

RESEARCH ARTICLE

Effect of Data Characteristics Inconsistency on Medium and Long-Term Runoff Forecasting by Machine Learning

PING AI^{1,3}, CHUANSHENG XIONG¹, KE LI², YANHONG SONG³, SHICHENG GONG¹, AND ZHAOXIN YUE⁴

¹College of Hydrology and Water Resources, Hohai University, Nanjing 210098, China

²Chengdu Engineering Corporation Ltd., Chengdu 610072, China

³College of Computer and Information, Hohai University, Nanjing 211100, China

⁴College of Computer and Software, Nanjing Vocational University of Industry Technology, Nanjing 210023, China

Corresponding author: Chuansheng Xiong (xcs123@163.com)

This work was supported in part by the Key Research and Development Project of Jiangsu Province under Grant BE2020729, and in part by the School Research Fund of Nanjing Vocational University of Industry Technology under Grant YK21-05-05.

ABSTRACT In the application of medium and long-term runoff forecasting, machine learning has some problems, such as high learning cost, limited computing cost, and difficulty in satisfying statistical data assumptions in some regions, leading to difficulty in popularization in the hydrology industry. In the case of a few data, it is one of the ways to solve the problem to analyze the data characteristics consistency. This paper analyzes the statistical hypothesis of machine learning and runoff data characteristics such as periodicity and mutation. Aiming at the effect of data characteristics inconsistency on three representative machine learning models (multiple linear regression, random forest, back propagation neural network), a simple correction/improvement method suitable for engineering was proposed. The model results were verified in the Danjiangkou area, China. The results show that the errors of the three models have the same distribution as the periodic characteristics of the runoff periods, and the correction/improvement based on periodicity and mutation characteristics can improve the forecasting accuracy of the three models. The back propagation neural network model is most sensitive to the data characteristics consistency.


INDEX TERMS Danjiangkou reservoir, data characteristics consistency, machine learning, medium and long-term runoff forecasting, mutation, characteristics, periodicity characteristics.

I. INTRODUCTION

Medium and long-term runoff forecasting is a quantitative or qualitative analysis of the state of runoff in a certain period in the future based on the available information (tested or analyzed information). It is one of the essential bases for national economic and social development scientific planning. It has a high demand in flood and drought prevention, reservoir scheduling, hydropower plant operation and shipping management. The medium and long-term runoff forecasting is the primary reference and vital link to full use of water resources,

realizing the optimal operation of reservoirs and bringing into play the economic benefits of power stations.

There are limitations of natural phenomena. In medium and long-term runoff processes, the prediction period of river flood forecasting cannot exceed the propagation time of river flood waves, and the prediction period of rainfall-runoff forecasting cannot exceed the basin confluence time. However, most short-term runoff forecasting methods mainly deduce through solid causality. These causal relationships do not work well over more extended periods. Therefore, the medium and long-term process of runoff is not a simple combination of multiple short-term processes. Consequently, it is challenging to meet the application requirements by applying traditional physically driven methods (e.g., short-term

The associate editor coordinating the review of this manuscript and approving it for publication was Frederico Guimarães .

runoff forecasting methods) to solve the medium and long-term runoff forecasting problems because of the low accuracy of the results. At this time, data-driven strategies (such as machine learning) weakened the physical relationship between factors represented and came into the limelight.

Machine learning algorithms belong to the algorithmic model in the data-driven model. Data-driven models are mainly based on statistics, which start from initial data or observations and apply heuristic rules to find and establish relationships between interior characteristics to discover some theorems or laws. The algorithms consider that the data generation process is so complex and unknown that it is challenging to establish mathematical or physical relationships. Therefore, the constructed algorithms are always tricky to explain the reasons.

Machine learning methods have received attention from scholars researching medium and long-term runoff forecasting since their introduction. Among them, neural networks [1], Support Vector Machine (SVM) [2], Random Forest (RF) [3], and aggregation algorithms [4] have been used in flood analysis and simulation for a long time. These methods are effective, but it has been found that there are many limitations to the method. At this time, the improvement of the algorithm and the integration with other methods become the focus. Such as combining traditional hydrological models like the Xin'an Jiang model [5] or improving based on an intelligent algorithm like Particle Swarm Optimization (PSO) [6].

With the maturity of deep learning techniques, machine learning methods have higher performance. Deep learning models represented by Convolutional Network and Temporal Network are analyzed in space and time, respectively, improving performance significantly. Wavelet Coupled Neural Network (WNN) [7], Long Short-Term Memory (LSTM) [8], Deep Belief Network (DBN) [9], and Extreme Learning Machine Algorithm (ELM) [10] and other models have good accuracy in the study of rainfall-runoff relationship and climate factor-rainfall relationship.

Although the accuracy of the machine learning model is getting higher and higher in medium and long-term runoff forecasting, there are still some things that could be improved at the application level. First, although deep learning has high precision, it requires workers with a particular computer foundation to master it. Current machine learning models cannot co-model and learn between rivers with wide gaps, which requires high learning costs. Secondly, hydrologists need to be responsible for the result of runoff forecasting. However, the results of machine learning models cannot be fully trusted and only serve as a reference because of no interpretability, so the computational cost provided by machine learning is relatively limited. Finally, some non-key rivers have less data, which is challenging to meet the statistical hypothesis of the data of high-precision machine learning. Therefore, regarding runoff forecasting, the amount of computation and accuracy of machine learning models should match the amount of data and demand. However,

the current trend in research is to apply better and more complex models to solve problems. Researchers are primarily concerned with the structure and performance of models so that everyone wants to do the model work rather than the data [11]. In cases where the data type is insufficient to provide sufficient information, they lack a systematic understanding of the consistency of invisible, trying, and taken-for-granted data characteristics in machine learning, leading to ambiguous data characteristics and, thus, a learning process with deviation. At this point, data characteristics consistency is essential.

Therefore, this paper analyses whether the runoff sequence meets the data characteristics consistency regarding periodicity and mutation. Then analyze the influence of different runoff characteristics on the simulation results of three machine learning models (multiple linear regression, random forest, and back propagation neural network). Finally, the results of the machine learning model are improved and compared by periodic characteristics and mutation characteristics to show the impact of inconsistent data characteristics on the medium and long-term runoff forecasting based on machine learning. This study aims at alerting researchers to inconsistencies in data characteristics on medium and long-term runoff forecasting and provides a simple modification method to provide a data-driven reference for runoff forecasting in the absence of data.

II. METHODOLOGY

A. DATA CHARACTERISTICS CONSISTENCY

1) STATISTICAL HYPOTHESIS OF MACHINE LEARNING

To some extent, machine learning has solved the difficulty of getting analytical solutions when hydrological process simulation occurs. The traditional paradigm of machine learning research can be defined as follows [12]:

$$f^* = \operatorname{argmin}_{f_{\theta} \in \mathfrak{F}} \{ \mathbb{E}_{\mathcal{D}} [l(f_{\theta}(x), y)] + \lambda p(f_{\theta}) \} \quad (1)$$

where $\mathcal{D} = \{x_i, y_i\}$ is a dataset; $f^* : x \rightarrow y$ is a possibly existed mapping; argmin is an optimization algorithm; \mathfrak{F} is a hypothetical space; l is a loss function; p is a regularizer; f_{θ} is a function in \mathfrak{F} with parameter θ .

An excellent machine learning model has the best loss function, optimization algorithm, complete data, regular term, reasonable assumption space. So, the model's applicability should consider these five points, not just the structure and parameters. This paper focuses on the machine learning model in the big data environment, which attempts to simulate all cases of functions by fitting them to the data. A prerequisite to achieving this goal is to satisfy the statistical hypothesis of the data on machine learning, that is, the independence hypothesis on loss function, the large capacity hypothesis on hypothetical spaces, and the completeness hypothesis on training data. Under these hypotheses, machine learning models will assume that the sample data are large enough and good enough to cover a variety of extreme cases and that there is enough information for all characteristic changes to

be simulated. So, the more complex machine learning models will always be based on big data.

2) RUNOFF DATA CHARACTERISTICS

These hypotheses are not entirely valid in the current research on medium and long-term runoff forecasting. We cannot observe the complete life cycle of rainfall runoff, making the class distribution imbalanced and partial data available. Compared with the whole life cycle of runoff, the data sequence is short. Small samples make the frequency and weight of various characteristics not necessarily conform to the general law, leading to data characteristics often being unbalanced.

On the other hand, runoff data are easily disturbed. Due to climate change or human intervention, the runoff process's environmental characteristics are prone to change, and the data characteristics are very likely to vary significantly. The naturally recorded/collected data are guaranteed to be neither from the same distribution nor independent of each other in different times and spaces, resulting in data heterogeneity. Due to the particularity of medium and long-term runoff process data, it is challenging to label many labels like image recognition so that the data are accurate.

Therefore, small sample data will be most input in the current learning of medium and long-term runoff. It is common for data series to be distributed in a different distribution on medium and long-term runoff forecasting. Different distributions can easily lead to the instability of eigenvalues in machine learning. The model itself finds this situation challenging because of insufficient data. This effect caused by the inconsistency of data characteristics is difficult to solve by the improved algorithm, so it needs to be analyzed from the characteristics of the data itself.

3) PERIODICITY AND MUTATION

Periodicity and mutation are the common cases of inconsistent data characteristics in runoff series. These two situations are sometimes tricky to mined by the machine learning model. Periodicity is the expression of long-term cyclic fluctuations in the response area and is one of the main characteristics of runoff evolution. The mutation is the appearance of discontinuous jumps in the form of changes in the sequence, usually caused by hydrological series subjected to changes in the substratum and climate change [13]. Periodicity analysis often utilizes wavelet analysis, and mutation analysis uses the M-K test.

Wavelet analysis [14] is a non-smooth time sequence analysis method that can analyze the multi-timescale resolution characteristics and the variation characteristics of hydrological elements by identifying the period. The main contents of wavelet analysis are wavelet function and wavelet transform. The wavelet coefficients are obtained by wavelet transformation, and they can reflect the periodic transformation characteristics of the wavelet function. The medium and long-term runoff process belongs to the non-stationary time sequence and multi-time scale structure, and it has the elements of

randomness and regionality. The characteristics match well with the aspects of a wavelet function, so it is often used to obtain the periodic characteristics of hydrologic sequence.

Mann-Kendall (M-K) test [15] is a nonparametric test method recommended by the World Meteorological Organization and is widely used in mutation research. It is one of the trend analysis methods of runoff time sequence [16]. Once proposed, this method has been respected by many scholars and is widely used to analyze the runoff trend, temperature, precipitation, and water quality. A nonparametric M-K mutation test does not require the analysis of samples to obey a specific distribution. It is also not disturbed by other outliers, so it is often used in hydrology.

B. MODEL UPDATING METHOD

1) MACHINE LEARNING ALGORITHM

Multiple linear regression (MLR) algorithm [17] can be used to study the quantitative linear statistical relationship between multiple predictors and predictors. After mastering a certain amount of hydrometeorological data, the regression model can obtain the closeness of the relationship between objective variables and the structural state. Assuming that the predictor is Y and the set of predictors is X , the linear regression modulus is:

$$Y = X\beta + \varepsilon \quad (2)$$

where:

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}; \quad X = \begin{bmatrix} 1x_{11}x_{12}\cdots x_{1m} \\ 1x_{21}x_{22}\cdots x_{2m} \\ \vdots \\ 1x_{n1}x_{n2}\cdots x_{nm} \end{bmatrix},$$

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_m \end{bmatrix}; \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \quad (3)$$

Where n is the sample size, that is, the series length of the forecasting quantity Y ; m is the number of factors; β is the forecasting factor coefficient, estimated by the least square method; ε is a random error.

The MLR algorithm is the most basic and straightforward in multiple regression analysis. Forecasting or estimating by the algorithm is more effective and practical than a single variable. However, the choice of factor and the expression of the factor is only a conjecture. The interpretation of the relationship between the data is often different from person to person, and different analysts will get different conclusions. Because the MLR algorithm has the characteristics of fast modeling speed and simple calculation in the case of a large amount of data, the algorithm's result is a reference result that can be quickly obtained.

The random forest (RF) algorithm [18] is a classification algorithm to create a forest by random sampling. Its essence

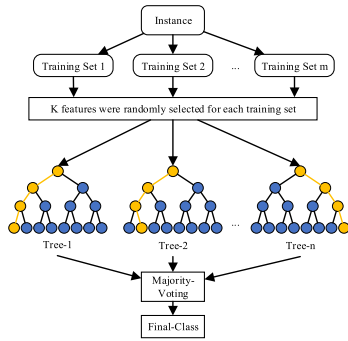


FIGURE 1. Schematic diagram of RF algorithm.

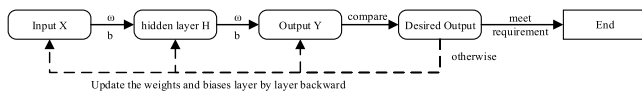


FIGURE 2. Schematic diagram of BPNN algorithm.

is a classifier combination method combining bagging integrated learning theory and decision tree classification. Where bagging is used to generate multiple classifiers/models that independently learn and make predictions, the decision tree makes decisions based on a series of questions about eigenvalues. Schematic diagram of RF algorithm is shown in Figure 1.

As a classification regression intelligence algorithm, the RF algorithm can process the data of higher dimensions, effectively fit the correlation between related variables without feature selection, and the model can still be trained efficiently when the characteristic dimension of the sample is very high. However, in some noisy sample sets, the RF algorithm easily falls into overfitting. Due to its high efficiency and ease of use, the RF algorithm is widely used in hydrological forecasting.

The back propagation neural network (BPNN) algorithm [19] is one of many artificial neural network algorithms. The BPNN’s learning process consists of forward and backward propagation. In the forward propagation process, the state of neurons in each layer only affects neurons in the next layer. When the desired output layer cannot get the expected output, the reverse propagation process will be transferred. The error signal will be returned to modify the connection weight of neurons in each layer, and the signal will result forward again. When the error reaches expectations, the network’s learning process ends. Schematic diagram of BPNN algorithm is shown in Figure 2.

The BPNN algorithm has strong nonlinear mapping ability and is particularly suitable for solving problems with complex internal mechanisms. BP neural network has a high degree of self-learning and adaptive ability. During training, it can automatically extract the output and “reasonable rules” between the output data by learning and adaptively remembering the learned content in the weight of the network. However, the BP algorithm is a local search optimization method to solve the

extreme global value of the complex nonlinear function. The algorithm will likely fall into the local extreme value, causing the training failure. The ability of network approximation and generalization is closely related to the typicality of learning samples, so selecting typical sample samples from the problems to form training sets is challenging.

MLR, RF, and BPNN are representative algorithms in machine learning. They have simple models, fast calculation speed, low learning cost and are suitable for non-computer practitioners. The three algorithms are data-driven and do not apply too much physical meaning of data, so they are complementary to the traditional physically-driven hydrological forecasting model. The three algorithms are the fundamental algorithms of machine learning and have high scalability. Therefore, this paper selects these three algorithms for improvement and comparison to research the effect of data characteristics inconsistency on medium and long-term runoff forecasting by machine learning.

2) IMPROVEMENT METHODS OF THE MODEL

In order to objectively evaluate the accuracy of the forecasting model, this paper uses the mean absolute percentage error (MAPE) value to test the accuracy of forecasting models to reflect the model’s applicability to the long-term runoff forecasting of Danjiangkou Reservoir.

When the sequence has a periodic or mutation characteristics, it represents a change in the characteristics of the runoff on the scale of time. When the sequence has a periodicity or mutation characteristics, it represents a change in the runoff characteristics. It is hard to know whether machine learning models have found the effect of these characteristics. However, modifying the sequence according to its periodicity or mutation might be an excellent decision.

Periodic characteristics are mainly the result of some large-scale factors. If there are extensive periods, the data sequence length may be only a few cycles, so that the volatility of the cycle is not enough to cause the model’s attention, but also makes the results not conform to the periodic analysis. The simulation results of machine learning are corrected based on the periodic analysis. The forecasting results are divided into different sets according to different periods, and the correction values are increased or decreased respectively so that the mean MAPE value of the set is the smallest, that is,

$$y = \min \frac{1}{n} \sum_{i=1}^n k_i = \min \frac{1}{n} \sum_{i=1}^n \left| \frac{(P_i + x) - O_i}{O_i} \right| \quad (4)$$

where y represents mean value of MAPE of different sets; k_i is the modified value of element i in different sets; n is the number of elements of different sets; O_i represents the measured value of element i ; P_i represents the forecasting value of the first factor, x is correction values.

The mutation is runoff characteristics have undergone more severe changes, that is, the sequence has multiple different characteristics of the stage, so the treatment method is divided runoff sequence into multiple segments, then analyzed separately.

III. STUDY AREA AND DATA

A. STUDY AREA

Danjiangkou Reservoir is the water source of China's South-to-North Water Diversion Project and a national-level water conservation area. The control area of the reservoir is in the middle and upper reaches of Han River in China, at 110° E and 30° N. The control basin area is 95,200 km², accounting for about 60% of the total basin area of Han River. The average annual runoff of the Danjiangkou Reservoir control area is 37.9 billion m³, accounting for about 75% of the total annual runoff of the Han River.

The control area of Danjiangkou Reservoir belongs to the East Asian subtropical monsoon climate zone, which is mainly influenced by the cold high pressure of the Eurasian continent in winter and the western Pacific subtropical high pressure in summer. Precipitation in the reservoir mainly comes from two warm and humid air currents from the southeast and southwest. Precipitation distribution during the year is not uniform.

Danjiangkou Reservoir is mainly divided into two phases to complete the construction. The initial project was completed in 1973, with a standard storage level of 157 m and a storage capacity of 17.45 billion m³. The dam-raising project started in 2005 and was completed in 2012. After the rise, the standard storage level was raised to 170 m, the reservoir capacity increased to 29.05 billion m³, and the total reservoir capacity increased to 33.91 billion m³. The main task of Danjiangkou Water Conservancy Hub is water supply and flood control, and the secondary task is power generation and shipping.

B. DATA SOURCES

The primary data applied in this paper are runoff sequence and atmospheric circulation factors. The runoff sequence is a month-by-month runoff sequence of the Danjiangkou inlet from 1969 to 2017. The main influencing factor is the Huanglongtan hydropower station, which is regulated in the upstream season of the reservoir. The runoff sequence data were provided by the Hydrological Bureau of Changjiang Water Resources Commission of The Ministry of Water Resources, China. There are 130 atmospheric circulation factors, including 88 atmospheric circulation indices, 26 SST indices, and 16 other indices such as cold air, typhoon number, and Southern Oscillation index, covering the period from 1968 to 2011. After eliminating the missing items, 97 meteorological factors were selected. The weather factor index data were provided by the National Climate Center of China (<http://data.cma.cn/>).

IV. RESULTS

A. ANALYSIS OF RUNOFF DATA FEATURE INCONSISTENCY

1) PERIODICITY

The Morlet wavelet [14] is used to transform the inflow runoff sequence of Danjiangkou Reservoir. The variation

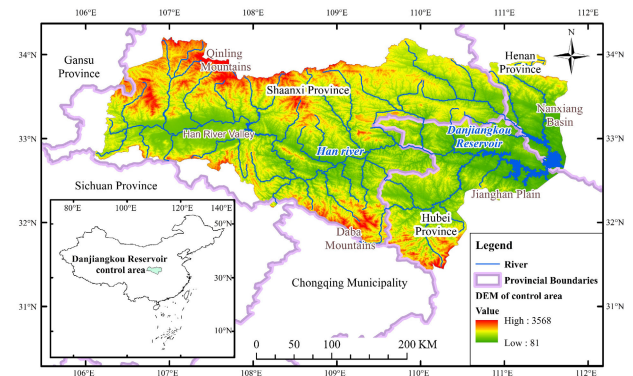


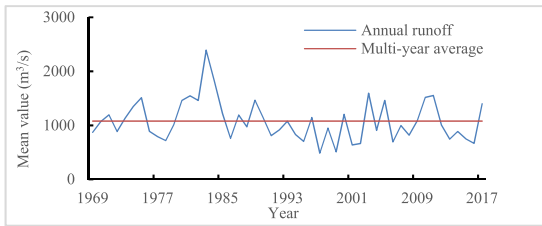
FIGURE 3. Schematic diagram of the Danjiangkou Reservoir control area.

characteristics of the inflow runoff are displayed in the wavelet transform domain and are characterized by the real number part coefficient of the wavelet transform. The fluctuation surface reflected in the wavelet transform domain is projected to the (a – b) plane in the form of contour lines. The real number part of $W_f(a, b)$ reflects the size. The positive and negative values of $W_f(a, b)$ reflect the excessive water and less water in the runoff. The alternation of peaks and valleys corresponds to the changes in the runoff. The wavelet variance diagram reflects the distribution of fluctuation energy of runoff time sequence with scale a, which is used to determine the main period existing in the runoff evolution process. Draw runoff sequence of Danjiangkou Reservoir, $W_f(a, b)$ real-time frequency diagram and wavelet variance $\text{Var}(a)$ curve of annual runoff, as shown in Figure 4.

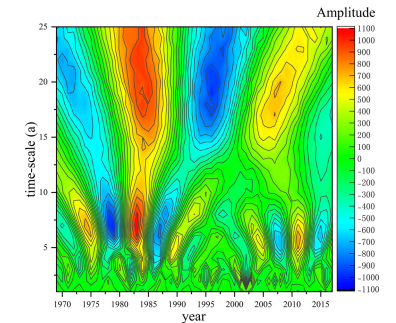
It can be seen from the Figure 4 that the annual inflow runoff of Danjiangkou Reservoir mainly shows the periodicity oscillation and variation characteristics of the two-time scales of 6-8 years and 19-22 years. The first peak value of the inflow runoff sequence is about 20 years, indicating that the periodicity oscillation around 20 years is the strongest, the first cycle of annual runoff cycle change, namely the primary cycle. The second peak value is about eight years, indicating that the periodicity oscillation of about eight years is sub-strong, which is a small periodicity oscillation. The oscillation frequency has been more intense over the years, and it is the second principal period of annual runoff periodicity change.

Related studies have shown that for the same hydrological sequence, the period value will change when intercepting different sequence lengths for wavelet analysis [20], and the period results identified by choosing different pairs of wavelet functions to vary. According to the conclusion of the Morlet wavelet cycle analysis and the actual situation, the inflow runoff in Danjiangkou Reservoir can be divided into 22-year cycles, as shown in Figure 5.

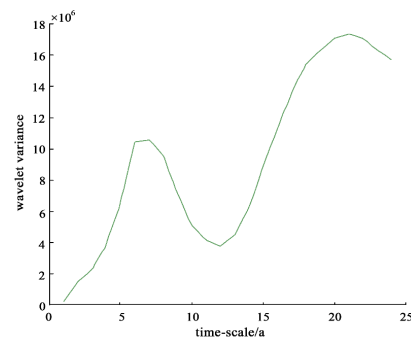
Combined with Figure 5, the statistical data for every 11 years shows that the Danjiangkou Reservoir runoff sequence has alternating oscillations above and below the mean value for about 11 years, showing a trend of



(a) Runoff sequence of Danjiangkou Reservoir



(b) Frequency of wavelet coefficients real-time



(c) Wavelet variance curve

FIGURE 4. Wavelet analysis results of runoff data series in Danjiangkou Reservoir.

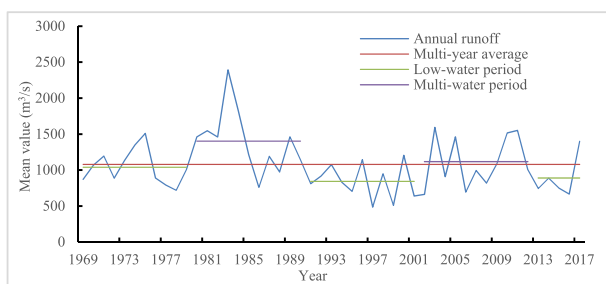


FIGURE 5. Period distribution of inflow runoff of Danjiangkou Reservoir.

“less - more - less - more.” Based on the cyclical characteristics, the runoff from Danjiangkou Reservoir can be divided into 11 years, which are called Multi-water Period and the Low-water Period. Characteristics table of quasi-periodic oscillation of annual runoff from Danjiangkou Reservoir is formed in Table 1.

Danjiangkou Reservoir inflow runoff sequence about the cycle and each period in the annual inflow is further analyzed and counted. The annual runoff was compared with the mean

TABLE 1. Characteristics table of quasi-periodic oscillation of annual runoff from Danjiangkou Reservoir.

| Period (year) | 1969-1979 | 1980-1990 | 1991-2001 | 2002-2012 | 2013-2017 | Multi-year average |
|--------------------------------|-----------|-----------|-----------|-----------|-----------|--------------------|
| Trend | Less | More | Less | More | Less | 1078.861 |
| Mean value (m ³ /s) | 1038.878 | 1402.674 | 842.674 | 1117.437 | 889.175 | |

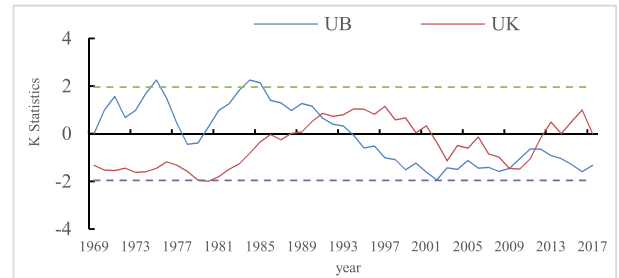


FIGURE 6. M-K statistic curve of annual runoff in Danjiangkou Reservoir.

annual runoff, and Wet Year, Normal Year, and Dry Year were divided according to the amount of water. The number of different types of years in each cycle is as follows:

In the Multi-water Period, the number of Wet Years is more than that of Dry Years and Normal Years. During 1969 ~ 1979, only one continuous two years of Dry Years occurred, but the magnitude is small, so it is considered abnormal. The proportion of Wet Years is less than that of Dry Years from 2002 to 2012. However, the water departure of Wet Years is more extensive, and the absolute value of the water departure of Dry Years is smaller, which can also better reflect the characteristics of the Multi-water Period. During the Low-water Period, the proportion of Dry Years is more significant than that of Wet Years and Normal Years, and no consecutive Wet Years have occurred.

2) MUTATION

The nonparametric M-K test method was used to identify and analyze the mutation of inflow runoff of Danjiangkou Reservoir from 1969 to 2015. When $\alpha = 0.05$, the critical value $U_{0.05} = \pm 1.96$, and the M-K statistical curve of annual runoff in Danjiangkou Reservoir is shown in Figure 6.

It can be seen from Figure 6 that UF_k and UB_k have three points of intersection, so there are three mutation points in the annual runoff sequence of Danjiangkou Reservoir during 1956 - 2017, which are 1990, 2008, and 2012, respectively, which is consistent with the construction and elevation time of reservoir engineering.

B. RESULTS OF MACHINE LEARNING

1) FACTOR SCREENING

The screening of forecasting factors is the basis of long-term runoff forecasting using machine learning, and a subset of compact and informative inputs can significantly enhance

the model performance [21]. Mathematical statistics analysis is one of the main methods of factor screening, and the single-factor correlation coefficient method effectively measures the correlation between the two variables, which is widely used [22].

The screening of forecasting factors is based on 97 meteorological factors. The single-factor correlation coefficient method sets the confidence level. The correlation between monthly runoff in 39 years from 1969 to 2017 and meteorological factors in the previous year is analyzed. According to the screening principle, five positive and five negative correlation factors with the highest absolute correlation coefficient value each month are selected as the selected factors. The primary screening principle is as follows:

- (a) Selected factors selected must be prominent relevant factors;
- (b) Each selected correlation factor appears at most once per month;
- (c) The factors of the same month in the previous year were selected up to five times each month.

2) FORECASTING RESULTS AND ERROR ANALYSIS WITHOUT CORRECTION

The model sets 1969 - 2007 as the Training Period and 2008 - 2017 as the Forecasting Period.

The MLR model for medium and long-term runoff forecasting of inflow runoff in Danjiangkou Reservoir is established using the formula between the selected factor and inflow runoff. The parameters are calibrated according to the periodic fitting of the model rate. The F-test and the multiple correlation coefficient R-test are used to determine the fitting the ten selected factors situation., Each month's β value and β_0 are determined and the factors of elimination take β is 0. The calculation formula of the multivariate linear regression model for medium and long-term runoff forecasting in Danjiangkou Reservoir is as follows:

$$y = \beta_0 + \sum_{i=1}^{10} X_i \beta_i \quad (5)$$

A RF model for medium and long-term forecasting of inflow runoff in Danjiangkou Reservoir is established. The model takes the selected factor as the input of the model and realizes the model forecasting through the random forest toolbox of MATLAB. The number of sub-forecasting models M is set to 6000, and the number of variables N selected for regression tree node division is set to 2000.

A BPNN model for medium and long-term runoff forecasting of Danjiangkou Reservoir is established. According to the screening of forecasting factors, it is determined that the input layer n is 10, the hidden layer is $2n + 2$, and the output layer is 1. The activation function selects the S-type function and assigns a random number in the interval $(-1, 1)$ to each connection weight. The target value error ε and the maximum learning number M are given, where $\varepsilon = 0.05$, the learning rate $\lambda = 0.1$, and the maximum learning number $M = 5000$.

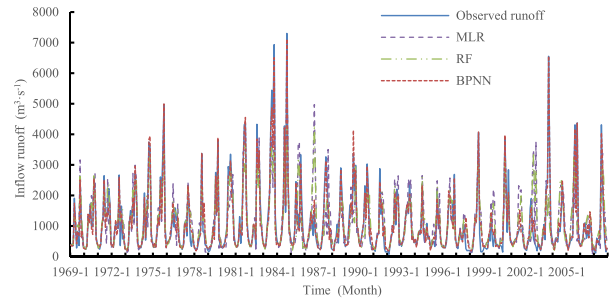


FIGURE 7. Training results obtained using different models without correction.

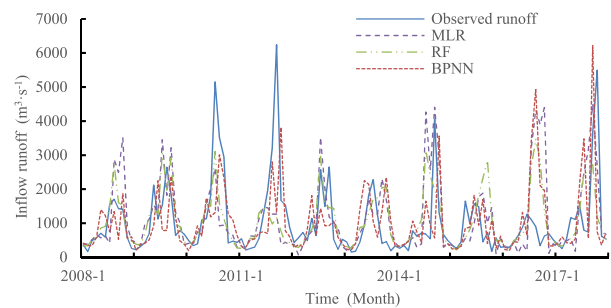


FIGURE 8. Forecasting results obtained using different models without correction.

TABLE 2. The number of different types of years in each cycle.

| Period | The number of Wet Years | The number of Normal Years | The number of Dry Years |
|----------------------------------|-------------------------|----------------------------|-------------------------|
| Low-water Period (1969-1979) | 2 | 0 | 9 |
| Multi-water Period (1980-1990) | 6 | 3 | 2 |
| Low-water Period (1991-2001) | 0 | 2 | 9 |
| Multi-water Period (2002-2012) | 4 | 3 | 4 |
| Low-water Period (2013-2017) | 1 | 1 | 3 |

Through the three methods, the training results and forecasting results of inflow runoff of Danjiangkou Reservoir are shown in Figure 7 and Figure 8.

According to Figure 7 and Figure 8, MAPE values of each month in the Training Period and the Forecasting Period are calculated, and the results are shown in Table 3.

It can be seen from Table 3. that the overall forecasting results in the Training Period are better than those in the verification period, but the forecasting results are not ideal. For each month in the Training Period, the effects of the three models are relatively sound from January to April and poor from August to November. In the Forecasting Period, the best months are January, February, June, and December, followed by May and April, and the months with poor forecasting results are between July and October. The overall forecasting results of the non-flood season are better than those of the flood season, which also shows the influence of more water and less water on the model. Comparing the three machine learning models, BPNN has the best effect in the Training

TABLE 3. MAPE of different models using different models without correction.

| Models | MLR | | RF | | BPNN | |
|--------------|-----------------|--------------------|-----------------|--------------------|-----------------|--------------------|
| | Training period | Forecasting period | Training period | Forecasting period | Training period | Forecasting period |
| Month | | | | | | |
| Jan. | 18.80% | 27.50% | 19.10% | 27.30% | 11.00% | 29.60% |
| Feb. | 28.70% | 46.70% | 26.50% | 36.30% | 21.90% | 46.20% |
| Mar. | 22.00% | 71.10% | 20.10% | 62.60% | 8.90% | 63.60% |
| Apr. | 27.90% | 52.20% | 20.00% | 46.50% | 10.00% | 61.70% |
| May. | 46.00% | 47.00% | 37.70% | 37.00% | 21.60% | 56.90% |
| Jun. | 51.60% | 27.20% | 44.10% | 36.00% | 19.30% | 44.80% |
| Jul. | 46.40% | 130.70% | 40.80% | 95.00% | 9.10% | 106.40% |
| Aug. | 77.30% | 166.90% | 66.20% | 153.50% | 28.50% | 102.50% |
| Sept. | 65.70% | 192.90% | 74.70% | 176.80% | 36.80% | 120.60% |
| Oct. | 116.40% | 229.80% | 103.10% | 150.30% | 55.50% | 195.90% |
| Nov. | 63.10% | 98.40% | 44.40% | 59.00% | 32.40% | 103.10% |
| Dec. | 30.30% | 43.60% | 27.10% | 35.20% | 13.10% | 47.50% |
| monthly mean | 49.50% | 94.50% | 43.60% | 76.30% | 22.30% | 81.60% |

TABLE 4. Statistical table of error per cycle.

| | Period | MLR | RF | BPNN |
|-------------|--------------------|------|------|------|
| Training | Multi-water Period | -8% | -9% | -3% |
| Period | Low-water Period | 9% | 9% | 2% |
| Forecasting | Multi-water Period | -10% | -13% | -19% |
| Period | Low-water Period | 40% | 27% | 25% |

Period, and RF has the best effect in the Forecasting Period. However, generally, it does not meet the accuracy requirements of engineering forecasting.

3) CORRECTION BY PERIODIC CHARACTERISTICS

The inflow runoff of Danjiangkou Reservoir has the alternate phenomenon between Multi-water Period and Low-water Period, and its cycle is 11 years, which makes the runoff in each cycle have a strong tendency to wet or dry. The error of inflow runoff forecasting of Danjiangkou Reservoir is calculated according to the periodicity law, and the results are shown in Table 4.

It can be seen from Table 4 that the forecasting values of the three models have apparent positive and negative differences under different cycles. This phenomenon shows that the forecasting results of the three models are less in the rainy season and more in the rainy season. The forecasting deviation and dry and wet characteristics show significant regularity changes. Therefore, it has specific theoretical feasibility to amend the forecasting results in different periods.

The annual runoff simulation results of the three models are modified based on the periodic characteristics. The forecasting results of the Training Period are divided into sets according to different months, periods (Multi-water Period and Low-water Period), and years (wet year, normal year, and dry year). The correction values of medium and long-term runoff forecasting based on periodicity are obtained according to the formula. Correction values of periodic characterization are shown in Table 5.

The forecasting results of machine learning are added with the correction value. Training results using different models based on the correction by periodic characteristics are shown in Figure 9. Forecasting results using different models based on the correction by periodic characteristics are shown in

TABLE 5. Correction values of periodic characterization.

| Period | Month | Wet year (mm) | | | Normal year (mm) | | | Dry year (mm) | | |
|--------------------|-------|---------------|-------|------|------------------|------|------|---------------|--------|------|
| | | MLR | RF | BPNN | MLR | RF | BPNN | MLR | RF | BPNN |
| Multi-water Period | Jan. | 4 | 6 | -2 | 114 | 73 | 29 | -68 | -60 | 14 |
| | Feb. | 19 | 1 | 13 | 123 | 131 | 63 | 23 | 26 | -3 |
| | Mar. | 87 | 79 | 32 | 43 | 38 | -16 | 85 | 72 | -19 |
| | Apr. | 70 | 25 | 50 | 308 | 227 | 37 | 136 | 62 | -55 |
| | May. | -183 | -155 | -36 | 961 | 892 | 25 | 526 | 439 | 213 |
| | Jun. | 129 | 23 | 50 | 184 | 261 | 44 | 205 | 122 | 2 |
| | Jul. | 933 | 907 | 31 | -358 | -215 | 159 | -601 | -658 | 99 |
| | Aug. | 849 | 910 | 53 | -411 | -477 | 249 | -1,147 | -1,134 | -270 |
| | Sept. | 767 | 1,018 | 240 | 131 | 153 | 37 | -658 | -1,127 | -200 |
| | Oct. | 666 | 907 | 193 | -91 | -18 | 64 | -1,400 | -1,192 | -49 |
| | Nov. | -39 | 29 | 121 | -125 | 99 | -108 | -180 | -275 | -136 |
| | Dec. | 75 | 70 | 46 | -59 | 39 | -1 | -189 | -133 | -14 |
| Low-water Period | Jan. | -8 | -7 | -17 | 2 | -45 | -37 | 4 | 11 | -7 |
| | Feb. | -21 | -11 | -30 | -43 | -62 | -47 | -20 | -6 | 0 |
| | Mar. | -52 | -50 | -13 | -67 | -80 | -25 | 8 | -11 | 6 |
| | Apr. | -115 | -59 | -35 | -56 | -37 | 27 | -73 | -7 | -27 |
| | May. | -157 | -157 | -122 | 43 | 85 | -8 | -123 | -105 | 23 |
| | Jun. | -72 | -33 | -100 | 172 | 97 | 124 | -217 | -155 | -14 |
| | Jul. | -127 | 105 | -15 | -285 | -258 | 12 | -101 | -166 | -25 |
| | Aug. | 405 | 335 | 444 | 372 | 51 | -18 | -451 | -366 | -160 |
| | Sept. | -81 | -189 | 196 | -367 | -101 | -74 | -279 | -400 | -160 |
| | Oct. | 351 | 573 | 232 | -189 | -300 | -111 | -352 | -429 | -158 |
| | Nov. | 267 | 312 | 106 | 92 | -25 | -37 | -86 | -67 | -2 |
| | Dec. | 27 | 43 | 32 | -13 | -23 | 14 | -13 | -27 | -12 |

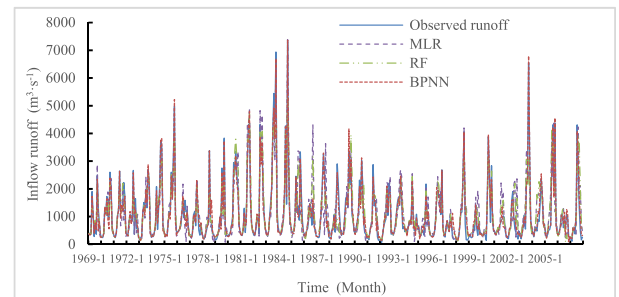


FIGURE 9. Training results obtained using different models based on the correction by periodic characteristics.

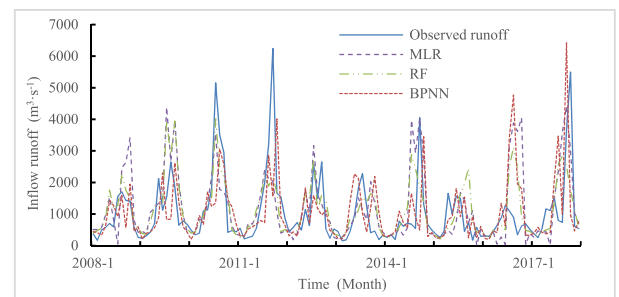


FIGURE 10. Forecasting results obtained using different models based on the correction by periodic characteristics.

Figure 10. The comparison of monthly mean relative error correction is shown in Table 6.

As seen from Table 6, after correction by periodic characteristics, the model results have been corrected to a certain extent, and the accuracy has improved slightly both in the Training and Forecasting periods. RF is the model with the greatest improvement in the Training period, while BPNN is

TABLE 6. Results of different models based on the correction by periodic characteristics.

| Period | Models | Before | After | Performance improvement rate |
|--------------------|--------|--------|--------|------------------------------|
| Training period | MLR | 49.50% | 44.00% | 5.50% |
| | RF | 43.60% | 35.90% | 7.70% |
| | BPNN | 22.30% | 20.10% | 2.30% |
| Forecasting period | MLR | 94.50% | 93.80% | 0.70% |
| | RF | 76.30% | 76.20% | 0.10% |
| | BPNN | 81.60% | 80.00% | 1.60% |

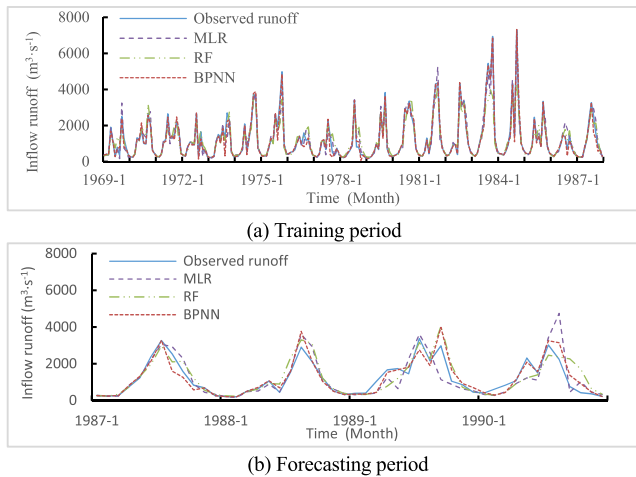


FIGURE 11. Training and forecasting results based on the improvement by mutation characteristics from 1969 to 1990.

the model with the greatest improvement in the Forecasting period.

4) IMPROVEMENT BY MUTATION CHARACTERISTICS

In addition to the periodic characteristics, the inflow runoff of Danjiangkou Reservoir has mutation characteristics. Since the Danjiangkou dam was reconstructed from 2005 to 2012, the dam heightening project may be one of the reasons for the mutation of runoff sequence, which directly leads to the change of the runoff characteristics, thus affecting the results of medium and long-term runoff forecasting.

Based on the mutation characteristics, the inflow runoff sequence of Danjiangkou Reservoir from 1969 to 2017 can be divided into four sections: 1969 - 1990, 1991 - 2008, 2009 - 2011, and 2012 - 2017. Due to a short sequence and a period of dam elevation, Runoff is likely to be unstable between 2009 and 2011, so this paper uses the three-stage sequence of 1969 - 1990, 1991 - 2008 and 2012 - 2017 to carry out medium and long-term runoff forecasting research.

The MLR, RF, and BPNN models for medium and long-term forecasting of inflow runoff in Danjiangkou Reservoir in 1969 - 1990, 1991 - 2008, and 2012 - 2017 are established respectively and take the last two years of each sequence as Forecasting Period. The simulation results of the three models are shown in Figure 11 to Figure 13.

It can be seen from Figure 11 to Figure 13 that the Training periods of the model are highly overlapping in different three periods, and the learning of data characteristics is good.

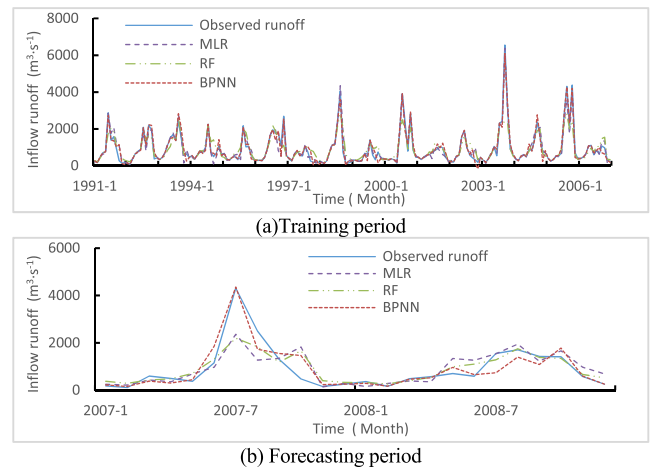


FIGURE 12. Training and forecasting results based on the improvement by mutation characteristics from 1991 to 2008.

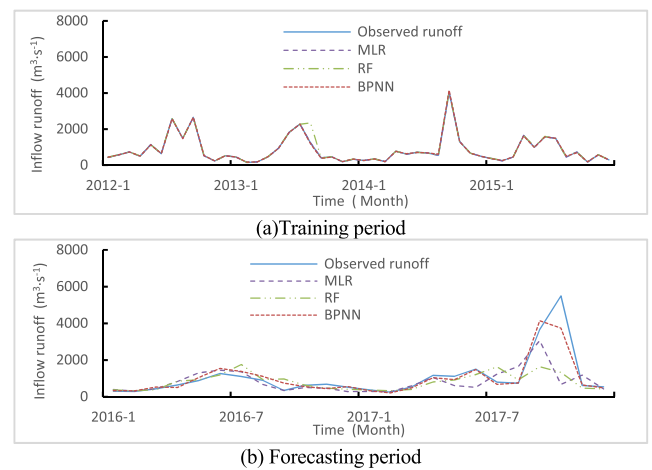


FIGURE 13. Training and forecasting results based on the improvement by mutation characteristics from 2012 to 2017.

In contrast, the Forecasting period results of the model still have some errors, but they are consistent with the observed runoff trend. The MAPE values of results of medium and long-term runoff forecasting of MLR, RF, and BPNN methods based on the improvement by mutation characteristics of Danjiangkou Reservoir are shown in Table 7.

The Training Period and Forecasting Period accuracy of the three machine learning models are improved because of the improvement by mutation characteristics. During the Training Period, MAPE values of MLR and BPNN were less than 20%, and the RF model was less than 30%, of which MAPE values of the three models were less than 1% from 2012 to 2017. During the Forecasting Period, MAPE values of MLR and RF models for 1969 - 1990 and 2012 - 2017 were less than 50%, and MAPE values for 1991 - 2008 was more than 50%; MAPE values of BPNN in the validation period were less than 30% during the three periods, and MAPE values in 2012 - 2017 were less than 20%.

TABLE 7. Results of different improved models based on the improvement by mutation characteristics.

| Models | MLR | | RF | | BPNN | |
|--------------|-----------------|--------------------|-----------------|--------------------|-----------------|--------------------|
| | Training period | Forecasting period | Training period | Forecasting period | Training period | Forecasting period |
| Jan. | 5.7% | 30.1% | 9.3% | 27.2% | 1.3% | 12.9% |
| Feb. | 8.7% | 42.5% | 12.9% | 55.4% | 1.4% | 22.1% |
| Mar. | 3.6% | 27.5% | 9.2% | 31.2% | 0.7% | 28.6% |
| Apr. | 5.8% | 24.1% | 11.4% | 22.9% | 4.0% | 15.5% |
| May. | 7.5% | 62.5% | 16.4% | 38.2% | 2.6% | 17.4% |
| Jun. | 13.4% | 48.8% | 22.5% | 26.9% | 10.6% | 20.9% |
| Jul. | 10.8% | 24.1% | 14.9% | 42.5% | 1.2% | 20.0% |
| Aug. | 14.8% | 57.2% | 24.6% | 11.8% | 17.3% | 21.2% |
| Sept. | 24.5% | 22.8% | 37.8% | 81.1% | 18.0% | 47.5% |
| Oct. | 29.6% | 94.4% | 40.1% | 120.3% | 22.9% | 75.9% |
| Nov. | 18.1% | 42.3% | 24.4% | 53.2% | 18.1% | 22.2% |
| Dec. | 10.4% | 50.9% | 13.2% | 38.8% | 7.3% | 12.4% |
| monthly mean | 12.7% | 43.9% | 19.7% | 45.8% | 8.8% | 26.4% |

C. DISCUSSION

Machine learning is one of the main methods to solve the medium and long-term runoff forecasting problem. Because of the characteristics of the medium and long-term runoff process and the demands of the hydrological industry, the analysis of data characteristics consistency is critical. This paper analyzes runoff characteristics of periodicity and mutation. Then contrasts the effect of data characteristics inconsistency on three representative machine learning models (MLR, RF, BPNN). Finally builds corresponding simple improvement methods for better use in hydrographic stations with only small sample data.

Through wavelet analysis, we found that the inflow runoff sequence of Danjiangkou Reservoir has about 11 years of fluctuating alternation between the upper and lower mean values, showing the trend of ‘less- more -more - less.’ This phenomenon is related to the periodic recharge of groundwater and global oscillations indices. However, the machine learning model does not find and learn the characteristics, so the results have apparent periodic errors. The primary possibility is that the amount of data is too small and the period is too large, so there are only five periods in the sequence, making the characteristics challenging to be mined. Through the M-K test, we found that the inflow runoff sequence of Danjiangkou Reservoir has multiple discontinuity points, which divide the overall trend of the sequence into multiple segments. In many cases, the abrupt change of runoff sequence is caused by the construction of water conservancy projects. After the construction of hydraulic engineering, the characteristics of most rivers will change significantly. There will be severe cognitive bias if the machine learning model does not pick up the change.

Due to rivers’ natural and social attributes, it is a common phenomenon that rivers have periodicity and mutation characteristics. However, because of small sample data, some features of runoff itself cannot be recognized and learned by the machine learning model. Therefore, it is suitable to study the characteristic inconsistency of runoff data.

After modifying the model, the MAPE values of the unimproved scheme, the periodic characteristics-based correction

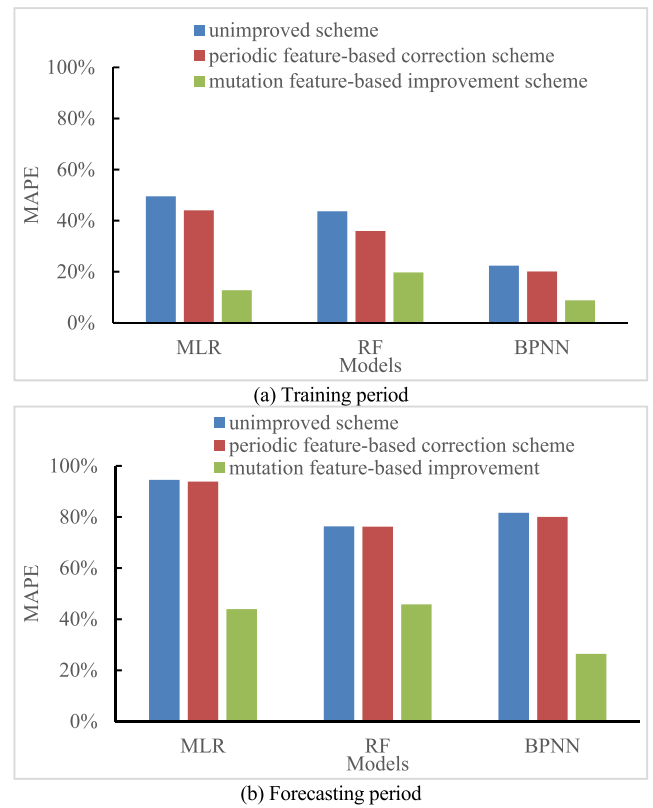


FIGURE 14. Comparison of different improvement schemes for different models.

scheme, and the mutation characteristics-based improvement scheme are compared as follows:

It can be seen from Figure 14 that when the method is not improved, the simulation effect of BPNN is better in the Training Period, followed by RF, and the simulation effect of MLR is the worst. The simulation accuracy of the three models decreased significantly in Forecasting Period, and MAPE values exceeded 70%, among which MLR had the worst accuracy. The result shows that the runoff process is not a simple linear process, so the effect of the MLR method is the worst. Moreover, there is a large gap between the effect of the Forecasting Period and the Training Period, indicating that the medium and long-term runoff process is very complex, and the simple machine learning model is challenging to simulate well. Hence, the model needs to be optimized.

After periodic characteristics correct the simulation results, the accuracy is improved to a certain extent in the Training Period, and the performance improvement rate is about 5%. The accuracy is also improved in the Forecasting Period, but the effect is not apparent; the performance improvement rate is about 1%. The reasons for the poor correction effect of models may include: a too large error in the forecasting model itself, the change of the research scale, the neglect of the impact of early climate change, and the aperiodic characteristics of the main influencing factors. The reasons make the correction value challenging to match the simulation value of the Forecasting Period perfectly. However, this correction

method still has a specific improvement effect, even if the simplest machine learning model is used.

Furthermore, it also shows that the periodic variation trend does affect the forecasting results. Therefore, the data characteristics inconsistency caused by periodic characteristics may affect the accuracy of medium and long-term runoff forecasting based on machine learning. Among them, the forecasting effect of RF is generally close to that of BPNN, and MLR is the worst. The reason may be that the over-fitting problem caused by RF and BPNN methods in the Training Period has not been well solved, and MLR determines the regression coefficient according to the principle of the least square method. Moreover, the improved method does not fully consider the independence of each factor, so the constructed set does not fully match.

The medium and long-term runoff forecasting results based on the improvement by mutation characteristics of Danjiangkou Reservoir show that the simulation accuracy of each model is significantly improved during the Training Period and Forecasting Period, and the simulated MAPE values are all less than 50%. The result shows that climate change or underlying surface change may lead to changes in the distribution of runoff, which makes the piecewise forecasting obtain better data characteristics consistency and improves the simulation accuracy. Among them, BPNN has the best forecasting effect, followed by RF, and MLR has the worst forecasting effect.

Each sequence segment becomes more linear after segmenting, so MLR has the most significant increase. However, the segmentation method also strengthens the trend consistency of each sequence and reduces the randomness. Therefore, the RF method has improved, but the improvement effect is the lowest of the three algorithms. The neural network method fits the data through a 'piecewise' linear function by establishing a hidden layer, so when the data characteristics are more consistent, the simulation effect of the neural network will be the best.

After improvement by mutation, the characteristics of inflow runoff of Danjiangkou Reservoir are more related to meteorological factors, and the consistency of sequences is good. However, because it is difficult to obtain the mutation at the end of the sequence, the applicability of the improved method is poor, and the improved method can only be used as a trend reference for medium and long-term runoff forecasting schemes.

Overall, in the study area of Danjiangkou Reservoir, the inconsistent characteristics of the inflow runoff sequence significantly impact medium and long-term runoff forecasting based on machine learning (MLR, RF, and BPNN), and the improvement of mutation characteristics is better than that of periodic characteristics. For the complex medium and long-term runoff situation, the model work and the data work are also needed. When the amount of data is insufficient to provide complete information, characteristics consistency analysis for data can effectively improve the model's accuracy. In this paper, three kinds of algorithms in machine

learning are applied to illustrate the demand and influence of data characteristics consistency. Furthermore, there is a significant improvement space in the algorithm itself. In future research, the sensitivity of better models to data characteristics consistency can be analyzed.

V. CONCLUSION

The medium and long-term runoff forecasting is the primary reference and essential link to full use of water resources, realizing the optimal operation of reservoirs, and bringing into play the economic benefits of power stations. The machine learning method has become one of the main methods to solve the problem of medium and long-term runoff forecasting. Due to the complexity of medium and long-term runoff processes and the inability to obtain entire sample attributes, in addition to the matching degree between the model and the hydrological process, the inconsistent data characteristics may also affect the accuracy of the results.

Thus, this paper takes the inflow runoff of Danjiangkou Reservoir in China as the research object and analyzes the runoff sequence' periodicity characteristics and mutation characteristics. Then, three representative machine learning models (MLR, RF, BPNN) are used for medium and long-term runoff forecasting. Finally, the influence of data characteristics consistency on machine learning based on periodicity characteristics and mutation characteristics in medium and long-term runoff forecasting are analyzed and compared. The results show that the accuracy of each model is improved to some extent. The significant findings in this paper are: (1) Machine learning results may produce errors due to inconsistent data characteristics, such as the forecasting error has the same cycle and trend as the runoff periods. (2) Correction by periodic or improvement by mutation can increase the accuracy of machine learning. (3) Among the three models, BPNN is the most affected by data characteristics consistency, followed by MLR, and RF is the least affected.

DATA AVAILABILITY

Hydrological data used to support the findings of this study were supplied by the Hydrology Bureau of Yangtze River Water Conservancy Commission of China and the National Climate Center of China (<http://data.cma.cn/>).

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

REFERENCES

- [1] R. Taormina, K.-W. Chau, and B. Sivakumar, "Neural network river forecasting through baseflow separation and binary-coded swarm optimization," *J. Hydrol.*, vol. 529, pp. 1788–1797, Oct. 2015.
- [2] S.-Y. Liong and C. Sivapragasam, "Flood stage forecasting with support vector machines," *J. Amer. Water Resour. Assoc.*, vol. 38, no. 1, pp. 173–186, Feb. 2002.
- [3] H. Huang, Z. Liang, B. Li, D. Wang, Y. Hu, and Y. Li, "Combination of multiple data-driven models for long-term monthly runoff predictions based on Bayesian model averaging," *Water Resour. Manage.*, vol. 33, no. 9, pp. 3321–3338, Jul. 2019.

- [4] H. I. Erdal and O. Karakurt, "Advancing monthly streamflow prediction accuracy of CART models using ensemble learning paradigms," *J. Hydrol.*, vol. 477, pp. 119–128, Jan. 2013.
- [5] J. Wang, P. Shi, P. Jiang, J. Hu, S. Qu, X. Chen, Y. Chen, Y. Dai, and Z. Xiao, "Application of BP neural network algorithm in traditional hydrological model for flood forecasting," *Water*, vol. 9, no. 1, p. 48, Jan. 2017.
- [6] P. Ai, Y. Song, C. Xiong, B. Chen, and Z. Yue, "A novel medium- and long-term runoff combined forecasting model based on different lag periods," *J. Hydroinformatics*, vol. 24, no. 2, pp. 367–387, Mar. 2022.
- [7] M. Shoaib, A. Y. Shamseldin, S. Khan, M. M. Khan, Z. M. Khan, T. Sultan, and B. W. Melville, "A comparative study of various hybrid wavelet feed-forward neural network models for runoff forecasting," *Water Resour. Manage.*, vol. 32, no. 1, pp. 83–103, Jan. 2018.
- [8] F. Kratzert, D. Klotz, C. Brenner, and K. Schulz, "Rainfall-runoff modelling using long short-term memory (LSTM) networks," *Hydrol. Earth Syst. Sci.*, vol. 22, no. 11, pp. 6005–6022, 2018.
- [9] Z. Yue, "Mid- and long-term runoff forecasting based on improved deep belief networks model," *J. Hydroelectr. Eng.*, vol. 39, no. 10, pp. 33–46, Jan. 2020.
- [10] Z. Yue, P. Ai, D. Yuan, and C. Xiong, "Ensemble approach for mid-long term runoff forecasting using hybrid algorithms," *J. Ambient Intell. Humanized Comput.*, vol. 13, no. 11, pp. 5103–5122, Jul. 2020.
- [11] N. Sambasivan, "‘‘Everyone wants to do the model work, not the data work’’: Data cascades in high-stakes AI," in *Proc. Chi Conf. Hum. Factors Comput. Syst.*, 2021, pp. 1–15.
- [12] Z. Xu, "Ten fundamental problems for artificial intelligence: Mathematical and physical aspects," *Scientia Sinica Informationis*, vol. 51, no. 12, pp. 1967–1978, Jan. 2021.
- [13] Y. Zhou, "Advances in the research methods of abrupt changes of hydrological sequences and their applications in drainage basins in China," *Prog. Geography*, vol. 30, no. 11, pp. 1361–1369, Jan. 2011.
- [14] E. Ghaderpour, T. Vujadinovic, and Q. K. Hassan, "Application of the least-squares wavelet software in hydrology: Athabasca river basin," *J. Hydrol., Regional Stud.*, vol. 36, Aug. 2021, Art. no. 100847.
- [15] Z. Zhou, L. Wang, A. Lin, M. Zhang, and Z. Niu, "Innovative trend analysis of solar radiation in China during 1962–2015," *Renew. Energy*, vol. 119, pp. 675–689, Apr. 2018.
- [16] M. S. Zaghoul, E. Ghaderpour, H. Dastour, B. Farjad, A. Gupta, H. Eum, G. Achari, and Q. K. Hassan, "Long term trend analysis of river flow and climate in northern Canada," *Hydrology*, vol. 9, no. 11, p. 197, Nov. 2022.
- [17] R. I. Esha and M. A. Imteaz, "Assessing the predictability of MLR models for long-term streamflow using lagged climate indices as predictors: A case study of NSW (Australia)," *Hydrol. Res.*, vol. 50, no. 1, pp. 262–281, Feb. 2019.
- [18] C. Shijun, W. Qin, Z. Yanmei, M. Guangwen, H. Xiaoyan, and W. Liang, "Medium- and long-term runoff forecasting based on a random forest regression model," *Water Supply*, vol. 20, no. 8, pp. 3658–3664, Dec. 2020.
- [19] P. Ai, Y. Song, C. Xiong, B. Chen, and Z. Yue, "A novel medium- and long-term runoff combined forecasting model based on different lag periods," *J. Hydroinformatics*, vol. 24, no. 2, pp. 367–387, Mar. 2022.
- [20] D. Wang and Y. Zhu, "Research on cryptic period of hydrologic time series based on MEM1 spectral analysis," *J. China Hydrol.*, vol. 2, pp. 19–23, Jan. 2002.
- [21] H. Lyu, M. Wan, J. Han, R. Liu, and C. Wang, "A filter feature selection method based on the maximal information coefficient and gram-Schmidt orthogonalization for biomedical data mining," *Comput. Biol. Med.*, vol. 89, pp. 264–274, Oct. 2017.
- [22] L. Yang, F. Tian, and H. Hu, "Modified esp with information on the atmospheric circulation and teleconnection incorporated and its application," *J. Tsinghua Univ. Sci. Tech.*, vol. 53, pp. 606–612, Jan. 2013.



PING AI received the Ph.D. degree in water conservancy and hydropower project from Hohai University, Nanjing, China, in 2002. He is currently a Professor with the College of Hydrology and Water Resource, Hohai University, Nanjing, China. His current research interests include hydrology and water resources, hydroinformatics, and big data. He is a Distinguished Member of Chinese Computer Society, in 2018.



CHUANSHENG XIONG received the B.S. degree in harbor, water channels and coastal engineering from the Changsha University of Science and Technology, Changsha, China, and the M.S. degree in water conservancy engineering from Hohai University, Nanjing, China, where he is currently pursuing the Ph.D. degree in hydrology and water resources engineering. His current research interests include hydroinformatics, big data, and artificial intelligence applied in water field.



KE LI received the B.S. degree in harbor, water channels and coastal engineering from the Changsha University of Science and Technology, Changsha, China, and the M.S. degree in water conservancy engineering from Hohai University, Nanjing, China. He is currently working with Chengdu Engineering Corporation Ltd., China. His research interests include water conservancy automation and water information.



YANHONG SONG received the B.S. degree in mathematics and applied mathematics from the Henan University of Urban Construction, Pingdingshan, China, and the M.S. degree in probability theory and mathematical statistics from the North China University of Water Resources and Electric Power, Zhengzhou, China. She is currently pursuing the Ph.D. degree in computer science and technology with Hohai University, Nanjing, China. Her current research interests include data mining and water conservancy information.



SHICHENG GONG received the B.S. degree in harbor, water channels and coastal engineering from the Shandong University of Science and Technology, Qingdao, China. He is currently pursuing the M.S. degree in hydrology and water resources engineering with Hohai University, Nanjing, China. His current research interests include hydroinformatics, big data, and artificial intelligence applied in water field.



ZHAOXIN YUE received the B.S. degree in electronic and information engineering from Sanjiang University, Nanjing, China, the M.S. degree in pattern recognition and intelligent system from the Sichuan University of Science and Engineering, Zigong, China, and the Ph.D. degree in hydroinformatics from Hohai University, Nanjing. He is currently working with the College of Computer and Software, Nanjing Vocational University of Industry Technology, Nanjing. His current research interests include data mining, hydrological forecasting, and water resources information.

• • •