is the average value of forecasted precipitation and multiyear average precipitation at time n at all grid points in the region, and $\Delta f_{m,n} = f_{m,n} - \overline{f_m}$.

2) *Graded test Pg score:* The graded test Pg score is mainly used to assess the magnitude proximity between forecasted and observed precipitation anomaly percentage, and it is a qualitative grade evaluation standard for operational forecast by China Meteorological Administration since January 1, 2010. Precipitation trend forecast is judged according to the six-grade scoring system. See Table I for detailed classification standards.

See Table II for specific inspection criteria of the Pg score, with the lowest score of 0 and the highest score of 100. When the sign and magnitude of the forecasted and observed anomaly percentage are the same, the score is 100 points. When the difference between the forecasted and observed magnitude is one level, 20 points will be deducted. If there is a difference of two levels, 40 points will be reduced, and so on, until it is reduced to 0 points. When the forecasted and observed anomaly symbols are inconsistent, 20 points will be deducted on the basis of magnitude reduction until it is reduced to 0 points. The Pg score encourages abnormal prediction. When the prediction is abnormal and the difference between prediction and actual

[1][Online]. Available: https://cmdp.ncc-cma.net/Monitoring/cn_index_130.php

TABLE I
GRADING STANDARD OF PG SCORE

1	2	3	4	5	6
$\Delta R \leq -50\%$	$-50\% < \Delta R \leq -20\%$	$-20\% < \Delta R < 0$	$0 \leq \Delta R < 20\%$	$20\% \leq \Delta R < 50\%$	$\Delta R \geq 50\%$

TABLE II
SCORING SYSTEM OF PG SCORE

Forecasted \ Observed	1	2	3	4	5	6
1	100	$80+10$	60	20	0	0
2	$80+10$	100	80	40	20	0
3	60	$80+10$	100	60	40	20
4	20	40	60	100	$80+10$	60
5	0	20	40	80	100	$80+10$
6	0	0	20	60	$80+10$	100

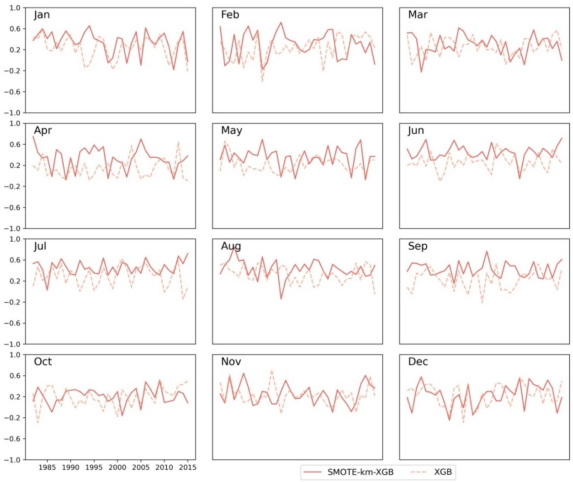

Fig. 8. ACC score of SMOTE-km-XGB and XGB.

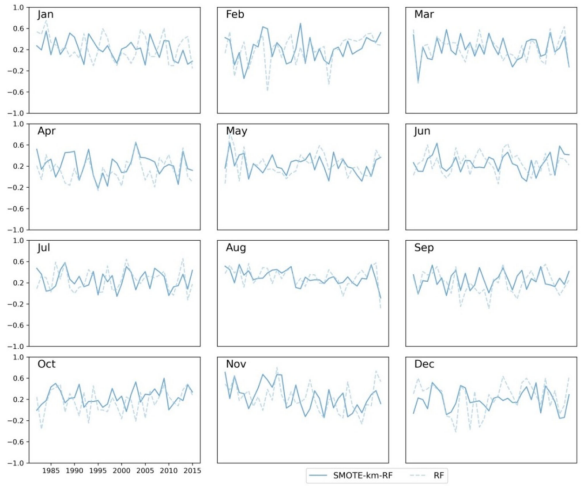

Fig. 9. ACC score of SMOTE-km-RF and RF.

measurement is one order of magnitude, ten points can be added to the aforementioned score.

A. *Results of the Shallow ML Forecasting Model*

To verify the rationality and effectiveness of the SMOTE data augmentation algorithm and K-means clustering algorithm in improving precipitation forecasting accuracy, four ML prediction models are used to verify and compare the results. This section introduces two shallow ML models. The SMOTE data augmentation algorithm and the K-means clustering algorithm use the XGB model to forecast precipitation, which is abbreviated as SMOTE-km-XGB for convenience. The SMOTE data augmentation algorithm and the K-means clustering algorithm are used to forecast precipitation with the RF model, which is abbreviated as SMOTE-km-RF for convenience. Figs. 8 and 9 show the ACC score comparison between the aforementioned two methods and the original method without SMOTE-km.

The ACC score in Fig. 8 shows that the prediction performance of the SMOTE-km-XGB method is superior to that of the XGB method in January, February, April, June, July, and September. In most years, the ACC value exceeds that of the XGB method, even more than double in some months, and there is a clear trend of improving accuracy in other months. The fluctuation range of the ACC value is smaller than that of the SMOTE-km-XGB, and the model shows higher stability and robustness, which shows that the coupling of the SMOTE data augmentation algorithm and K-means clustering is beneficial for expanding the sample datasets and improving the forecasting accuracy when the XGB model is used for precipitation forecasting. At the same time, the ACC scores of both the SMOTE-km-XGB and XGB methods are higher in June and July than those of other months, while the ACC scores in December are generally lower. A possible reason is that the precipitation level is larger in June and July during the flood season, and the ACC value does not fluctuate as much as when the precipitation level is small through calculation. The ACC score is more sensitive when the precipitation level is small. Generally, the SMOTE-km-XGB method has achieved reasonable application results in the study of precipitation forecasts in river basins. The ACC score in Fig. 9 shows that the gap of the SMOTE-km coupling algorithm in the RF precipitation prediction model is not significant, and SMOTE-km-RF only shows a weak advantage. Searching for the reason for the depth may have a great relationship with the principle of the RF model itself. As seen from the introduction

of RF, when using sample datasets to construct decision trees, RF uses bootstrap resampling technology to randomly sample the sample datasets to obtain sample data subsets, which is also a way of virtually increasing the number of samples. Therefore, the SMOTE-km algorithm did not greatly improve the forecasting accuracy. However, judging from the fluctuation range of the ACC value, the SMOTE-km-RF method has a smaller fluctuation range and is more stable, which means that if the forecasting accuracy of the RF model is not good in some special years, using the SMOTE-km method to expand the sample data has advantages and changes that cannot be underestimated. From the perspective of monthly changes, the ACC scores in July and August are more stable; their values basically fluctuate within the range of [0.1,0.6], which may be due to the stable changes of the forecast model in ACC scores when the precipitation level is large. Different from the quantitative calculation of the ACC index, the Pg score mainly focuses on the forecasting accuracy of the model prediction value in order of magnitude, which is a qualitative index. The closer its value is to 100, the better the forecasting accuracy is. The deeper blue is in the figure, the better the forecasting accuracy is; the deeper red is, the worse the forecasting accuracy is. To simplify the layout, the values of the Pg score of the SMOTE-km-XGB, XGB, SMOTE-km-RF, and RF precipitation forecasting models are drawn in the same picture, but they are still separately compared and analyzed, as shown in Fig. 10. The left half of the Pg score shows that the forecast accuracy of the SMOTE-km-XGB method in January, February, May, June, and July is much higher than that of XGB. In January, among 167 grid points in the whole basin, approximately 80% of the grid points scored above 80, only six grid points scored below 70, and no grid points scored below 60. However, because the XGB score has two grid points whose scores are lower than 50, the prediction results are almost difficult to adopt, only three grid points exceed 80, and most other grid points are between [60,70]. In June, more than half of the grid points had a Pg score above 90, and the Pg score of the whole basin was not less than 70. When only the XGB method is used for forecasting, the Pg score is between 70 and 80 at most grid points, and few grid points exceed 80. Among statistical data of 167 grids, only one grid has a Pg score above 90. It is fully explained that the precision of the SMOTE-km data expansion algorithm is improved when it is applied to the XGB model to forecast precipitation. It can be seen from the right half of the Pg score chart that the SMOTE-km-RF method is slightly better than the RF method in February, April, July, and October. In February, more than 70% of the grid score in the SMOTE-km-RF model scored above 80, a few grid points scored above 80 in the RF model, but most grid points scored lower than those in the SMOTE-km-RF model. There is a large difference between the two methods in April. Only one lattice of SMOTE-km-RF has a Pg score lower than 70, and approximately half of the lattice scores are higher than 80, while the scores of nearly half of the grids of the RF method are lower than 70, and the scores of three grids are lower than 60. However, the Pg score shows that the SMOTE-km-RF method is slightly inferior to the RF method in May and December, which is consistent with the ACC score.

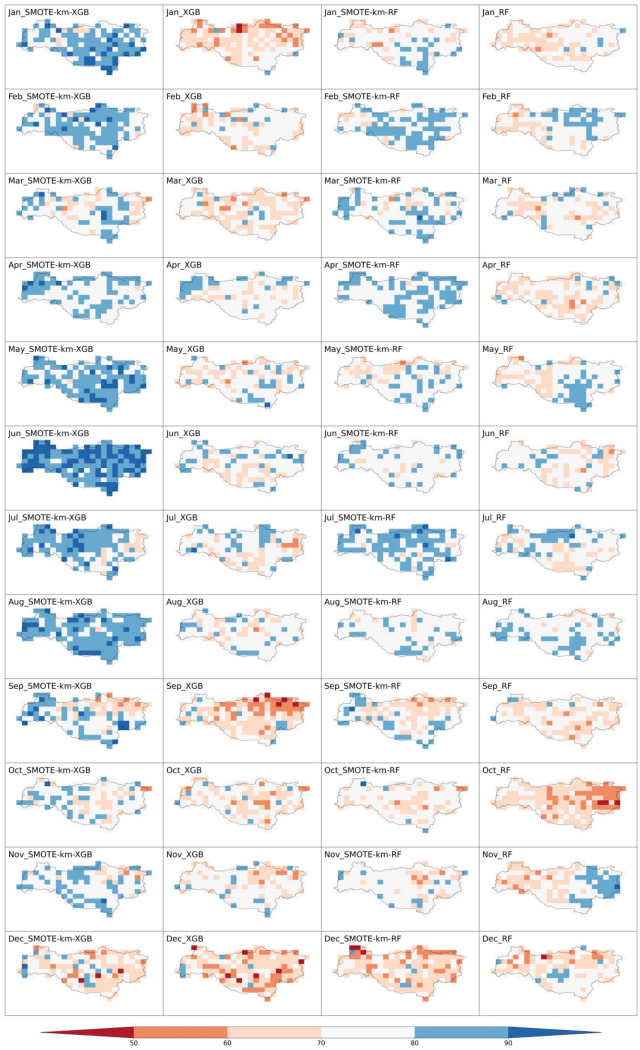

Fig. 10. Pg score of SMOTE-km-XGB, XGB, SMOTE-km-RF, and RF.

B. Results of Deep Learning Forecasting Model

This section introduces the prediction results of two deep learning models, which are abbreviated as SMOTE-km-RNN and SMOTE-km-LSTM. Figs. 11 and 12 show the ACC score comparison between the aforementioned two methods and the original method without SMOTE-km.

The ACC score in Fig. 11 shows that the SMOTE-km method has no obvious contribution to the improvement of accuracy in the RNN precipitation forecasting model. Even in May and August, the ACC score of the SMOTE-km-RNN precipitation forecasting model is lower than that of the RNN. The reason may be that although the sample capacity is increased when expanding the sample data by the SMOTE-km data augmentation algorithm, the noise data of the sample data are also invisibly increased. When the deep learning model connects the multilayer neurons to the sample data, it constantly calculates through the activation function. In other words, the existence of noise data additionally increases the calculation error of the model. For deep learning, the monthly series data in hydrology still cannot meet the requirements in terms of data quantity. Therefore, it

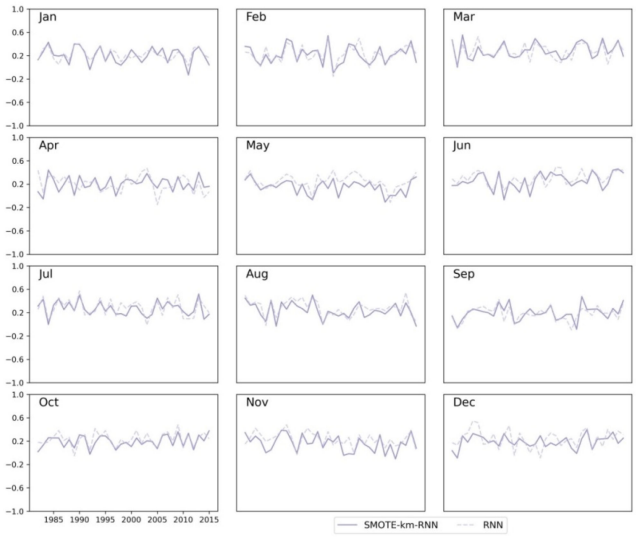

Fig. 11. ACC score of SMOTE-km-RNN and RNN.

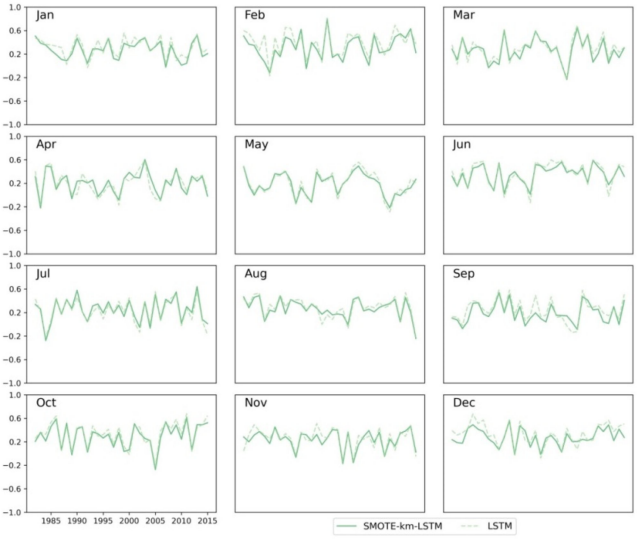

Fig. 12. ACC score of SMOTE-km-LSTM and LSTM.

may not be helpful for the deep learning model to increase a small part of the sample data by the data amplification method as it also virtually increases the loss of the model, which wastes the sample data preprocessing time and increases the model error. Coincidentally, in Fig. 12, when the SMOTE-km method is applied to the LSTM precipitation forecasting model, the ACC value is almost the same as that forecasted by directly using LSTM. Different from the RNN model, the SMOTE-km-LSTM precipitation forecasting model does not show worse results than the LSTM precipitation forecasting model in individual months. This may be related to the difference in structure between the LSTM and RNN models. Compared with the RNN model, the LSTM model has a layer of memory units, which makes the LSTM model have a memory function. Therefore, the forecasting model has strong stability and is weak under the influence of redundant data and noise data. The left half of Fig. 13 shows the Pg score of the RNN precipitation forecasting model,

Fig. 13. Pg score of SMOTE-km-RNN, RNN, SMOTE-km-LSTM, and LSTM.

which uses the SMOTE-km algorithm to expand sample data and does not use the data expansion algorithm, which is similar to the ACC score. In May and August, the forecasting accuracy of the SMOTE-km-RNN model is lower than that of the RNN model, and there is little difference in the ACC value between the two models in other months. In May, more than half of the grids in the RNN model scored more than 80, while only a few grid points scored more than 80 in the SMOTE-km-RNN model, and one grid even scored less than 50. Therefore, its accuracy was poor. In the right half of Fig. 13, the Pg scores of the LSTM precipitation forecasting model are expanded with the SMOTE-km algorithm and without the data expansion algorithm. In February, the Pg scores of the SMOTE-km-LSTM are quite different from those of LSTM, and the Pg scores of almost all grid points are lower than LSTM. It can be roughly observed from Fig. 13 that the forecasting accuracy of the LSTM precipitation forecasting model in the Danjiangkou Basin is slightly better than that of the RNN precipitation forecasting model, which is worthy of attention and development in future research. Generally, the SMOTE-km data augmentation algorithm has not achieved considerable results in deep learning. From the ACC score and Pg score, the prediction accuracy does not show obvious

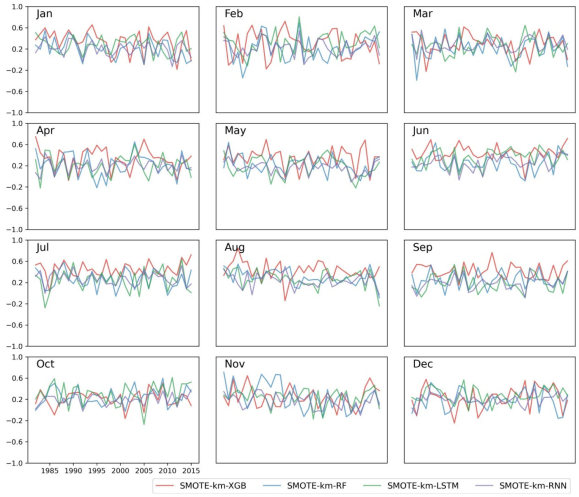

Fig. 14. ACC score of SMOTE-km-XGB, SMOTE-km-RF, SMOTE-km-RNN, and SMOTE-km-LSTM.

differences between months, which is somewhat different from the results of shallow ML precipitation forecasting.

C. Results of Shallow and Deep Learning Forecasting Model

To compare and analyze the accuracy of four precipitation forecasting models: SMOTE-km-XGB, SMOTE-km-RF, SMOTE-km-RNN, and SMOTE-km-LSTM, the ACC scores and Pg scores of the four models were compared (see Figs. 14 and 15). It is easy to see that SMOTE-km-XGB outperforms the other three models in more months, and the forecasting accuracy of the two shallow ML models is better than that of the two deep learning models, especially in January, June, July, and August. In February, when the Pg scores of most grids of the two shallow ML models exceeded 80, the Pg scores of the two deep learning models were generally low, especially for the RNN model, where the Pg scores of most grid points were lower than 70 and three grid points were lower than 50. In June, the ACC score and Pg score of SMOTE-km-XGB were significantly higher than those of the other three models. In July, the Pg scores of the two shallow ML models exceeded 80 in most grids, while the Pg scores of the two deep learning models were lower than 70 in most grids. At the same time, the forecasting accuracy of the shallow ML model is similar in 167 grids in the whole basin, while the two deep learning models are different. The differential value of the Pg score in the same month can even reach more than 50, which is not very accurate for the forecasting of grid precipitation data. The aforementioned results show that when the amount of sample data is very small, the deep learning model does not show the advantages of deep mining, and the forecasting accuracy cannot reach the accuracy of the shallow ML model. The aforementioned results show that the SMOTE-km-XGB method is more suitable than the other three methods for precipitation prediction in the basin studied in this article.

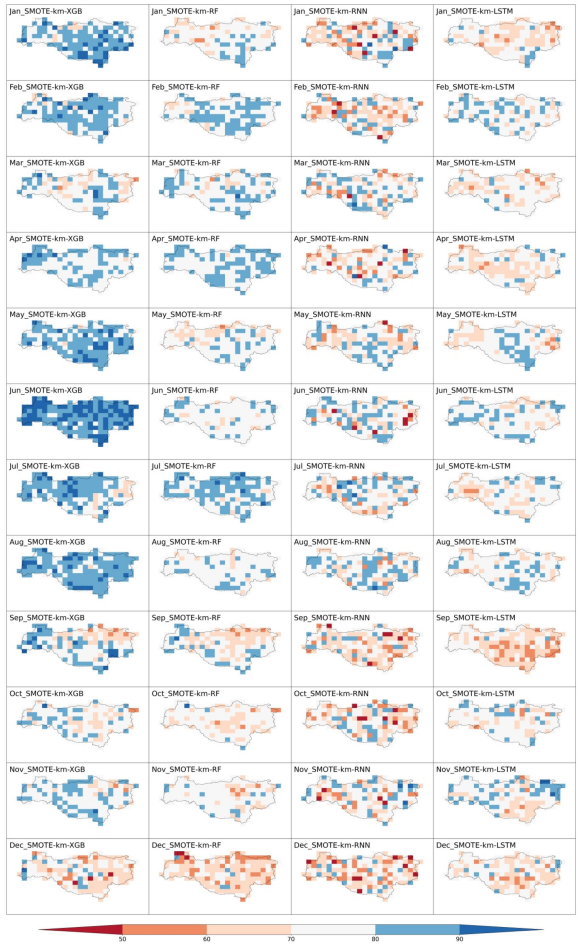

Fig. 15. Pg score of SMOTE-km-XGB, SMOTE-km-RF, SMOTE-km-RNN, and SMOTE-km-LSTM.

VI. Conclusion

In view of the characteristics of the ML model, the amount of data used in the model construction directly affects the accuracy of ML modeling. Therefore, this article constructs a data augmentation algorithm based on the K-means clustering algorithm and SMOTE to expand the precipitation series to expand the data within a reasonable range to meet the basic needs of the ML model and improve the performance of medium- and long-term precipitation forecasting. First, the sample datasets are clustered by the K-means clustering algorithm, and the sample data are grouped into three categories. Then, the SMOTE data augmentation algorithm is used to expand the categories with a small number of samples in the class according to the oversampling rate of the original datasets, and the final augmented samples make the sample data more balanced when the sample data amount increases. Taking the augmented sample data as input, four ML models (XGB, RF, RNN, and LSTM) are used to compare the forecasting results before and after the augmented precipitation data. The differences between shallow ML and deep learning are compared, and the differences, advantages, and disadvantages between shallow ML and deep learning models are discussed in depth. The results of the case study of the Danjiangkou River basin are summarized as follows.

1) The forecasting results of two shallow ML methods (RF and XGB) show that the forecasting accuracy is improved after the sample data are processed by the K-means clustering algorithm and SMOTE data augmentation algorithm. In most years, the ACC and Pg scores of SMOTE-km-XGB and SMOTE-km-RF exceed those of XGB and RF. Furthermore, compared with the other three methods, SMOTE-km-XGB method is more suitable for precipitation forecasting in the basin studied in this article.
2) The forecasting results of the two deep learning methods (RNN and LSTM) show that the sample data processed by the K-means clustering algorithm and SMOTE data augmentation algorithm have not achieved considerable results in deep learning. The possible reason is that when the sample data are very small, the requirements of the deep network cannot be met.
3) This study improves the accuracy of precipitation forecasting by expanding and balancing the information of sample data, and provides a new research idea for improving the accuracy of medium- and long-term hydrological forecasting.

REFERENCES

[1] A. J. Simmons, K. M. Willett, P. D. Jones, P. W. Thorne, and D. P. Dee, "Low frequency variations in surface atmospheric humidity, temperature, and precipitation: Inferences from reanalyses and monthly gridded observational data sets," *J. Geophys. Res., Atmos.*, vol. 115, no. D1, pp. 1–21, Jan. 2010.

[2] G. Thompson, P. R. Field, R. M. Rasmussen, and W. D. Hall, "Explicit forecasts of winter precipitation using an improved bulk microphysics scheme. Part II: Implementation of a new snow parameterization," *Monthly Weather Rev.*, vol. 136, no. 12, pp. 5095–5115, Dec. 2008.

[3] W. Volker *et al.*, "The convective and orographically-induced precipitation study (COPS): The scientific strategy, the field phase, and research highlights," *Quart. J. Roy. Meteorological Soc.*, vol. 137, pp. 3–30, Jan. 2011.

[4] J. Guo *et al.*, "Investigation of near-global daytime boundary layer height using high-resolution radiosondes: First results and comparison with ERA-5, MERRA-2, JRA-55, and NCEP-2," *Atmospheric Chem. Phys.*, vol. 21, no. 22, pp. 17079–17097, Nov. 2021.

[5] S. Saha *et al.*, "The NCEP climate forecast system version 2," *J. Climate*, vol. 27, no. 6, pp. 2185–2208, Mar. 2014.

[6] A. Cottrill *et al.*, "Seasonal forecasting in the Pacific using the coupled model POAMA-2," *Weather Forecasting*, vol. 28, no. 3, pp. 668–680, 2013.

[7] D. Botes, J. R. Mecikalski, and G. J. Jedlovec, "Atmospheric infrared sounder (AIRS) sounding evaluation and analysis of the pre-convective environment," *J. Geophysical Res. Atmos.*, vol. 117, 2012, Art. no. D09205.

[8] J. Guo *et al.*, "Viurnal variation and the influential factors of precipitation from surface and satellite measurements in Tibet," *Int. J. Climatol.*, vol. 34, pp. 2940–2956, 2014.

[9] M. Min *et al.*, "Estimating summertime precipitation from Himawari-8 and global forecast system based on machine learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 5, pp. 2557–2570, May 2019.

[10] X. Zou, Z. Qin, and F. Weng, "Improved coastal precipitation forecasts with direct assimilation of GOES-11/12 imager radiances," *Monthly Weather Rev.*, vol. 139, no. 1, pp. 3711–3728, Dec. 2011.

[11] M. Wakin, *Dimensionality Reduction*. New York, NY, USA: Wiley, 2007.

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

[13] J. E. Shortridge, S. D. Guikema, and B. F. Zaitchik, "Machine learning methods for empirical streamflow simulation: A comparison of model accuracy, interpretability, and uncertainty in seasonal watersheds," *Hydrol. Earth Syst. Sci.*, vol. 7, no. 20, pp. 2611–2628, 2016.

[14] K. Fang and C. Shen, "Full-flow-regime storage-streamflow correlation patterns provide insights into hydrologic functioning over the continental US," *Water Resour. Res.*, vol. 53, 8064–8083, 2017.

[15] P. S. Yu *et al.*, "Comparison of random forests and support vector machine for real-time radar-derived rainfall forecasting," *J. Hydrol.*, vol. 552, pp. 92–104, 2017.

[16] Z. Liang *et al.*, "Long-term streamflow forecasting using SWAT through the integration of the random forests precipitation generator: Case study of Danjiangkou reservoir," *Hydrol. Res.*, vol. 49, no. 5, pp. 1513–1527, 2018.

[17] Z. M. Yaseen *et al.*, "Stream-flow forecasting using extreme learning machines: A case study in a semi-arid region in Iraq," *J. Hydrol.*, vol. 542, pp. 603–614, 2016.

[18] L. Breiman, "Random forest," *Mach. Learn.*, vol. 45, pp. 5–32, 2001.

[19] L. Schoppa, M. Disse, and S. Bachmair, "Evaluating the performance of random forest for large-scale flood discharge simulation," *J. Hydrol.*, vol. 590, Sep. 2020, Art. no. 125531.

[20] J. M. Sadlera, J. L. Goodalla, M. M. Morsyab, and K. Spencerc, "Modeling urban coastal flood severity from crowd-sourced flood reports using Poisson regression and random forest," *J. Hydrol.*, vol. 559, pp. 43–55, Apr. 2018.

[21] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.

[22] R. Noori *et al.*, "Assessment of input variables determination on the SVM model performance using PCA, Gamma test, and forward selection techniques for monthly stream flow prediction," *J. Hydrol.*, vol. 401, no. 3-4, pp. 177–189, 2011.

[23] Z. He *et al.*, "A comparative study of artificial neural network, adaptive neuro fuzzy inference system and support vector machine for forecasting river flow in the semiarid mountain region," *J. Hydrol.*, vol. 509, pp. 379–386, 2014.

[24] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, Aug. 2016, pp. 785–794.

[25] S. Abba *et al.*, "Evolutionary computational intelligence algorithm coupled with self-tuning predictive model for water quality index determination," *J. Hydrol.*, vol. 587, Aug. 2020, Art. no. 124974.

[26] Y. Li *et al.*, "A multi-model integration method for monthly streamflow prediction: Modified stacking ensemble strategy," *J. Hydroinform.*, vol. 22, no. 2, pp. 310–326, Mar. 2020.

[27] G. Ke *et al.*, "LightGBM: A highly efficient gradient boosting decision tree," *Adv. Neural Inf. Process. Syst.*, vol. 30, pp. 3146–3154, 2017.

[28] R. Sutinen and M. Middleton, "Soil water drives distribution of northern Boreal conifers Picea abies and Pinus sylvestris," *J. Hydrol.*, vol. 588, 2020, Art. no. 125048.

[29] J. Cai *et al.*, "An assembly-level neutronic calculation method based on LightGBM algorithm," *Ann. Nucl. Energy*, vol. 150, 2020, Art. no. 107871.

[30] J. J. Hopfield, "Neural networks and physical systems with emergent collective computational abilities," *Proc. Nat. Acad. Sci.*, vol. 79, no. 8, pp. 2554–2558, 1982.

[31] Z.C. Lipton, J. Berkowitz, and C. Elkan, "A critical review of recurrent neural networks for sequence learning," 2015, *arXiv:1506.00019*.

[32] J. Zhang *et al.*, "Developing a long short-term memory (LSTM) based model for predicting water table depth in agricultural areas," *J. Hydrol.*, vol. 561, pp. 918–929, 2018.

[33] Z. Xiang, J. Yan, and I. Demir, "A rainfall-runoff model with LSTM-based sequence-to-sequence learning," *Water Resour. Res.*, vol. 56, no. 1, 2020, Art. no. e2019WR025326.

[34] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, 2015, pp 234–241.

[35] Z. Mushtaq and S. Su, "Environmental sound classification using a regularized deep convolutional neural network with data augmentation," *Appl. Acoust.*, vol. 167, Oct. 2020, Art. no. 107389.

[36] M. Frid-Adar *et al.*, "Synthetic data augmentation using GAN for improved liver lesion classification," in *Proc. IEEE 15th Int. Symp. Biomed. Imag.*, Apr. 2018, pp. 289–293.

[37] X. Peng, Z. Tang, F. Yang, R. S. Feris, and D. Metaxas, "Jointly optimize data augmentation and network training: Adversarial data augmentation in human pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 2226–2234. [Online]. Available: https://arxiv.org/abs/1805.09707

[38] F. Chen *et al.*, "Self-supervised data augmentation for person re-identification," *Neurocomputing*, vol. 415, pp. 48–59, Nov. 2020.

[39] B. Xiang, A. Kopa, Z. Fu, and A. B. Apsel, "Theoretical analysis and practical considerations for the integrated time-stretching system using dispersive delay line (DDL)," *IEEE Trans. Microw. Theory Techn.*, vol. 60, no. 11, pp. 3449–3457, Nov. 2012.
[40] A. A. El-Sayed *et al.*, "Handling autism imbalanced data using synthetic minority over-sampling technique (SMOTE)," in *Proc. 3rd World Conf. Complex Syst.*, Nov. 23–25, 2015, pp. 1–5.
[41] Y. Abdulaziz and S. Ahmad, "Infant cry recognition system: A comparison of system performance based on mel frequency and linear prediction cepstral coefficients," in *Proc. Int. Conf. Inf. Retrieval Knowl. Manage.*, Shah Alam, Malaysia, Mar. 2010, pp. 260–263.
[42] F. V. D. Heijden *et al.*, *Unsupervised Learning*. New York, NY, USA: Wiley, 2005.
[43] S. Del Rio, V. Lopez, J. M. Benitez, and F. Herrera, "On the use of mapReduce for imbalanced Big Data using random forest," *Inf. Sci.*, vol. 285, pp. 112–137, Nov. 2014.
[44] C. T. Su and Y. H. Hsiao, "An evaluation of the robustness of MTS for imbalanced data," *IEEE Trans. Knowl. Data Eng.*, vol. 19, no. 10, pp. 1321–1332, Oct. 2007.
[45] L. Breiman, "Bagging predictors," *Mach. Learn.*, vol. 24, pp. 123–140, Aug. 1996.
[46] T. Ho, "The random subspace method for constructing decision forests," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 8, pp. 832–844, Aug. 1998.
[47] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE Trans. Neural Netw.*, vol. 5, no. 2, pp. 157–166, Mar. 1994.
[48] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
[49] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional LSTM and other neural network architectures," *Neural Netw.*, vol. 18, vo. 5-6, pp. 602–610, 2005.
[50] J. Wu, X.-J. Gao, Y.-L. Xu, and J. Pan, "Numerical simulation and uncertainty analysis of regional climate change in east Asia and southeast Asia," *Chin. Acad. Meteorological Sci.*, vol. 8, no. 3, pp. 147–152, Aug. 2015.
[51] X. Ying *et al.*, "A daily temperature dataset over China and its application in validating a RCM simulation," *Adv. Atmospheric Sci.*, vol. 4, no. 26, pp. 153–162, 2009.
[52] M. New, M. Hulme, and P. Jones, "Representing twentieth-century space-time climate variability. Part II: Development of 1901-96 monthly grids of terrestrial surface climate," *J. Climate*, vol. 13, no. 12, pp. 829–856, 2000.

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

Since 2021, she has been a Postdoctoral Fellow with the School of Geographic and Biologic Information, Nanjing University of Posts and Telecommunications, Nanjing. Her recent research interests include medium- and long-term hydrological forecasting, machine learning, deep learning, and flood forecasting.

Donglai Jiao received the Dr. Eng. degree in geographic information science from Nanjing Normal University, Nanjing, China, in 2009.

He is currently a Professor with the School of Geographic and Biologic Information, Nanjing University of Posts and Telecommunications, Nanjing. His recent research interests include Internet of Things geographic information system, spatio-temporal Big Data processing and analysis, etc.

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

Since 2020, he has been an Engineer with the Hydrology and Water Resources Department, Nanjing Hydraulic Research Institute, Nanjing. His recent research interests include multisource precipitation fusion and hydrology simulation.

Guan Gui (Senior Member, IEEE) received the Dr. Eng. degree in information and communication engineering from the University of Electronic Science and Technology of China, Chengdu, China, in 2012.

Since 2015, he has been a Professor with the College of Telecommunications and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing, China. His recent research interests include artificial intelligence, deep learning, non-orthogonal multiple access, wireless power transfer, and physical layer security.