Water Supply

IWA PUBLISHING

Water Supply Vol 22 No 4, 3921 doi: 10.2166/ws.2022.021

Runoff forecasting model based on CEEMD and combination model: a case study in the Manasi River, China

Lian Lian

College of Information Engineering, Shenyang University of Chemical Technology, Shenyang 110142, China
E-mail: lianlian_syuct@163.com

ABSTRACT

Accurate forecasting of runoff is necessary for water resources management. However, the runoff time series consists of complex nonlinear and non-stationary characteristics, which makes forecasting difficult. To contribute towards improved forecasting accuracy, a novel combination model based on complementary ensemble empirical mode decomposition (CEEMD) for runoff forecasting is proposed and applied in this paper. Firstly, the original runoff series is decomposed into a limited number of intrinsic mode functions (IMFs) and one residual based on CEEMD, which makes the runoff time series stationary. Then, approximate entropy is introduced to judge the complexity of each IMF and residual. According to the calculation results of approximate entropy, the high complexity components are predicted by Gaussian process regression (GPR), the medium complexity components are predicted by support vector machine (SVM), and the low complexity components are predicted by autoregressive integrated moving average model (ARIMA). The advantages of each forecasting model are used to forecast the appropriate components. In order to solve the problem that the forecasting performance of GPR and SVM is affected by their parameters, an improved fireworks algorithm (IFWA) is proposed to optimize the parameters of two models. Finally, the final forecasting result is obtained by adding the forecasted values of each component. The runoff data collected from the Manasi River, China is chosen as the research object. Compared with some state-of-the-art forecasting models, the comparison result curve between the forecasted value and actual value of runoff, the forecasting error, the histogram of the forecasting error distribution, the performance indicators and related statistical indicators show that the developed forecasting model has higher prediction accuracy and is able to reflect the change laws of runoff correctly.

Key words: approximate entropy, combination model, complementary ensemble empirical mode decomposition, improved fireworks algorithm, runoff forecasting

HIGHLIGHTS

- CEEMD is introduced to obtain the IMF components and residual component.
- Each IMF component and residual component is analyzed by approximate entropy, and suitable forecasting models are determined.
- An IFWA is proposed to optimize the parameters of SVM and GPR models.
- The excellent performance of the proposed model is evaluated by comparing the predictive results to other state-of-the-art comparative models.

1. INTRODUCTION

1.1. Background

Runoff forecasting is the basis of reasonable and effective management of water resources and improving the utilization rate of water resources. It allows a careful analysis of the current hydrological situation, accurate prediction of the hydrological situation in the medium and long term, and establishing the movement trend of runoff change, so as to make the most effective use of water resources and finally maximize the interests of the national economy. High precision runoff forecasting can provide a scientific basis for water resources planning, sustainable utilization of water resources, and output optimization of hydropower stations, irrigation and water environment management (Zeng *et al.* 2020). The runoff evolution process is highly complex. Changes of rainfall, climate change, solar activity, human activities and other surface factors directly or indirectly cause changes of runoff. With the increasing impact of human activities and climate change on runoff, the runoff time series presents nonlinear, non-uniformity and uncertainty. Under this background, high-precision runoff forecasting is full of

challenges, and the development of a high-precision forecasting model is the main difficulty of runoff forecasting research (Ling *et al.* 2020). It is difficult to analyze the characteristics of runoff change by traditional forecasting methods. Therefore, the study of high-precision runoff forecasting model is of great significance for the efficient utilization of water conservancy projects and the sustainable development of water resources (Xie *et al.* 2019).

The key to runoff forecasting is to correctly reveal the development law of the studied hydrological system and establish forecasting models based on it. So far, there are many runoff forecasting models, which can be roughly divided into process-driven and data-driven models. Because data-driven models generally do not need to consider the physical genetic mechanism of the runoff formation process, they are the most commonly used prediction models. However, different runoff series have their own characteristics, and the applicable forecasting models will be different. We need to analyze, try, test and use other processes, and then get a suitable forecasting model. This paper focuses on the forecasting of runoff and proposes a forecasting model based on the complementary ensemble empirical mode decomposition (CEEMD) and combination model.

1.2. Literature review

The key problem of runoff forecasting is to reveal the development law of the hydrological system and establish the forecasting model based on it. At present, there are many runoff prediction models, most of which are data-driven models. The data-driven model is the most commonly used forecasting model because it does not need to consider the physical mechanism of the runoff formation process. However, different runoff time series have their own characteristics, and the applicable forecasting models are also different. For the runoff forecasting problem, scholars usually find a suitable forecasting model after analyzing, experimenting, and testing. Runoff forecasting models can be divided into the following categories.

A. Multiple regression analysis model. Regression analysis is a model that takes into account the changes of the predicted objects affected by the impact factors, and is one of the most common models used in runoff forecasting. Because the factors affecting runoff forecasting are very complex and diverse, it is often not enough to consider only one prediction factor. In general, it is necessary to consider the common influence of multiple forecasting factors on the forecasting object. Therefore, a multiple regression model is often used in runoff time series prediction (Niu *et al.* 2016; Bitew *et al.* 2019). However, in practical engineering application, many variables or parameters affecting runoff are difficult to obtain, which will limit the application of the regression analysis model.

B. Time series forecasting model. Time series analysis models generally need to establish a dynamic model that can describe the development trend of the phenomenon. Using the extrapolation of the dynamic model in time, we can predict the trend of this phenomenon in the future. Runoff forecasting based on a time series model forecasts the future runoff by looking for the evolution law of runoff. The typical time series-based forecasting models include the autoregression model (Jiang *et al.* 2018) and autoregressive moving average model (Chua & Wong 2011) for stationary time series, and autoregressive integrated moving average (ARIMA) model (Nigam *et al.* 2014; Dhote *et al.* 2018) for non-stationary time series. Because runoff series often show nonlinear characteristics, the establishment of a nonlinear prediction model can analyze the structure characteristics of data more objectively. The threshold autoregressive model (TAR) is the most widely used and mature nonlinear time series forecasting model (Amiri 2015; Sabzevari 2017). The main disadvantage of the time series model is that it is not suitable for forecasting time series with strong nonlinearity and random characteristics. Runoff time series are affected by many external factors, so they have strong nonlinear and random characteristics, and the model based on time series has inherent defects.

C. Grey model. The grey model is a kind of system in which some information is known and the other part is unknown or unascertained. The reason why runoff forecast is regarded as a grey system is that it contains a lot of unknown information and there is a limitation of system cognition (Huang & Shen 2013; Li *et al.* 2016). However, the grey model is only applied to problems with exponential growth trend. The actual runoff time series does not conform to the exponential distribution law, which makes the prediction of runoff based on a grey model difficult.

D. Fuzzy models. The fuzzy pattern recognition forecasting method and fuzzy logic forecasting method are common fuzzy-based models in runoff forecasting. In their study, Barreto-Neto & de Souza Filho (2008) proposed a fuzzy rule-based model to estimate runoff in a tropical watershed using the Soil Conservation Service Curve Number model. The evaluation of runoff derived from fuzzy and Boolean methods demonstrated that the former provided calculated runoff closer to the measured runoff in the watershed, confirming the suitability of the fuzzy theory in modeling natural phenomena (Barreto-Neto & de Souza Filho 2008). In the literature (Mahabir *et al.* 2003), the applicability of fuzzy logic modeling techniques for forecasting water supply was investigated. By applying fuzzy logic, a water supply forecast was created that

classified potential runoff into three forecast zones. Based on the modeling results in these two basins, it is concluded that fuzzy logic has a promising potential for providing reliable water supply forecasts. However, the application of the fuzzy logic method in runoff forecasting is limited because it has obvious subjectivity when fuzzifying information.

E. Neural networks. As a forecasting model with universal approximation ability, many runoff forecasting models based on neural networks have been proposed in recent years. These forecasting models include the extreme learning machine (Niu *et al.* 2018; Cheng *et al*. 2020), RBF neural network (Wu 2018), fuzzy neural network (Shi *et al.* 2016), and Elman neural network (Li *et al.* 2019). Although the performance of the traditional neural network is excellent, it is difficult to determine the structure of the network. At the same time, the training process of the network is complex and it is easy to reach the local optimum. At the same time, the demand for sample data is very high. In recent years, with the maturity and improvement of the deep learning algorithm, many scholars have proposed some runoff forecasting models based on a deep learning model. The typical models include long-short term memory (LSTM) (X. Chen *et al.* 2020; de la Fuente *et al.* 2021), deep belief network (Yue *et al.* 2020), convolution neural network (S. Chen *et al.* 2020; Song 2020), and recurrent neural network (Shoaib *et al.* 2016). The deep learning model needs a large number of sample resources, and requires the participation of a graphics processing unit (GPU) in the training process. The cost is high and the modeling process takes a long time, which greatly limits the application of the deep learning model.

F. Support vector machine (SVM) and its related models. SVM can realize the minimization of empirical risk and structural risk. The main highlight of SVM is to solve the problems of small sample, high dimension and nonlinear pattern recognition, and it can solve the problems of over learning and dimension disaster. The least squares support vector machine (LSSVM) inherits the advantages of SVM, and transforms the quadratic programming problem in SVM into the solution of linear equations. Therefore, some SVM (Sharifi *et al*. 2017; Liang *et al.* 2018), LSSVM (Zhao *et al.* 2017) and their improved models are applied to the forecasting of runoff. However, the biggest drawback of SVM and LSSVM models is that their parameters are difficult to determine. General grid search methods are time-consuming and unstable. At present, there is no unified method to determine the optimal model parameters.

G. Combination forecasting model. With the rapid development of various forecasting models, scholars have fully recognized the advantages of each model, and clearly recognized that each model has certain disadvantages. Because the runoff is a complex large system, there are many factors affecting it, which means it is difficult to fully mine all aspects of useful information of runoff data using a single forecasting model. Therefore, a variety of combination forecasting models have been proposed. These combination forecasting models include BP neural network and particle swarm optimization-SVM (Song *et al.* 2020), coupling gated recurrent unit combining improved complete ensemble empirical decomposition with additive noise (Sibtain *et al.* 2021), group method of data handling and wavelet-based analyzed data (Moosavi *et al.* 2017), long short-term memory neural network and ant lion optimizer model (Yuan *et al.* 2018), among others. On the other hand, the combination of decomposition algorithm and some forecasting models is also the research direction of the runoff combination forecasting model. The related decomposition algorithms have variational mode decomposition (He *et al.* 2019; He *et al.* 2020), empirical mode decomposition (Zhao & Chen 2015), and ensemble empirical mode decomposition (Niu *et al.* 2019; X. Q. Zhang *et al.* 2020). For the combination forecasting model, how to choose the appropriate decomposition algorithm and how to determine which kind of forecasting model is applied to predict the components generated after decomposition are very worthy of discussion. At present, there are some problems in these combination forecasting models of runoff, such as improper selection of decomposition algorithm and forecasting model. As a whole, the combination forecasting model is a topic worthy of further discussion and research.

1.3. The contributions

According to the characteristics of randomness and complexity of runoff time series, this paper proposes a novel forecasting model based on the complementary empirical mode decomposition and combination model. (a) As an improvement of empirical mode decomposition algorithm and ensemble empirical mode decomposition algorithm, by adding positive and negative white noise to the original signal, CEEMD can not only solve the mode confusion problem of empirical mode decomposition or ensemble empirical mode decomposition, but also cancel the white noise residue. CEEMD can decompose the runoff data into relatively stable components with different frequencies. (b) The approximate entropy of components after CEEMD processing is calculated, the complexity and nonlinearity of each component is determined. The component with large approximate entropy indicates that it has high complexity, strong nonlinear and non-stationary characteristics, therefore

Gaussian process regression (GPR) is determined as the forecasting model. The component with medium approximate entropy shows that its complexity is general, and it has the characteristics of linear and nonlinear superposition, therefore SVM is selected as the prediction model. The component with small approximate entropy indicates that it has low complexity, approximate linearity and correlation, so the autoregressive integrated moving average model (ARIMA) is selected as the forecasting model. (c) An improved fireworks algorithm (IFWA) is proposed to optimize the parameters of GPR and SVM models. (d) After getting the forecasting results of each model, the final runoff forecasting value is obtained by superposition of multiple models.

In conclusion, the proposed novel forecasting model combines the ensemble signal decomposition technology, linear and nonlinear forecasting model, optimization algorithm, with multiple models fusion technology to make contributions as follows:

1. CEEMD is introduced to obtain the intrinsic mode function (IMF) components and residual component of the original runoff time series. The components after decomposing remove the long correlation and emphasize the different local characteristics of the original runoff time series. The complexity of runoff system modeling is reduced.
2. Each IMF component and residual component is analyzed by approximate entropy, and ARIMA, SVM and GPR are selected as forecasting models of different components.
3. An IFWA is proposed to solve the problem that the parameters of SVM and GPR models are difficult to determine. The forecasting accuracy of each component is further improved.
4. The excellent performance of the proposed model is evaluated by comparing the predictive results to other state-of-the-art comparative models. The effectiveness of the forecasting model is comprehensively judged by using the comparison curve, performance indicators, and the statistical significance.

2. DECOMPOSITION AND ANALYSIS OF RUNOFF BASED ON CEEMD

2.1. Dataset

In order to verify the performance of the forecasting model, this paper collected 250 groups' monthly runoff data of Hongshanzui hydrological observation station of Manasi River in Xinjiang Uygur Autonomous Region of the People's Republic of China from January 1975 to November 1995. The runoff data is shown in Figure 1.

From the results of Figure 1, it can be seen that the monthly runoff time series shows quasi periodic fluctuations. This kind of fluctuation is not a strict periodic motion, but some kind of irregular motion. Each large-scale fluctuation has a small range fluctuation, which is related to the overall dynamic characteristics of runoff. Large and violent fluctuations also have cyclical components. A large number of runoff values rise and fall sharply. The runoff between two adjacent sampling times is very different. Therefore, runoff presents periodic and random changes. CEEMD decomposes the original runoff data into several IMFs with different time scales. Choosing a suitable forecasting model for each IMF can reduce the complexity of modeling and improve the accuracy of prediction.

2.2. CEEMD algorithm

On the basis of empirical mode decomposition (EMD), CEEMD can not only solve the mode confusion problem of EMD, but also cancel the white noise residue by adding positive and negative white noise to the original signal. The main steps are as follows (Y. A. Zhang *et al.* 2020).

Step 1 By adding positive and negative white noise I_i and $-I_i$ to the original signal X_i, the synthesized signals P_i and N_i are obtained.

$$\begin{cases} P_i = X_i + I_i \\ N_i = X_i - I_i \end{cases} \tag{1}$$

Step 2 EMD (Tian & Chen 2021) is used to decompose the synthetic signal obtained from Equation (1).

$$\begin{cases} \{C_{1j}^{+}, C_{2j}^{+}, \cdots, C_{Mj}^{+}\} \\ \{C_{1j}^{-}, C_{2j}^{-}, \cdots, C_{Mj}^{-}\} \end{cases} \tag{2}$$

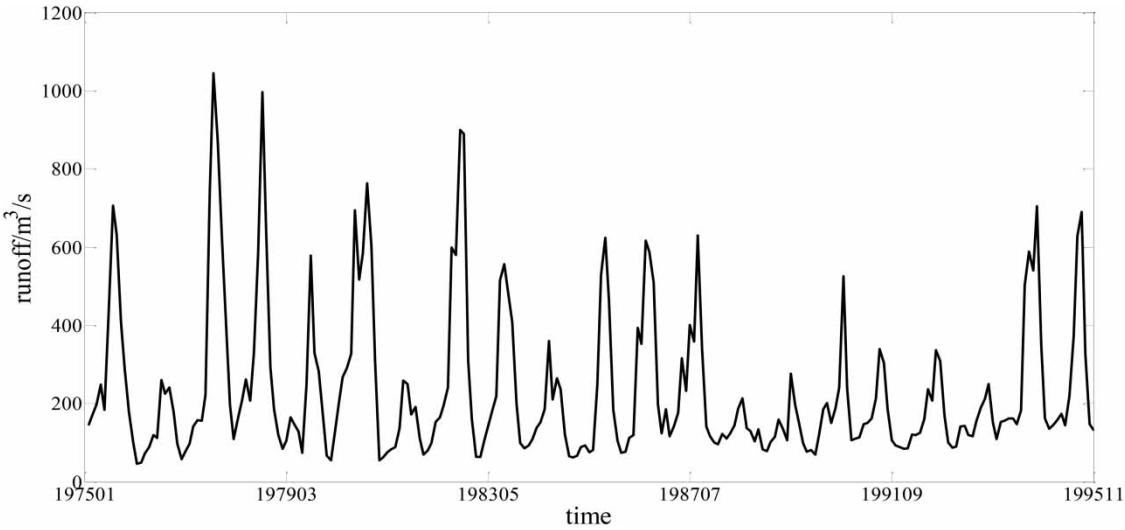

Figure 1 | Runoff time series.

where C_{ij}^{+} is the j-th IMF or residual of the synthesized signal after adding positive white noise signal in the i-th test, C_{ij}^{-} is the j-th IMF or residual of the synthesized signal after adding negative white noise signal in the i-th test, M is the sum of IMF and residual.

Step 3 Repeat steps (1) and (2) M times to obtain M IMF and residual.

$$\begin{cases} P_i = X_i + I_i \\ N_i = X_i - I_i \end{cases} \tag{3}$$

Step 4 The total mean value of all IMF and residual is calculated, which is the IMF and residual after CEEMD decomposes the runoff time series.

$$C_j = \frac{1}{2M}\sum_{i=1}^{M}(C_{ij}^{+} + C_{ij}^{-}) \tag{4}$$

Then the original runoff time series can be decomposed into the following formula.

$$X = \sum_{j=1}^{M} C_j \tag{5}$$

For the 250 groups' runoff data, the former 200 groups' data is chosen as training set, the latter 50 groups' data is chosen as test set. CEEMD is used to decompose the training set, and six IMFs and one residual component are obtained. The decomposition results are as shown in Figure 2.

From the CEEMD decomposition results of Figure 2, it can be seen that each IMF component and residual has different characteristics, some of which are complex and some are simple. Therefore, it is necessary to analyse each component to select a suitable forecasting model.

2.3. Determination of forecasting model

In this paper, the approximate entropy algorithm is introduced to analyze the complexity of each component. Approximate entropy is an algorithm used to calculate the complexity of time series (Sun *et al.* 2020). Approximate entropy describes the possibility of generating new patterns when the dimension increases from m to $m+1$ after multidimensional space reconstruction of one-dimensional time series. Approximate entropy uses a non-negative number to represent the complexity of

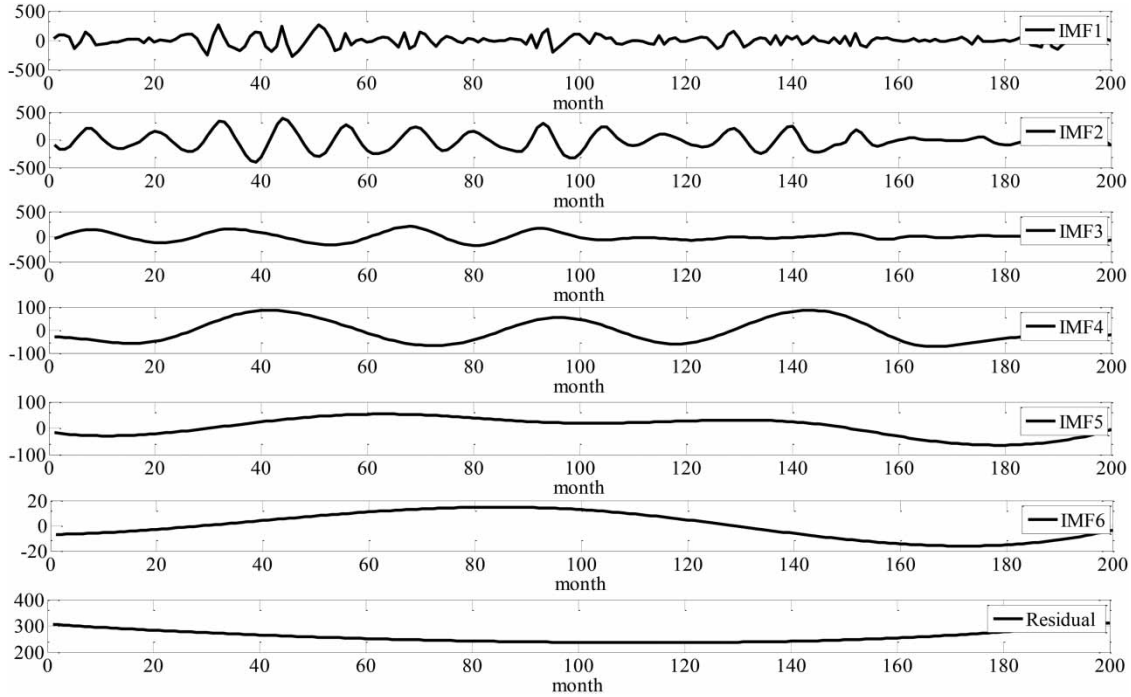

Figure 2 | The runoff decomposition results based on CEEMD.

a time series. The greater the complexity of the sequence, the greater the approximate entropy. The calculation steps of the algorithm are as follows.

1. For the time series x_i with length n, the phase space of m dimension is reconstructed, and the vector is

$$Y(i) = (x_i, x_{i+1}, \cdots, x_{i\_m-1}), \quad i = 1, 2, \cdots, n-m+1 \tag{6}$$

2. The distance d_{ij} between vectors $Y(i)$ and $Y(j)$ is defined as the maximum distance between components and can be expressed as

$$d_{ij} = \max||x_{i+k-1} - x_{j+k-1}||, \quad k = 1, 2, \cdots, m \tag{7}$$

The following formula is defined for the vector sequence $Y(i)$.

$$C_i^m(r) = \frac{\sum_{j=1}^{n-m+1} \theta(r - d_{ij})}{n-m+1} \tag{8}$$

where function $\theta()$ is 1 when the variable is greater than 0 and 0 when it is not greater than 0; $C_i^m(r)$ represents the probability that the distance between the remaining vectors $Y(i)$ and $Y(j)$ is less than r when the window length is m and the allowable deviation is r with $Y(i)$ as the center. Therefore, it represents the degree of correlation between $Y(i)(j \neq i)$ and $Y(j)$, that is, the regularity degree of vector sequence $Y(i)$.

1. Calculate the following formula.

$$h^m(r) = (n-m+1)^{-1} \sum_{i=1}^{n-m+1} \ln[C_i^m(r)] \tag{9}$$

where $h^m(r)$ is the average correlation degree of vector sequence $Y(i)$. Then the approximate entropy is

$$C_{ApEn} = h^m(r) - n^{m+1}(r) \tag{10}$$

The value of approximate entropy is related to the values of n, m and r. According to experience, when m is 2 r is 0.2 times the standard deviation value of the original time series. The obtained ApEn can be used to characterize the irregularity and complexity of time series. The approximate entropy values of each IMF component and residual are shown in Figure 3.

It can be seen from the results in Figure 3 that IMF1 and IMF2 have large approximate entropy values, so it is necessary to adopt a model with good performance for complex and strongly nonlinear objects. In this paper, GPR is selected as the forecasting model. IMF3 and IMF4 have moderate approximate entropy, and the forecasting model needs to have good nonlinear and linear fitting ability, so SVM is selected as the forecasting model. However, IMF5, IMF6 and residual have small approximate entropy, so we need to choose a model with good performance for linear non-stationary time series. ARIMA is selected as the forecasting model.

3. METHODOLOGY

In this section, ARIMA, SVM and GPR models used in the developed forecasting model are briefly introduced.

3.1. ARIMA model

The ARIMA model can effectively analyze the correlation of periodic non-stationary data sequences. The ARIMA model has good forecasting ability for linear sequences (Alzyout & Alsmirat 2020). Therefore, ARIMA is suitable for the forecasting of the IMFs with larger Hurst exponents.

For non-stationary time series, the basic approach of the ARIMA model is to make it a stationary series by using multiple difference operations; the number of differential is d. The ARIMA model with p, q as parameters is used to model the stationary sequence. After inverse transform to get the original sequence, the ARIMA prediction equation with p, d, q as parameters is expressed as:

$$y_t = \theta_0 + \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + \cdots + \varphi_p y_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \cdots - \theta_q \varepsilon_{t-q} \tag{11}$$

where y_t is the sample value of the time sequence, ϕ_i and θ_i are the model parameters, ε_t is white noise with independent normal distribution.

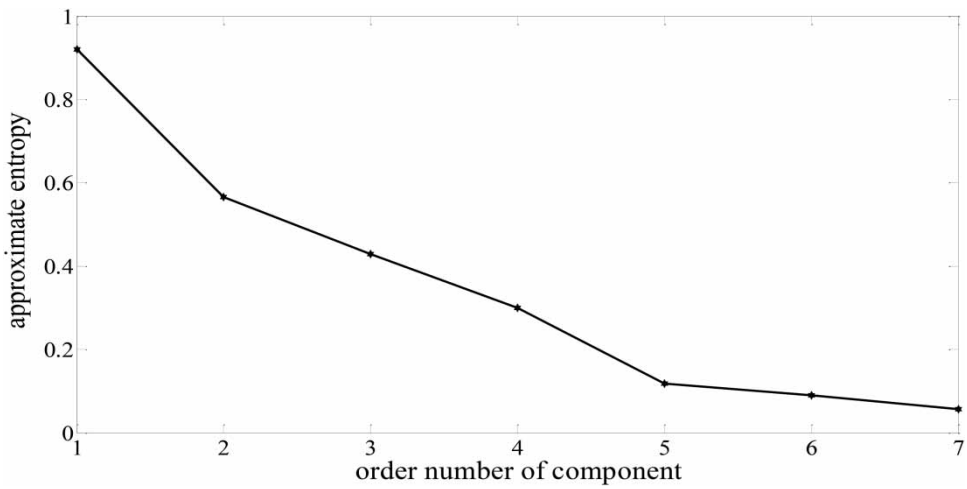

Figure 3 | The approximate entropy values of each IMF component and residual.

After smoothing the time series, the autocorrelation function (ACF) and the partial correlation function (PAC) of the original time series is calculated. For the time sequence y_t, there is auto covariance:

$$\gamma_k = \frac{1}{N}\sum_{j=1}^{N-k} y_k y_{t+k} \tag{12}$$

Autocorrelation function:

$$\rho = \frac{\gamma_k}{\gamma_0} \tag{13}$$

Partial correlation function:

$$\left\{\begin{array}{l} \alpha_{11} = \rho_1 \\ \alpha_{k+1,k+1} = (\rho_{k+1} - \sum \rho_{k+1-j}\alpha_{kj})\times \\ \qquad (1 - \sum_{j=1}^{k} \rho_j \alpha_{kj})^{-1} \\ \alpha_{k+1,j} = \alpha_{kj} - \alpha_{k+1,k+1} \times \alpha_{k,k-j+1} \end{array}\right\} \tag{14}$$

The model order can be determined through the cutoff property of ρ_k and α_k. Parameter identification of time series can be obtained by least squares estimation, through the parameters estimation of $\varphi_1, \varphi_2, \cdots, \varphi_p, \theta_1, \theta_2, \cdots, \theta_q$, making the following formula minimum:

$$\sum_{t=1}^{N} \alpha_t^2 = \sum_{t=1}^{N} (\theta_q^{-1}(Z)\varphi_p(Z)\nabla^d y_t)^2 \tag{15}$$

3.2. SVM model

SVM has good sparsity and excellent generalization, and can deal with practical problems with nonlinear, small sample, local minimum and high dimension. Some scholars point out that the performance of SVM is better than the traditional neural network algorithm (Fang *et al.* 2021). SVM is chosen as the prediction model for IMF components with median approximate entropy. Assuming the nonlinear regression function can be expressed as:

$$f(\cdot) = \boldsymbol{\omega}^{\mathrm{T}}\Phi(\cdot) + b \tag{16}$$

then the optimization problem can be expressed by the following formula:

$$\min Q(\boldsymbol{\omega}, b, \xi, \xi^*) = \frac{1}{2}\|\boldsymbol{\omega}\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \tag{17}$$

Constraint conditions are:

$$\left\{\begin{array}{l} y_i - \boldsymbol{\omega}^{\mathrm{T}}\Phi(x_i) - b \le \varepsilon + \xi_i \\ \boldsymbol{\omega}^{\mathrm{T}}\Phi(x_i) + b - y_i \le \varepsilon + \xi_i^* \\ \qquad \xi_i^*, \xi_i \ge 0 \end{array}\right. \tag{18}$$

where, C is the penalty coefficient, ξ and ξ^* are the non-negative slack variables. ξ and ξ^* are as follows:

$$\xi = \begin{cases} 0 & y - f(x) - \varepsilon \le 0 \\ y - f(x) - \varepsilon & others \end{cases} \tag{19}$$

$$\xi^* = \begin{cases} 0 & \varepsilon - y - f(x) \le 0 \\ \varepsilon - y - f(x) & others \end{cases} \tag{20}$$

where ε is an insensitive loss function.

The Lagrange function is introduced to obtain the following minimum value (Tian *et al.* 2021).

$$\min Q(\alpha_i, \alpha_i^*) = \frac{1}{2}\sum_{i,j=1}^{n}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)K(x_i \cdot x_j) + \varepsilon\sum_{i=1}^{n}(\alpha_i + \alpha_i^*) - \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)y_i \tag{21}$$

Constraint conditions are:

$$\begin{cases} \sum\limits_{i=1}^{n}(\alpha_i - \alpha_i^*) = 0, \quad i = 1, \cdots, n \\ 0 \le \alpha_i, \alpha_i^* \le C \end{cases} \tag{22}$$

$K(x_i, x_j)$ is a kernel function satisfying the Mercer conditions. The typical kernel function is the radial basis function shown in Equation (23).

$$K(x, x_i) = \exp\left(-\frac{\|x - x_i\|^2}{2\sigma^2}\right) \tag{23}$$

3.3. GPR model

GPR is a machine learning regression method developed in recent years. It has a strict statistical learning theoretical basis and is convenient to predict the development of things in the future. It has good adaptability to deal with complex problems such as high dimension, small sample size and nonlinearity, and has strong generalization ability (Xiao *et al.* 2021). Its basic principle is to assume a given training dataset $D = \{(\boldsymbol{x}_i, y_i)|i = 1, 2, \cdots, n\}$, where $\boldsymbol{x}_i$ is the i-th input vector in training data set D, y_i is the i-th target output in training data set D, and n is the number of samples in the training data set. Suppose that f is a Gaussian process, that is $f \sim GP(m, k)$, with m as mean function and k as covariance function. According to the definition of a Gaussian process, $f(\boldsymbol{x}_1),\ f(\boldsymbol{x}_2),\ \cdots,\ f(\boldsymbol{x}_n)$ obeys multivariate Gaussian distribution, and the mean vector of the multivariate Gaussian distribution is $m(\boldsymbol{x}_i)$ and the covariance matrix is $\boldsymbol{K}$.

$$f(\boldsymbol{x}_i) \sim N[m(\boldsymbol{x}_i), \boldsymbol{K}], i = 1, 2, \cdots, n \tag{24}$$

$$D : y_i = f(\boldsymbol{x}_i), i = 1, 2, \cdots, n \tag{25}$$

The actual target output $\boldsymbol{y}$ often contains some noise.

$$\boldsymbol{y} = f(\boldsymbol{x}_i) + \varepsilon_i \tag{26}$$

where $\varepsilon \sim N(0, \sigma_n^2)$. The problem is transformed into the observed training data $D : y_i = f(\boldsymbol{x}_i) + \varepsilon_i, i = 1, 2, \cdots, n$, which needs to predict the corresponding output value f_* in the test dataset $D_* = \{(\boldsymbol{x}_i, y_i)|i = n + 1, n + 2, \cdots, n + n_*\}$. The multivariate Gaussian distribution of the output vector $\boldsymbol{y}$ of the training data set and the prediction value f_* of the test dataset are as follows:

$$\begin{bmatrix} \boldsymbol{y} \\ f_* \end{bmatrix} \sim N\left(0, \begin{bmatrix} \boldsymbol{K} + \sigma_n^2\boldsymbol{I} & \boldsymbol{K}_*^{\mathrm{T}} \\ \boldsymbol{K}_* & \boldsymbol{K}_{**} \end{bmatrix}\right) \tag{27}$$

where,

$$\boldsymbol{K} = \begin{bmatrix} k(\boldsymbol{x}_1, \boldsymbol{x}_1) & k(\boldsymbol{x}_1, \boldsymbol{x}_2) & \cdots & k(\boldsymbol{x}_1, \boldsymbol{x}_n) \\ k(\boldsymbol{x}_2, \boldsymbol{x}_1) & k(\boldsymbol{x}_2, \boldsymbol{x}_2) & \cdots & k(\boldsymbol{x}_2, \boldsymbol{x}_n) \\ \vdots & \vdots & \ddots & \vdots \\ k(\boldsymbol{x}_n, \boldsymbol{x}_1) & k(\boldsymbol{x}_n, \boldsymbol{x}_2) & \cdots & k(\boldsymbol{x}_n, \boldsymbol{x}_n) \end{bmatrix}, \ \boldsymbol{K}_* = [k(\boldsymbol{x}_*, \boldsymbol{x}_1), k(\boldsymbol{x}_*, \boldsymbol{x}_2), \ \ldots, k(\boldsymbol{x}_*, \boldsymbol{x}_n)], \ \boldsymbol{K}_{**} = k(\boldsymbol{x}_*, \boldsymbol{x}_*).$$

According to the conditional distribution form of multivariate Gaussian distribution, the key formula of the Gaussian process prediction equation can be obtained as follows.

$$f_*|\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{X}_* \sim GP[m(\boldsymbol{x}_*), \mathrm{cov}(f_*)] \tag{28}$$

where, matrix $\boldsymbol{X}$ is composed of column vectors of training data input $\boldsymbol{x}_i$; matrix $\boldsymbol{X}_*$ is composed of column vectors of input $\boldsymbol{x}_{i*}$ of the test set.

$$m(\boldsymbol{x}_*) = \boldsymbol{K}_*(\boldsymbol{K} + \sigma_n^2 \boldsymbol{I})^{-1} y \tag{29}$$

$$\mathrm{cov}(f_*) = \boldsymbol{K}_{**} - \boldsymbol{K}_*(\boldsymbol{K} + \sigma_n^2 \boldsymbol{I})^{-1} \boldsymbol{K}_*^{\mathrm{T}} \tag{30}$$

In GPR, the covariance function is a symmetric function satisfying the Mercer condition. It is positive definite on a finite input set. Therefore, the covariance function is equivalent to the kernel function. Equation (29) is rewritten as follows.

$$m(\boldsymbol{x}_*) = \sum_{i=1}^{n} \boldsymbol{\alpha}_i \boldsymbol{K}(\boldsymbol{x}_i, \boldsymbol{x}_*) \tag{31}$$

where $\boldsymbol{\alpha} = [\boldsymbol{K}(\boldsymbol{X}, \boldsymbol{X}) + \delta_n^2 \boldsymbol{I}]^{-1} \boldsymbol{y}$.

In this paper, we choose the square exponential function as the kernel function.

$$\boldsymbol{K}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \delta_f^2 \exp\left(-\frac{1}{2l^2} (\boldsymbol{x}_i - \boldsymbol{x}_j)^2 \right) + \delta_n^2 \delta_{ij} \tag{32}$$

where l is the parameter of correlation measurement, δ_f^2 is the signal variance of kernel function, δ_n^2 is noise variance, δ_{ij} is the Kronecker symbol. Generally, l, δ_f^2 and δ_n^2 are called super hyper parameters. At present, the conjugate gradient method is generally used to determine the hyper parameters of GPR, but the conjugate gradient method easily reaches the local optimum by gradient descent, and the effect and convergence of the algorithm depend on the initial value.

4. IFWA ALGORITHM

From the introduction in the previous section, we can see that the performance of SVM and GPR is greatly affected by the model parameters. Inappropriate parameters will greatly affect their forecasting performance, and finally affect the forecasting accuracy of the final combined runoff. In this section, an IFWA is proposed to solve the problem of determining the parameters of the SVM and GPR.

FWA is a novel intelligent optimization algorithm proposed in 2010 (Tan & Zhu 2010). It has the characteristics of fewer parameters and better performance. Scholars have carried out in-depth research on this basis and applied it to many fields. FWA is mainly composed of explosion operator, mutation operator and selection strategy. The basic FWA has some problems such as too fast convergence and poor search performance. This paper makes the following improvements.

1. The adaptive dynamic explosion radius adjustment strategy is adopted to fully consider the evaluation information obtained in the search process, so as to avoid the situation that the explosion radius of the fireworks close to the optimal is too small and the number of sparks is greatest, which leads to easily reaching the local optimal solution. The adaptive

dynamic radius can be expressed as follows.

$$d_i = \frac{||x_i - x_{best}||}{d_{\max}} \tag{33}$$

$$A_i = A_{\min} + (A_{\max} - A_{\min})d_i \tag{34}$$

where x_i is the current position of the i-th fireworks, and x_{best} is the current position with the best fitness. $d_{\max}$ is the maximum value of x_{best} and other fireworks positions. $A_{\max}$ is the maximum allowable value of fireworks explosion radius, and $A_{\min}$ is the minimum value. The parameter setting method is as follows.

$$A_{\max} = \alpha(x_{\max} - x_{\min}) \tag{35}$$

$$A_{\min} = \beta(x_{\max} - x_{\min}) \tag{36}$$

where $x_{\max}$ and $x_{\min}$ are the upper and lower boundaries of the fireworks definition domain, α and β are scale factors. Due to the introduction of dynamic radius adjustment strategy, the explosion radius of individual fireworks will be adjusted adaptively according to the distance from the optimal fireworks location, which makes the search more refined and improves the search accuracy of the algorithm.

2. Due to the random Gaussian mutation spark, it is easy to stay near the original position or exceed the location boundary, which results in the global search ability being limited. In order to improve the global search ability of standard FWA, a differential vector is introduced according to the differential evolution algorithm to enhance the search ability of the optimal solution. The expression of the differential mutation operation in dimension k is as follows:

$$\bar{x}_{ik} = \omega_1 x_{best,k} + \omega_2(x_{j1,k} - x_{j2,k}) \tag{37}$$

where $x_{j1,k}$ and $x_{j2,k}$B are randomly selected fireworks individuals, ω_1 and ω_2 are differential scaling factors. The fitness function in the IFWA optimization process is as follows.

$$f_j = \sqrt{\sum_{i=1}^{N} \frac{1}{N}(y_i - \hat{y}_i)^2} \tag{38}$$

where N is the number of samples, y_i is the actual value of the sample, $\hat{y}_i$ is the forecasted value of the sample, j is the number of iterations. The implementation steps of IFWA proposed in this paper are as follows.

Step 1 The initialization of IFWA parameters, including the initial number of fireworks, the size of fireworks population, the maximum number of iterations, α and β.

Step 2 Set the number and value range of parameters to be optimized. The parameters to be optimized are coded, and then the first generation fireworks are generated randomly in the solution space.

Step 3 The information carried by each firework is updated into the prediction model, and the initial fitness value is calculated according to Equation (38). According to Equation (33), the dynamic explosion radius of each fireworks is calculated.

Step 4 According to Equation (37), the variation sparks with differential variables are generated randomly. The Euclidean distance between individuals of each generation is calculated, and the next generation population is updated according to a certain probability.

Step 5 Repeat Steps 3 and 4, and determine whether the termination conditions are met. After that, the optimization process is finished and the optimal parameters are output.

5. THE DEVELOPED FORECASTING MODEL

Combined with the above introduction, the diagram of the proposed forecasting model in this paper is shown in Figure 4. From Figure 4, the runoff time series is processed by EEMD. K IMF components and one residual component are produced. According to the approximate entropy of each IMF and residual, ARIMA, SVM, and GPR are introduced to predict these IMFs and residual. At the same time, IFWA is adopted to optimize the parameters of SVM (C and σ^2) and GPR (l, δ_f^2 and

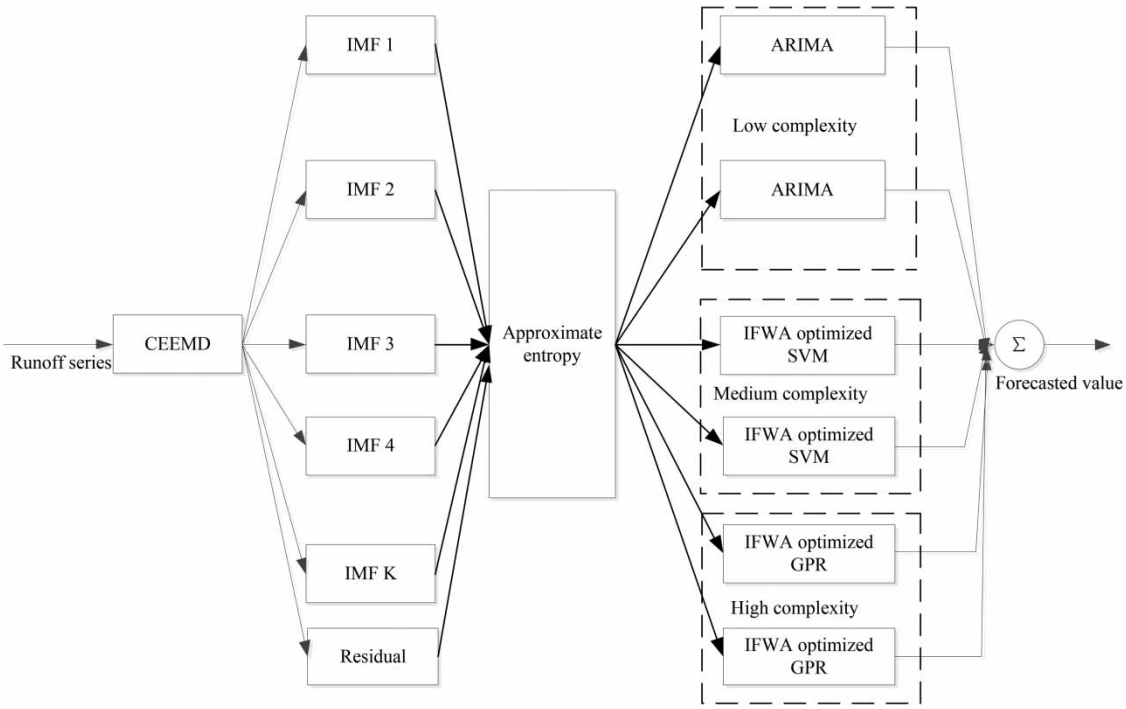

Figure 4 | The proposed forecasting model for runoff.

δ_n^2). After optimal models are obtained, the forecasted value of each forecasting model is added to get the final forecasting results. The implementation steps of the proposed forecasting model for runoff can be described as the following.

Step 1 Training process. Assuming that the length of original runoff is N. The original runoff is decomposed by CEEMD algorithm, and K IMF components and one residual component are obtained.

Step 2 The approximate entropy of IMF components is calculated. According to the different approximate entropy value, ARIMA, SVM, and GPR are chosen as the corresponding prediction model.

Step 3 ARIMA for each component is obtained by using the training set. Based on the training set, the parameters of SVM and GPR are optimized by IFWA.

Step 4 The prediction process. Assuming the current time is t, the length of input data is N. The input runoff is decomposed by CEEMD algorithm, and K IMF components and one residual are obtained. The established forecasting model is used to forecast each component. s is prediction step. The s forecasted values of each component are obtained. These s forecasted values of each model are added to obtain the final s forecasting results.

Step 5 Let $t = t + 1$. These s forecasting results are inserted at the head of the runoff queue; the oldest s ones in the tail of the runoff queue are removed. Then, return to **Step 4**, until all test set data is forecasted.

6. CASE STUDY

In this section, the effective of the proposed forecasting model for runoff is verified. The runoff dataset of Manasi River in Section 2 is used as the research object.

6.1. The performance indicators

In order to illustrate the effectiveness of the runoff forecasting model, the forecasting accuracy is measured using the following performance indicators: root mean square error (RMSE), mean absolute error (MAE), mean absolute percentile error (MAPE), relative root mean square error (RRMSE), square sum error (SSE), R^2 (R Square), Theil inequality coefficient

(TIC) and the index of agreement (IA) (Tian 2020a, 2021). The definitions of TIC and IA are as follows.

$$\mathrm{TIC} = \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(R(i)-\overline{R}(i))^2}}{\sqrt{\frac{1}{N}\sum_{i=1}^{N}R(i)^2}+\sqrt{\frac{1}{N}\sum_{i=1}^{N}\overline{R}(i)^2}} \tag{39}$$

$$\mathrm{IA} = 1 - \frac{\sum_{i=1}^{N}(R(i)-\overline{R}(i))^2}{\sum_{i=1}^{N}(|\overline{R}(i)-R\_mean|+|R(i)+R\_mean|)^2} \tag{40}$$

where, N is number of samples, $R(i)$ is actual value of runoff, $\overline{R}(i)$ is forecast value of runoff, R_mean is the average value of runoff.

In order to test the statistical significance of the forecasting model, the Wilcoxon Sign-Rank test and the Ranksum test are introduced. These two indicators can effectively evaluate the consistency between the forecast value and the actual value. Furthermore, the Pearson's test and the Diebold-Mariano (DM) test are used to test forecasting accuracy from the statistical perspective. The definition of these algorithms can be found in the corresponding references (Tian 2020b).

6.2. The results

The runoff dataset in Section 2 is chosen as the research object. We collected a total of 250 groups of runoff data. Generally, the data set of regression prediction performance of the model is divided into training set, verification set and test set, and their ratio is 6: 2: 2. Because the data set in this paper is small, and the idea of the CEEMD and combination prediction model is adopted at the same time, the verification set is not used, but only the training set and test set are used; 80% of the runoff data set is used as the training set and 20% of the runoff data set is used as the test set. Meanwhile, the order of data needs to be considered in the regression prediction of time series. There are complex associations between time series data. Therefore, the first 200 groups of runoff data are used as the training set and the last 50 groups of runoff data are used as the test set.

According to the calculation results of approximate entropy in Section 2.3, ARIMA, SVM and GPR are chosen as the forecasting model for different IMF components and residual component after decomposition of the CEEMD algorithm. The prediction step s is taken as 50. The parameters of ARIMA are determined by least squares estimation. The parameters of SVM and GPR are optimized by IFWA. In this study, for IFWA, the maximum number of iterations is set to 100, the population number is set to 30, α is 0.02, β is 0.005, ω_1 is 1, and ω_2 is 0.02. For GPR, the parameters to be optimized are $l \in [0, 10]$, $\delta_f^2 \in [0, 100]$, $\delta_n^2 \in [0, 100]$. For SVM, the parameters to be optimized are $C \in [0.001,\ 1000]$ and $\sigma^2 \in [0.001,\ 1000]$. After optimization, the parameters of these prediction models can be found in Table 1.

After the forecasting models of these components are achieved, 50 groups of test set data are tested. The forecasting results of these forecasting models for each component are shown in Figure 5. As can be seen from Figure 5, these forecasting models

Table 1 | The parameters of each forecasting model for IMF component and residual component

Component	Model	Parameters
IMF 1	GPR	l: 6.447; δ_f^2: 45.362; δ_n^2: 16.671
IMF 2	GPR	l: 4.092; δ_f^2: 61.008; δ_n^2: 24.207
IMF 3	SVM	C: 9.853; σ^2: 16.337
IMF 4	SVM	C: 12.352; σ^2: 25.408
IMF 5	ARIMA	p: 3; d: 1; q: 2
IMF 6	ARIMA	p: 2; d: 1; q: 1
Residual	ARIMA	p: 2; d: 1; q: 1

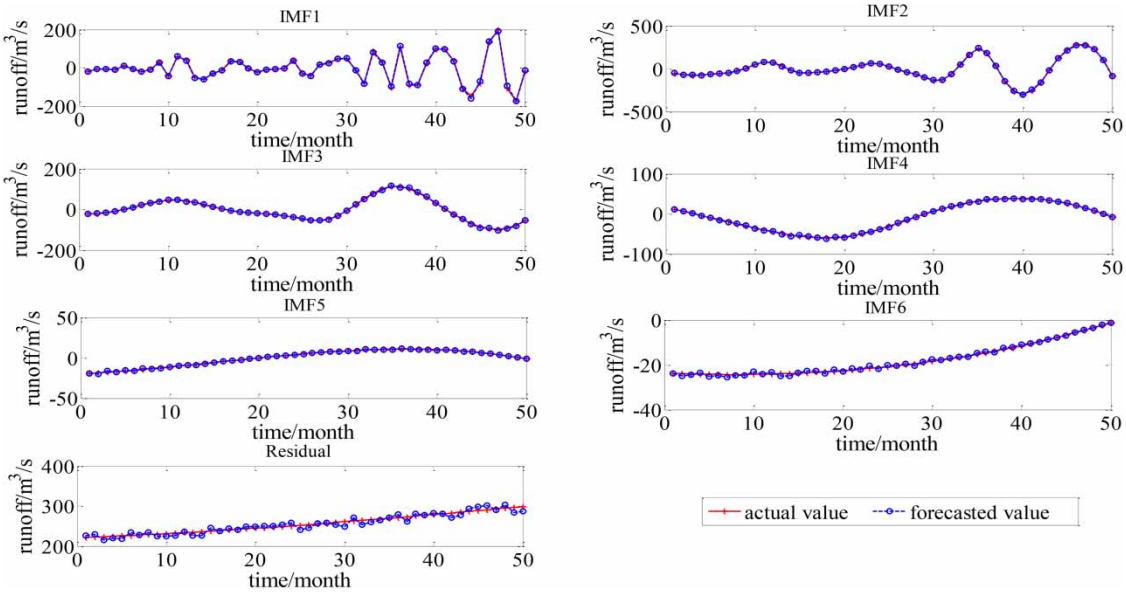

Figure 5 | The prediction results of each IMF component and residual component.

have good forecasting effect for each IMF component and residual component. Through these forecasting models, the change trend of each component can be well fitted.

When the forecasting results of each component are obtained, the final forecasting results can be obtained by adding these forecasting results. In order to verify the effectiveness of the proposed forecasting model, five state-of-the-art runoff forecasting models are chosen to compare. These comparative models are ARIMA (Dhote *et al.* 2018), TAR (Sabzevari 2017), SVM (Liang *et al.* 2018), LSTM (de la Fuente *et al.* 2021), and MEED-ARIMA (X. Q. Zhang *et al.* 2020). The detailed parameters of these comparative models are shown in Table 2.

The proposed forecasting model and other comparison models are used to predict the runoff data of 50 samples in the test set. The forecasting results are shown in Figure 6. It can be seen from the comparison results in Figure 6 that the predicted values of the proposed forecasting model can better fit the variation trend of runoff, and can accurately forecast the runoff at each sampling time.

Meanwhile, the forecasted error of each forecasting model is shown in Figure 7. The results in Figure 7 further show that the designed forecasting model has smaller prediction error, which means that the forecasting accuracy of the proposed forecasting model is the highest. It can also be seen from Figure 7 that the forecasting error range of the forecasting model in this paper is [−20.4191, 21.5435], the forecasting error range of ARIMA is [−74.7433, 36.2671], the forecasting error range of TAR is [−113.8104, 47.4017], the forecasting error range of SVM is [−30.3388, 37.6182], the forecasting error range of LSTM is [−55.4537, 39.6386], and the forecasting error range of MEED-ARIMA is [−35.3607, 37.8586]. After comparison, compared with other comparison models, the prediction error of this paper is reduced.

Figure 8 shows the histogram of the forecasting error distribution. It can be seen from this figure that the forecasting error distribution of the proposed forecasting model is more concentrated around the abscissa zero, which means that the number

Table 2 | The parameters of the comparative models

Comparative model	Parameters
ARIMA	p: 5; d: 1; q: 4
TAR	n: 5; k: 5.371
SVM	C: 25.741; σ^2: 18.007
LSTM	number of input nodes: 150; number of output nodes: 50; mini batch size: 16; L1: 200; L2: 200
MEED-ARIMA	IMF1(p: 3; d: 2; q: 2); IMF2 (p: 3; d: 1; q: 2; IMF3 (p: 3; d: 1; q: 2); Trend (p: 3; d: 1; q: 2)

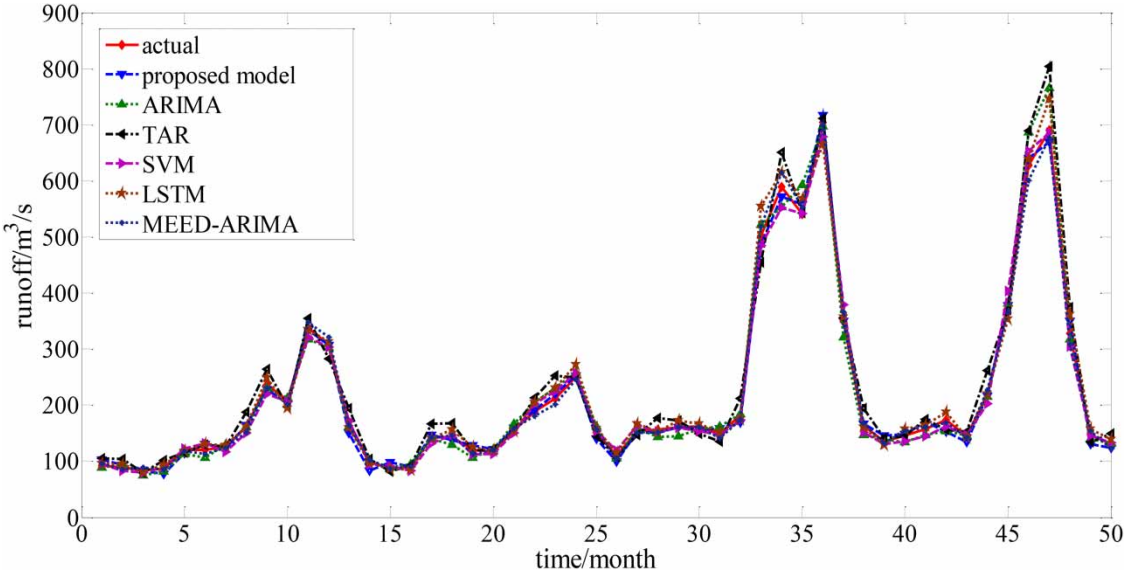

Figure 6 | The actual value and forecasted value comparison between the proposed model and comparison models.

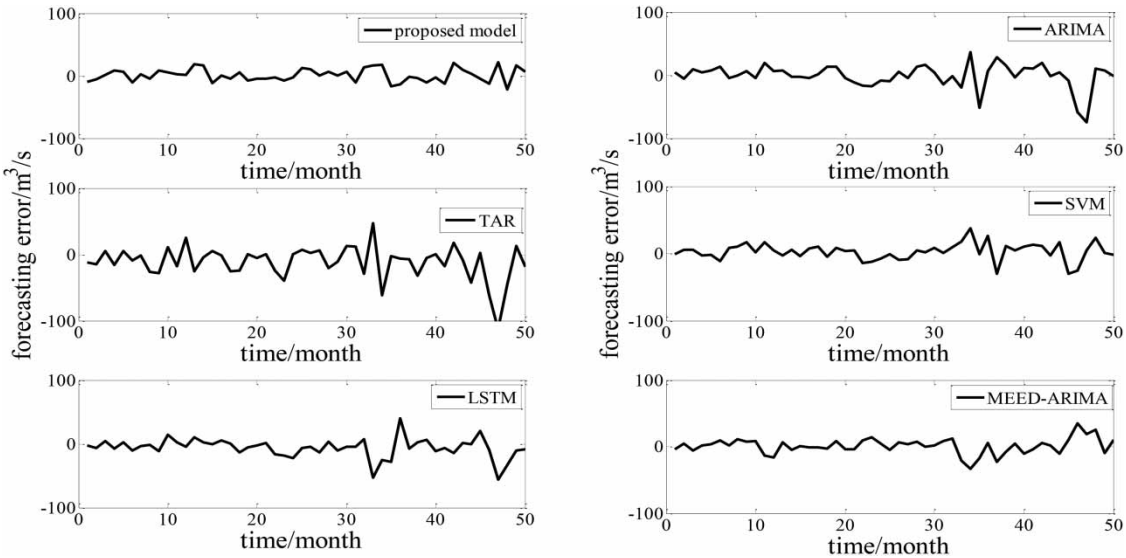

Figure 7 | The forecasting error comparison between the proposed model and comparison models.

of smaller forecasting errors is greater than that of larger forecasting errors. Therefore, the forecasting error distribution of the proposed forecasting model is centralized rather than decentralized, and the forecasting performance is better.

The box-plot of the forecasting error for each mentioned forecasting model for runoff is shown in Figure 9. As can be seen from Figure 9, compared with the existing comparison model, the proposed forecasting model has a smaller forecasting error. It is clear that the proposed forecasting model has a lower prediction error and a smaller difference. The proposed forecasting model has a higher level of stability.

Table 3 shows the comparison of the eight performance indicators between the proposed model and other comparison models. In order to show the effectiveness of the forecasting model, the performance indicators during training phases are provided in Table 4. It can be seen from the results in Tables 3 and 4 that the performance indicators of this paper are better than other comparison models in training and test phases. Specifically, as can be seen from the comparison results in this table, RMSE, MAE, MAPE, RRMSE, SSE, and TIC values of the proposed forecasting model are smaller than

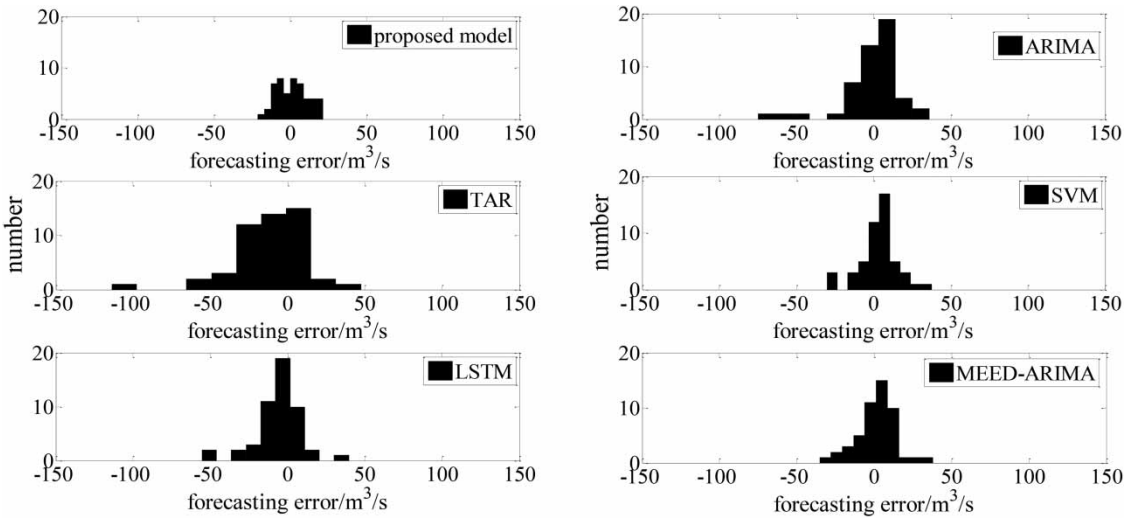

Figure 8 | Histogram of the forecasting error distribution.

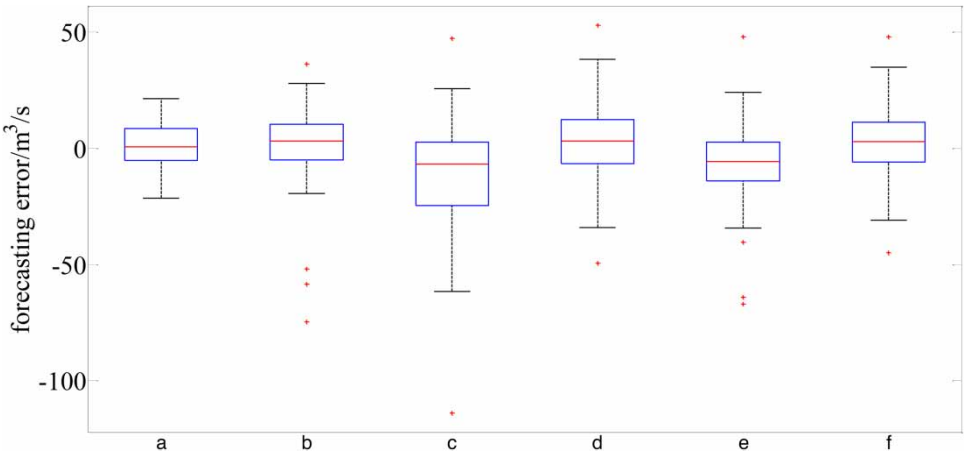

Figure 9 | The forecasting error box-plot distribution of models for runoff ((a): proposed model; (b): ARIMA; (c): TAR; (d): SVM; (e): LSTM; (f): MEED-ARIMA).

Table 3 | Comparison of the performance indicators between the proposed model and other comparison models (test phase)

Model	RMSE (m^3/s)	MAE (m^3/s)	MAPE (%)	RRMSE	SSE ($(m^3/s)^2$)	R^2	TIC	IA
Proposed model	**10.4193**	**8.5891**	**4.0122**	**0.0942**	**5.4281×10^3**	**0.9958**	**0.0190**	**0.9995**
ARIMA	19.1757	12.6628	5.5846	0.1176	18.3854×10^3	0.9872	0.0346	0.9984
TAR	27.6452	18.6944	8.6813	0.1915	38.2215×10^3	0.9741	0.0490	0.9968
SVM	12.8959	10.7082	5.6369	0.1143	8.31526×10^3	0.9914	0.0237	0.9985
LSTM	16.5479	11.1744	5.7865	0.1127	13.6924×10^3	0.9900	0.0298	0.9988
MEED-ARIMA	12.0232	10.8470	5.2037	0.1536	13.0648×10^3	0.9928	0.0288	0.9989

those of the other comparative models. The smaller the values of these performance indicators, the better the forecasting performance of the forecasting model. Meanwhile, R^2 and IA value of the proposed forecasting model are closer to 1 than the comparison models. The closer the value of R^2 and IA to 1, the better the regression forecasting performance of the model. Therefore, the forecasting accuracy of the proposed forecasting model for runoff is better than the other comparison models.

Table 4 | Comparison of the performance indicators between the proposed model and other comparison models (training phase)

Model	RMSE (m^3/s)	MAE (m^3/s)	MAPE (%)	RRMSE	SSE ($(m^3/s)^2$)	R^2	TIC	IA
Proposed model	**8.4639**	**5.4875**	**2.1084**	**0.0241**	**1.4328×10^4**	**0.9982**	**0.0135**	**0.9997**
ARIMA	20.6924	13.6110	5.7271	0.0667	8.5635×10^4	0.9900	0.0327	0.9985
TAR	33.3071	20.2336	7.9662	0.0968	2.2187×10^5	0.9755	0.0518	0.9962
SVM	16.9785	11.1892	4.4788	0.0527	5.7654×10^4	0.9930	0.0269	0.9990
LSTM	17.1655	11.2025	4.7237	0.0562	5.8931×10^4	0.9928	0.0272	0.9989
MEED-ARIMA	15.1384	9.9670	4.0823	0.0486	4.5834×10^4	0.9946	0.0272	0.9992

Pearson's test is introduced to measure the association strength between the actual value and the forecasted value of runoff. If Pearson's correlation coefficient is equal to 1, it indicates that the actual value and the prediction of the forecasting model have a linear relationship. On the other hand, if Pearson's correlation coefficient is equal to 0, there is no relationship between the actual value and the forecasted value of the forecasting model. Table 5 gives the results of the Pearson's test between the proposed model and comparison models. The result clearly shows that the result of Pearson's test of the proposed forecasting model is higher than those of the other prediction models.

The DM test values of these forecasting models for runoff are listed in Table 6. From the results in Table 6, the DM test value between the developed forecasting model and the other comparative forecasting models is greater than 0; the developed forecasting model significantly outperforms the other forecasting models at 1, 5, and 10% significance levels. Thus, it can reasonably be concluded that the developed forecasting model is superior to the other forecasting models.

In brief, the comparison between the forecasted value and the actual value, the comparison of the forecasting error and its histogram distribution, the comparison of the performance indicators, the Pearson's test and the DM test results all show that the forecasting model proposed in this paper has better forecasting accuracy and forecasting effect than other comparison models.

Table 5 | The Pearson's test results between the proposed model and comparison models

Model	Pearson's test coefficient (test phase)	Pearson's test coefficient (training phase)
Proposed model	**0.9989**	**0.9991**
ARIMA	0.9948	0.9956
TAR	0.9909	0.9902
SVM	0.9954	0.9967
LSTM	0.9961	0.9966
MEED-ARIMA	0.9983	0.9987

Table 6 | The DM test values of the forecasting models

	significance level (test phase)			significance level (training phase)		
DM (model 1, model 2)	1%	5%	10%	1%	5%	10%
DM (ARIMA, proposed model)	1.9963	2.9335	2.8536	5.1758	5.2608	5.0268
DM (TAR, proposed model)	2.4325	3.2743	3.6537	4.3627	4.8926	4.7780
DM (SVM proposed model)	2.7765	3.7485	3.7643	5.0239	5.3789	5.2364
DM (LSTM, proposed model)	2.4031	2.7643	3.8834	3.9958	4.8211	4.8752
DM (MEED-ARIMA, proposed model)	2.4436	2.7875	3.5463	3.2594	3.1687	3.2658

6.3. Discussion

From the obtained results, it can be shown that the forecasting performance of the proposed forecasting model is better than other comparative models. The analysis and discussion of the results are as follows.

1. As can be seen from Figure 6, the forecasting value of the proposed model almost coincides with the original runoff data. Meanwhile, the results in Figure 7 show that the forecasting error of the proposed model is also smaller than that of other forecasting models. In addition, the results of Figures 8 and 9 show that the forecasting error distribution of the proposed model is relatively uniform, and there are larger amounts of smaller errors.
2. The results in Tables 3 and 4 show that the RMSE, MAE, MAPE, RRMSE, SSE and TIC of the developed forecasting model are smaller than those of the comparative models. Meanwhile, the R^2 and IA of the developed forecasting model are closer to 1 than the comparative models. In these indicators, RMSE can reflect the degree of dispersion of a data set; MAE can avoid the problem of mutual cancellation of errors, so it can accurately reflect the actual prediction error; MAPE represents the percentage of the forecasting error to the actual value. A perfect forecasting model is represented by 0% of the MAPE, while a very poor forecasting model is represented by a MAPE greater than 100%; RRMSE represents the relative difference between the forecasting value and the actual value, and is a dimensionless statistic, which can be used to compare different variables; SSE calculates the sum of the error squares of the corresponding points between the fitting data and the actual data; R^2 represents the fitness of a prediction model by the change of data; TIC represents the fitting effect of the forecasting model; IA gives the ratio of the mean square error and potential error of the forecasting values to the actual value is subtracted from 1. These results show that the performance indicators of the developed forecasting model are better.
3. The results in Table 5 show that the Pearson's test results of the developed forecasting model are higher than those of the other comparative models. These results mean that compared with other forecasting models, the average probability that the forecasting value is equal to the actual value of the developed forecasting model is greater. Therefore, the forecasting value of the developed forecasting model is more consistent with the actual value of time series.
4. Table 6 shows that the DM test results of all comparative models are greater than the critical value of the 1, 5, and 10% significance level. Therefore, we can reject the null hypothesis at the 1, 5, and 10% significance level, and we believe that the developed forecasting model significantly outperforms all comparative models. Therefore, we can reasonably draw the conclusion that the developed forecasting model not only achieves higher forecasting accuracy, but also shows a significant difference in terms of the forecasting accuracy level, which further proves the superiority of the developed forecasting model.

7. CONCLUSION

This paper investigates the problem of runoff forecasting. Although many combination forecasting models are widely used to capture the statistical features of runoff, there are some issues when they are applied to applications. These issues mainly consist of the selection of the forecasting model and the optimization of forecasting model parameters. Motivated by these observations, we proposed a runoff forecasting model based on the CEEMD and combination model. In the proposed combination model, approximate entropy is used to determine the appropriate forecasting model for each IMF and residual component. And the IFWA is proposed to obtain optimal parameters of the GPR and SVM. The proposed forecasting model was evaluated on real runoff data. According to the evaluation, the proposed model can forecast runoff effectively.

However, some works are still necessary to improve the performance of the proposed model adequately, which will be our main work in the future. Extreme weather occurs frequently, and the impact of climate change on runoff is increasing, so the impact of meteorological factors on runoff cannot be ignored. In this paper, meteorological factors are not considered in runoff prediction. The next step will start from the meteorological factors, through the analysis and search for the atmospheric circulation anomaly factors which have significant impact on runoff, combined with the rainfall evaporation meteorological factors, to further carry out the meteorological-runoff forecasting research.

ACKNOWLEDGEMENTS

The authors thank the editor, the associate editor, and anonymous reviewers for their valuable comments that helped us to improve the manuscript.

FUNDING

This paper is supported by the Doctoral Scientific Research Foundation of Liaoning Province (Grant No. 20180540050).

DATA AVAILABILITY STATEMENT

Data cannot be made publicly available; readers should contact the corresponding author for details. The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request. And, the runoff data of this paper and the code of relevant standard forecasting model can be downloaded from the following link, https://zenodo.org/record/5798070#.YcMjqslBwa9.

REFERENCES

Alzyout, M. S. & Alsmirat, M. A. 2020 Performance of design options of automated ARIMA model construction for dynamic vehicle GPS location prediction. *Simulation Modelling Practice and Theory* **104**, 102148.

Amiri, E. 2015 Forecasting daily river flows using nonlinear time series models. *Journal of Hydrology* **527**, 1054–1072.

Barreto-Neto, A. A. & de Souza Filho, C. R. 2008 Application of fuzzy logic to the evaluation of runoff in a tropical watershed. *Environmental Modelling & Software* **23** (2), 244–253.

Bitew, M. M., Goodrich, D. C., Demaria, E., Heilman, P., Nichols, M., Levick, L., Unkrich, C. L. & Kautz, M. 2019 Multiparameter regression modeling for improving quality of measured rainfall and runoff data in densely instrumented watersheds. *Journal of Hydrologic Engineering* **24** (10), 04019036.

Chen, S., Dong, S. N., Cao, G. & Guo, J. T. 2020 A compound approach for monthly runoff forecasting based on multiscale analysis and deep network with sequential structure. *Water* **12** (8), 2274.

Chen, X., Huang, J. X., Han, Z., Gao, H. K., Liu, M., Li, Z. Q., Liu, X. P., Li, Q. L., Qi, H. G. & Huang, Y. G. 2020 The importance of short lag-time in the runoff forecasting model based on long short-term memory. *Journal of Hydrology* **589**, 125359.

Cheng, X., Feng, Z. K. & Niu, W. J. 2020 Forecasting monthly runoff time series by single-layer feedforward artificial neural network and grey wolf optimizer. *IEEE Access* **8**, 157346–157355.

Chua, L. H. C. & Wong, T. S. W. 2011 Runoff forecasting for an asphalt plane by Artificial Neural Networks and comparisons with kinematic wave and autoregressive moving average models. *Journal of Hydrology* **397**, 191–201.

de la Fuente, A., Meruane, V. & Meruane, C. 2021 Hydrological early warning system based on a deep learning runoff model coupled with a meteorological forecast. *Water* **11** (9), 1808.

Dhote, V., Satanand, M., Shukla, J. P. & Pandey, S. K. 2018 Runoff prediction using Big Data analytics based on ARIMA Model. *Indian Journal of Geo-Marine Sciences* **47** (11), 2163–2170.

Fang, F. X., Tian, H. & Li, Y. W. 2021 SVM strategy for mitigating low-order harmonics in isolated AC-DC matrix converter. *IEEE Transactions on Power Electronics* **36** (1), 583–596.

He, X. X., Luo, J. G., Zuo, G. G. & Xie, J. C. 2019 Daily runoff forecasting using a hybrid model based on variational mode decomposition and deep neural networks. *Water Resources Management* **33** (4), 1571–1590.

He, X. X., Luo, J. G., Li, P., Zuo, G. G. & Xie, J. C. 2020 A hybrid model based on variational mode decomposition and gradient boosting regression tree for monthly runoff forecasting. *Water Resources Management* **34** (2), 865–884.

Huang, L. M. & Shen, B. 2013 Low-flow runoff prediction using the Grey self-memory model. In *2nd International Conference on Energy and Environmental Protection*. pp. 3272–3278.

Jiang, Z. Q., Li, R. B., Ji, C. M., Li, A. Q. & Zhou, J. Z. 2018 Wavelet analysis-based projection pursuit autoregression model and its application in the runoff forecasting of Li Xiangjiang basin. *Hydrological Sciences Journal* **63** (12), 1817–1830.

Li, C. M., Zhu, L., He, Z. Y., Gao, H. M., Yang, Y., Yao, D. & Qu, X. Y. 2019 Runoff prediction method based on adaptive Elman neural network. *Water* **11** (6), 1113.

Li, C. Q., Li, Y. W., Wang, P. & Shen, J. 2016 Natural runoff prediction of the Yellow River in the future under climate change. In *International Conference on Advanced Education and Management Engineering*. pp. 177–179.

Liang, Z. M., Li, Y. J., Hu, Y. M., Li, B. Q. & Wang, J. 2018 A data-driven SVR model for long-term runoff prediction and uncertainty analysis based on the Bayesian framework. *Theoretical and Applied Climatology* **133**, 137–149.

Ling, L., Yusop, Z. & Chow, M. F. 2020 Urban flood depth estimate with a new calibrated curve number runoff prediction model. *IEEE Access* **8**, 10915–10923.

Mahabir, C., Hicks, F. E. & Fayek, A. R. 2003 Application of fuzzy logic to forecast seasonal runoff. *Hydrological Processes* **17** (18), 3749–3762.

Moosavi, V., Talebi, A. & Hadian, M. R. 2017 Development of a hybrid wavelet packet-group method of data handling (WPGMDH) model for runoff forecasting. *Water Resources Management* **31** (1), 43–59.

Nigam, R., Nigam, S. & Mittal, S. K. 2014 The river runoff forecast based on the modeling of time series. *Russian Meteorology and Hydrology* **39** (11), 750–761.

Niu, J. Y., Cao, B. Q. & Li, Y. X. 2016 A dynamic multiple regression approach for quantifying the relative impact of precipitation variations and streamflow generation conditions on runoff. *Journal of Water and Climate Change* **7** (4), 749–763.

Niu, W. J., Feng, Z. K., Cheng, C. T. & Zhou, J. Z. 2018 Forecasting daily runoff by extreme learning machine based on quantum-behaved particle swarm optimization. *Journal of Hydrologic Engineering* **23** (3), 04018002.

Niu, W. J., Feng, Z. K., Zeng, M., Feng, B. F., Min, Y. W., Cheng, C. T. & Zhang, J. Z. 2019 Forecasting reservoir monthly runoff via ensemble empirical mode decomposition and extreme learning machine optimized by an improved gravitational search algorithm. *Applied Soft Computing* **82**, 105589.

Sabzevari, T. 2017 Runoff prediction in ungauged catchments using the gamma dimensionless time-area method. *Arabian Journal of Geosciences* **10** (6), 131.

Sharifi, A., Dinpashoh, Y. & Mirabbasi, R. 2017 Daily runoff prediction using the linear and non-linear models. *Water Science and Technology* **76** (4), 793–805.

Shi, B., Hu, C. H., Yu, X. H. & Hu, X. X. 2016 New fuzzy neural network-Markov model and application in mid- to long-term runoff forecast. *Hydrological Sciences Journal* **61** (6), 1157–1169.

Shoaib, M., Shamseldin, A. Y., Melville, B. W. & Khan, M. M. 2016 A comparison between wavelet based static and dynamic neural network approaches for runoff prediction. *Journal of Hydrology* **535**, 211–215.

Sibtain, M., Li, X. S., Azam, M. I. & Bashir, H. 2021 Applicability of a three-stage hybrid model by employing a two-stage signal decomposition approach and a deep learning methodology for runoff forecasting at Swat River Catchment, Pakistan. *Polish Journal of Environmental Studies* **30** (1), 369–384.

Song, C. M. 2020 Hydrological image building using curve number and prediction and evaluation of runoff through convolution neural network. *Water* **12** (8), 2292.

Song, P. B., Liu, W. F., Sun, J. H., Wang, C., Kong, L. Z., Nong, Z. X., Lei, X. H. & Wang, H. 2020 Annual runoff forecasting based on multi-model information fusion and residual error correction in the Ganjiang River Basin. *Water* **12** (8), 2086.

Sun, L. T., Wang, X., Yang, A. F. & Huang, Z. T. 2020 Radio frequency fingerprint extraction based on multi-dimension approximate antropy. *IEEE Signal Processing Letters* **27**, 471–475.

Tan, Y. & Zhu, Y. C. 2010 Introduction to fireworks algorithm. In *1st International Conference on Advances in Swarm Intelligence*. pp. 355–364.

Tian, Z. D. 2020a Short-term wind speed prediction based on LMD and improved FA optimized combined kernel function LSSVM. *Engineering Applications of Artificial Intelligence* **91**, 103573.

Tian, Z. D. 2020b Preliminary research of chaotic characteristics and prediction of short-term wind speed time series. *International Journal of Bifurcation and Chaos* **30** (12), 2050176.

Tian, Z. D. 2021 Modes decomposition forecasting approach for ultra-short-term wind speed. *Applied Soft Computing* **105**, 107303.

Tian, Z. D. & Chen, H. 2021 Multi-step short-term wind speed prediction based on integrated multi-model fusion. *Applied Energy* **298**, 117248.

Tian, Z. D., Li, H. & Li, F. H. 2021 A combination forecasting model of wind speed based on decomposition. *Energy Reports* **7**, 1217–1233.

Wu, J. S. 2018 Co-evolution algorithm for parameter optimization of RBF neural networks for rainfall-runoff forecasting. In: *14th International Conference on Intelligent Computing*. pp. 195–206.

Xiao, F., Li, C. R., Fan, Y. X., Yang, G. R. & Tang, X. 2021 State of charge estimation for lithium-ion battery based on Gaussian process regression with deep recurrent kernel. *International Journal of Electrical Power & Energy Systems* **124**, 106369.

Xie, T., Zhang, G., Hou, J. W., Xie, J. C., Lv, M. & Liu, F. C. 2019 Hybrid forecasting model for non-stationary daily runoff series: a case study in the Han River Basin, China. *Journal of Hydrology* **577**, 123915.

Yuan, X. H., Chen, C., Lei, X. H., Yuan, Y. B. & Adnan, R. M. 2018 Monthly runoff forecasting based on LSTM-ALO model. *Stochastic Environmental Research and Risk Assessment* **32** (8), 2199–2212.

Yue, Z. X., Ai, P., Xiong, C. S., Hong, M. & Song, Y. H. 2020 Mid- to long-term runoff prediction by combining the deep belief network and partial least-squares regression. *Journal of Hydroinformatics* **22** (5), 1283–1305.

Zeng, F., Ma, M. G., Di, D. R. & Shi, W. Y. 2020 Separating the impacts of climate change and human activities on runoff: a review of method and application. *Water* **12** (8), 2201.

Zhang, X. Q., Tuo, W. & Song, C. 2020 Application of MEEMD-ARIMA combining model for annual runoff prediction in the Lower Yellow River. *Journal of Water and Climate Change* **11** (3), 865–876.

Zhang, Y. A., Yan, B. B. & Aasma, M. 2020 A novel deep learning framework: prediction and analysis of financial time series using CEEMD and LSTM. *Expert Systems with Applications* **159**, 113609.

Zhao, X. H. & Chen, X. 2015 Auto regressive and ensemble empirical mode decomposition hybrid model for annual runoff forecasting. *Water Resources Management* **29** (8), 2913–2926.

Zhao, X. H., Chen, X., Xu, Y. X., Xi, D. J., Zhang, Y. B. & Zheng, X. Q. 2017 An EMD-based chaotic least squares support vector machine hybrid model for annual runoff forecasting. *Water* **9** (3), 153.

First received 11 October 2021; accepted in revised form 24 December 2021. Available online 19 January 2022