Monthly runoff prediction using modified CEEMD-based weighted integrated model

Xinqing Yan, Yuan Chang, Yang Yang and Xuemei Liu

ABSTRACT

Due to the nonlinear characteristics of runoff data and the poor performance of the single prediction model, a weighted integrated modified complementary ensemble empirical mode decomposition (MCEEMD)-based model was proposed to predict the monthly runoff of three hydrological stations in the lower reaches of the Yellow River. In this model, particle swarm optimization (PSO) was used to optimize the parameters of support vector regression (SVR), back propagation neural network (BP), long short-term memory neural network (LSTM) that constitute it. The weight coefficients and frequency terms decomposed by MCEEMD were used to obtain the final prediction results. Results indicated that this model performs better than other models, with the Nash–Sutcliffe efficiency (NSE) reaching above 0.92, qualification rate (QR) reaching above 75% and all error indicators being minimal. In addition, considering the influence of extreme weather and climate anomalies, the integrated model combined the atmospheric circulation anomalies factors and the results can still be improved. It can be verified that this weighted integrated model can be used for the stable and accurate predication of medium- and long-term runoff.

Key words | integrated model, modified complementary ensemble empirical mode decomposition, monthly runoff prediction, particle swarm optimization, weight coefficient

HIGHLIGHTS

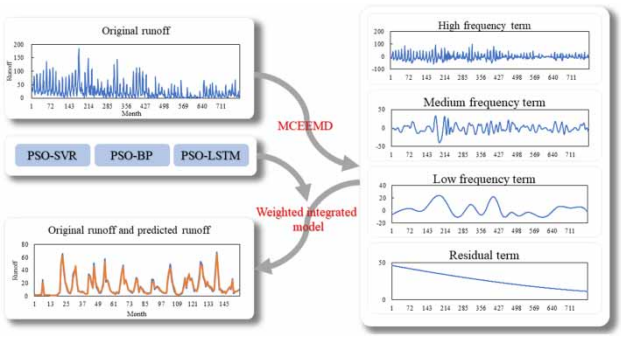
- Decomposition of runoff into smooth sequences using MCEEMD reduces the accumulation of errors associated with the CEEMD.
- A weighted integrated method was used to develop predictive models based on the MCEEMD decomposition.
- In view of the influence of extreme weather and abnormal climate on runoff prediction accuracy, the atmospheric circulation anomaly factors were used as the input data of the model to predict.

Xinqing Yan
Yuan Chang (corresponding author)
Xuemei Liu
School of Information Engineering,
North China University of Water Resources and Electric Power,
Zhengzhou,
China
E-mail: *786946984@qq.com*

Yang Yang
School of Water Conservancy,
North China University of Water Resources and Electric Power,
Zhengzhou,
China

doi: 10.2166/wcc.2020.274

GRAPHICAL ABSTRACT



INTRODUCTION

Runoff is an important condition of regional industrial and agricultural water supply, and also a restricting factor of regional socio-economic development. The measurement, calculation, and forecast of runoff are important tasks of water conservancy construction. For example, they are of great significance for the operation, planning, and dispatching of hydropower stations to predict monthly runoff accurately (Huang *et al.* 2015a, 2015b). Runoff prediction has received a lot of attention in recent years.

The traditional model usually uses mathematical statistics. Typically, Tesfaye *et al.* (2006) used a periodic autoregressive moving average model to produce a realistic simulation for monthly runoff data of the Fraser River in British Columbia. Ouyang *et al.* (2010) used the dynamic time warping distance method to perform a similarity search of floods at Shaliguilanke station in the Tarim River basin. With the development of data science, there are more and more studies on runoff prediction based on machine learning and neural networks. Typically, Cui (2013a, 2013b) proposed to improve the Elman neural network prediction model and the hidden layer back propagation neural network (BP) prediction model. Lu and Zhou (2014) screened input factors of a forecast model by using the mutual information method. In the BP neural network model, mean square error and mutual information are used as objective functions to measure the correlation between factors, optimize the final prediction factors, and apply the prediction factors to the runoff prediction of Biliu River in flood season; the model can identify multiple complex correlations between forecast factors and forecast quantity. Xiang *et al.* (2020) used long short-term memory neural network (LSTM) and the seq2seq structure to estimate hourly rainfall-runoff.

However, the generation process of runoff tends to be uncertain, highly nonlinear, and time varying, especially when extreme weather appears, thus the monthly streamflow series contains different frequency components (Huang *et al.* 2015a, 2015b). Similarly, the nonlinear character of runoff contradicts the data requirements of many models. Moreover, the single prediction model (i.e., a model using only one algorithm) has limitations, such as the over-fitting phenomenon and local optimality in neural network, and there is no uniform standard in kernel function selection and parameter calibration in the support vector machine method (Han *et al.* 2017).

In order to solve the above defects, this study has developed a weighted integrated model based on the modified complementary ensemble empirical mode decomposition (MCEEMD) to predict monthly runoff. In this study, the particle swarm optimization optimize support vector regression (PSO-SVR), particle swarm optimization optimize back propagation neural network (PSO-BP), and particle swarm optimization optimize long short-term memory neural network (PSO-LSTM) methods with excellent runoff prediction performance were selected as the compositions of the integrated model, and the runoff data

decomposed by MCEEMD were used as the training data. Then, the weights of each model's frequency terms were calculated according to the errors of the single algorithms, and the runoff prediction results of the weighted integrated model were finally obtained by adding each single model's weighted frequency terms. In addition, the optimal factors of atmospheric circulation anomaly factors were selected as the additional item of training data of the integrated model to predict runoff, which improved the prediction accuracy and stability of the model in extreme weather periods. Compared with other models, the weighted integrated model had a better prediction performance, especially during periods of extreme weather and weather anomalies, which can provide a reference for regional water resources allocation and regional water resources optimization. It also provided a new idea for the development of hydrological prediction.

METHODS

Modified complementary ensemble empirical mode decomposition (MCEEMD)

Huang *et al.* (1998) of the National Aeronautics and Space Administration proposed a method of data decomposition called empirical mode decomposition (EMD), which essentially identifies all vibration modes contained in a signal through characteristic time scales. The basic principle of EMD is to decompose the complex sequence of input into a finite number of intrinsic mode functions (IMF) from high to low frequencies and a residual term. The decomposed IMF contains local characteristic signals of the original signal at different time scales, and reduces the mutual interference between different trend information. IMF must meet with two characteristics:

1. the number of extremum points and zero crossings must be the same or at most one different throughout the data segment;
2. in any case, the mean value of the envelope defined by the local maxima and that of the local minima is zero.

However, EMD will lead to the frequent occurrence of mode aliasing in the decomposition process, which will undermine the physical significance of IMF (Han *et al.* 2017). Huang & Wu (2008) proposed the ensemble empirical mode decomposition (EEMD) to solve the defects of EMD. Although EEMD effectively solves the modal aliasing phenomenon of EMD by adding white noise, it is affected by residual noise in signal reconstruction. Yeh *et al.* (2010) improved the EEMD method and proposed the complementary ensemble empirical mode decomposition (CEEMD), which not only solved the problem of mode aliasing that EMD is prone to, but also avoided the influence of residual noise of EEMD in signal reconstruction. In recent years, CEEMD has been applied to many models for runoff prediction (Zhang *et al.* 2019).

However, the components obtained after the decomposition of CEEMD are too much, especially the nonlinear and non-stationary features of IMF1, which will bring a large number of calculations and error accumulation to the prediction model. MCEEMD reconstructed the IMFs decomposed by CEEMD through the fluctuation frequency and amplitude to obtain the high frequency (HF) term, medium frequency (MF) term, low frequency (LF) term, and residual term (Res) to address the above deficiencies. The process of decomposition for MCEEMD is as follows:

1. Adding a random sequence of positive and negative white noise $n_i(t)$ (i.e., the original signal plus the white noise plus the original signal minus the white noise) with the same amplitude and phase angle difference of π to the original signal $x(t)$. Adding a new noise with the same amplitude every time to get a new signal.

$$\begin{cases} m_i^+(t) = x(t) + n_i(t) \\ m_i^-(t) = x(t) - n_i(t) \end{cases} \tag{1}$$

2. EMD is used to decompose $m_i^+(t)$ and $m_i^-(t)$ to obtain two sets of signals composed of 2N IMFs components $C_j(t)$ and a residual term $r^i(t)$.

$$\begin{cases} m_i^+(t) = \sum_{j=1}^{N} C_j^{i+}(t) + r^{i+}(t) \\ m_i^-(t) = \sum_{j=1}^{N} C_j^{i-}(t) + r^{i-}(t) \end{cases} \tag{2}$$

3. Repeating step 1 and step 2, adding different positive and negative white noise sequences each time to get M (i.e.,

number of times white noise was added) group of IMF components and residual terms.

4. The IMF $C_j(t)$ and residual $r_j(t)$ of the original signal are averaged by the decomposition results.

$$\begin{cases} C_j(t) = \dfrac{1}{2M}\sum_{i=1}^{M}[C_j^{i+}(t) + C_j^{i-}(t)] \\ r_j(t) = \dfrac{1}{2M}\sum_{i=1}^{M}[r_j^{i+}(t) + r_j^{i-}(t)] \end{cases} \tag{3}$$

5. According to the fluctuation frequency and amplitude of IMF, the high frequency term $H(t)$, medium frequency term $M(t)$, and low frequency term $L(t)$ are obtained.

$$\begin{cases} H(t) = \sum_{j=1}^{2} C_j(t) \\ M(t) = \sum_{j=3}^{4} C_j(t) \\ L(t) = \sum_{j=5}^{7} C_j(t) \end{cases} \tag{4}$$

6. The original signal decomposition is expressed as follows:

$$x(t) = H(t) + M(t) + L(t) + r(t) \tag{5}$$

Support vector regression (SVR)

Support vector regression (SVR) is an application of SVM in regression prediction proposed by Vapnik (1998). SVR maps the original data to a new feature space through nonlinear mapping. In the feature's new space, a linear function can be found that constructs the mathematical relationship between the input and output values, and predicts the value through this function. SVR has high accuracy and strong generalization ability, so it is widely used in runoff prediction (Huang *et al.* 2015a, 2015b; Chu *et al.* 2016). The calculation formula of SVR is as follows:

$$f(x) = \sum_{i=1}^{l}(\alpha_i - \alpha_i*)K(x_i, x) + b \tag{6}$$

where l is the number of the data set; α_i and α_i* are lagrange multiplier; $K(x_i, x)$ is the kernel function; b is the deviation between the true value and the predicted value.

The solution of function $f(x)$ can be transformed into an optimized process:

$$\begin{cases} \min\left(\dfrac{1}{2}w^{\mathrm{T}}w\right) + C\sum_{i=1}^{n}(\xi_i + \xi_i*) \\ \text{s.t.} \quad \begin{array}{l} y_i - w\varphi(x_i) - b \le \varepsilon + \xi_i \\ w^{\mathrm{T}}\varphi(x_i) + b - y_i \le \varepsilon + \xi_i* \\ \xi_i \ge 0, \xi_i* \ge 0, i = 1, 2, \ldots, n \end{array} \end{cases} \tag{7}$$

where w is the weight vector; ξ_i and ξ_i* are relaxation factor; C is the penalty coefficient, an appropriate C can make the model have a better generalization ability; $\varphi(x_i)$ is the mapping from the input space to the feature space; ε is the constant deviation; n is the sample size. The optimized process is subjected to the constraints under it.

Back propagation neural network (BP)

BP is a multi-layer feed-forward network trained by error back propagation. BP uses the gradient search technology to minimize the mean square error between the actual output value and the expected output value of the network. BP is usually combined with other algorithms for runoff prediction (Lu & Zhou 2014). The structure of BP is shown in Figure 1.

The input signal X through the hidden layer node affects the output node, each nerve input includes input vector and

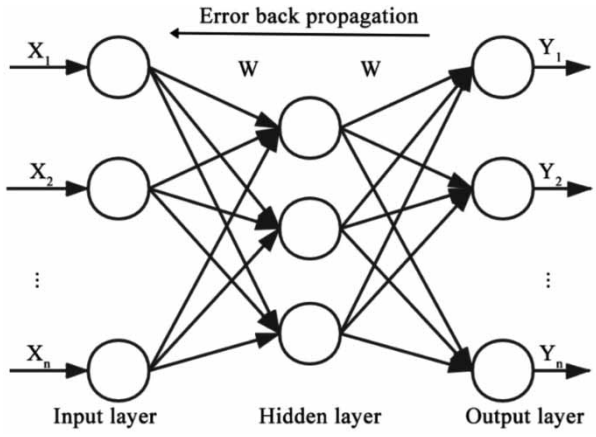


Figure 1 | The structure of back propagation neural network (BP).

the desired output vector. If the deviation vector is not as expected, the weight W and threshold value are adjusted in the opposite direction to reduce the error along the gradient direction. After repeated transmission, reduce the error to expectations and get the optimal solution Y.

Long short-term memory network (LSTM)

LSTM is an improved recurrent neural network (RNN). LSTM uses gates to control the memory process, making up for the loss caused by RNN gradient explosion and gradient disappearance to a large extent, and solving the problem that RNN cannot handle long-distance dependence. LSTM adds input gate, forgetting gate, output gate, and internal memory unit on the basis of RNN (Sepp & Jürgen 1997). The unit structure diagram of the LSTM model is shown in Figure 2.

The calculation formulas of forgetting door f_t, input door i_t and output door o_t are as follows:

$$f_t = \sigma(W_f x_t + U_i h_{t-1} + b_f) \tag{8}$$

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_{i-1}) \tag{9}$$

$$\tilde{C}_t = \tanh(W_C \times [h_{t-1}, x_t] + b_C) \tag{10}$$

$$C_t = f_t \times C_{t-1} + i_t \times \tilde{C}_t \tag{11}$$

$$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o) \tag{12}$$

$$h_t = o_t \times \tanh(C_t) \tag{13}$$

where x_t is the input layer; h_t is the hidden layer; W and U are weight parameters; b is the bias; C_t is the cell state. The input gate i_t is obtained through the activation function after the transformation between input layer x_t and output h_{t-1} of the previous hidden layer; the values of weight parameters W and U, bias b depend on the training results of the model; the result of the input gate is a vector, which is responsible for calculating the current input state based on the last output and this input; the calculation process of forgetting gate f_t and output gate o_t is similar to that of input gate i_t. Such a control process enables the LSTM model to quickly and accurately learn the long-term dependence between sequences, which has great advantages in the processing of time series runoff data.

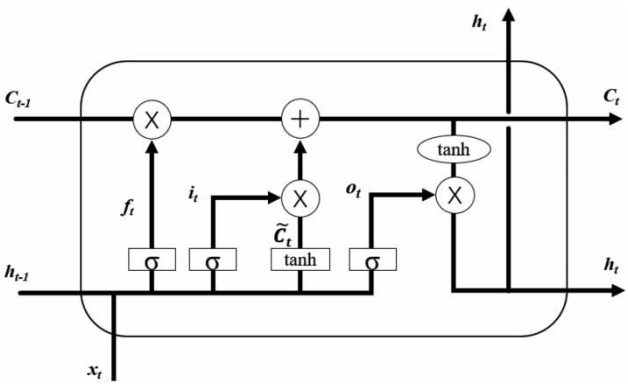


Figure 2 | The unit structure diagram of long short-term memory neural network (LSTM).

Particle swarm optimization (PSO)

Eberhart and Kennedy (2002) proposed the PSO algorithm, which can simulate the foraging behavior of birds, to realize intelligent problem solving. PSO regards the optimal solution of the problem as a particle, and obtains a set of random solutions after initialization, then iterates and finds the optimal solution by updating the velocity and position of the particle. The updated equations for the particle are as follows:

$$V_{i,j}^{t+1} = \omega V_{i,j}^{t} + c_1 r_{1,i,j}^{t}(\hat{y}_i^t - x_{i,j}^t) + c_2 r_{2,i,j}^{t}(\hat{y}_{i,j}^t - x_{i,j}^t) \tag{14}$$

$$x_{i,j}^{t+1} = x_{i,j}^{t} + V_{i,j}^{t+1} \tag{15}$$

where $V_{i,j}^t$ is the velocity of particle i in the j dimension at the t iteration; $V_{i,j}^t$ is the position of i; ω is the weight of inertia; c_1 and c_2 are learning factors; $y_{i,j}^t$ is the individual extremum point at the t iteration of the particle swarm; $\hat{y}_i^t$ is the global extremum; $r_{1,i,j}^t$ and $r_{2,i,j}^t$ are random numbers between 0 and 1.

The optimization process of PSO is as follows:

1. Initialize the parameters of the particle swarm, including population size, number of iterations, learning factor, and range of speed and location.
2. Choose the fitness function of PSO.
3. Create a particle randomly, including the various parameters of the algorithm.
4. Get the local and global optimal positions of the particle according to the fitness function.

5. Update the particle speed and position according to Equations (14) and (15). Update the local and global optimal values according to the new fitness function.
6. After reaching the maximum number of iterations, the optimal particle is taken as the parameter of the algorithm.

Weighted integrated model

The integrated model can efficiently utilize the effective information of each single prediction model (Jing & Zhang 2012). The basic idea of the weighted integrated model used in this study is to calculate the weights of frequency terms decomposed by MCEEMD from different models based on the predicted values of frequency terms, and get the weighted value as the final value of each frequency term according to the weight, and the weighted values of each frequency are added to obtain the final runoff prediction results. The calculation equations are as follows:

$$\overline{R} = \sum_{i=1}^{n} \alpha_i H_i + \sum_{i=1}^{n} \beta_i M_i + \sum_{i=1}^{n} \gamma_i L_i + \sum_{i=1}^{n} \delta_i Res_i \tag{16}$$

$$\alpha_i = \frac{\frac{1}{|e_{H,i}|}}{\sum_{i=1}^{n} \frac{1}{|e_{H,i}|}},\ \beta_i = \frac{\frac{1}{|e_{M,i}|}}{\sum_{i=1}^{n} \frac{1}{|e_{M,i}|}},\ \gamma_i = \frac{\frac{1}{|e_{L,i}|}}{\sum_{i=1}^{n} \frac{1}{|e_{L,i}|}},\ \delta_i = \frac{\frac{1}{|e_{R,i}|}}{\sum_{i=1}^{n} \frac{1}{|e_{R,i}|}} \tag{17}$$

where $\overline{R}$ is the the weighted integrated model obtained for the runoff; H_i is the result for the high frequency terms of each model; M_i is the result for the medium frequency terms of each model; L_i is the result for the low frequency terms of each model; Res_i is the result for the residual terms of each model; α_i, β_i, γ_i, δ_i are the weights of different frequency terms for each single model; n is the number of the single model; $e_{H,i}$, $e_{M,i}$, $e_{L,i}$, and $e_{R,i}$ are the errors between the predicted and true values of the frequency terms for different single models. The average relative error of the predicted values was used as the error in this study.

In order to improve the accuracy of runoff prediction, PSO optimize SVR based on MCEEMD (MCEEMD-PSO-SVR), PSO optimize BP based on MCEEMD (MCEEMD-PSO-BP), and PSO optimize LSTM based on MCEEMD (MCEEMD-PSO-LSTM) were used to develop a weighted integrated model in this study. The weighted integrated model is a process of 'Decomposition – Reconstruction – Optimization – Prediction – Weighted Integration – Reconstruction'. Figure 3 shows the framework of the model, including:

1. Decompose the original runoff data R into three frequency terms (HF, MF, LF) and one residual item Res by using MCEEMD.
2. Use each frequency term as the training data of each single model.
3. Optimize the different single models by using PSO.
4. Calculate the weights (α_i, β_i, γ_i, δ_i) of each frequency term from different single models by using the average relative error.

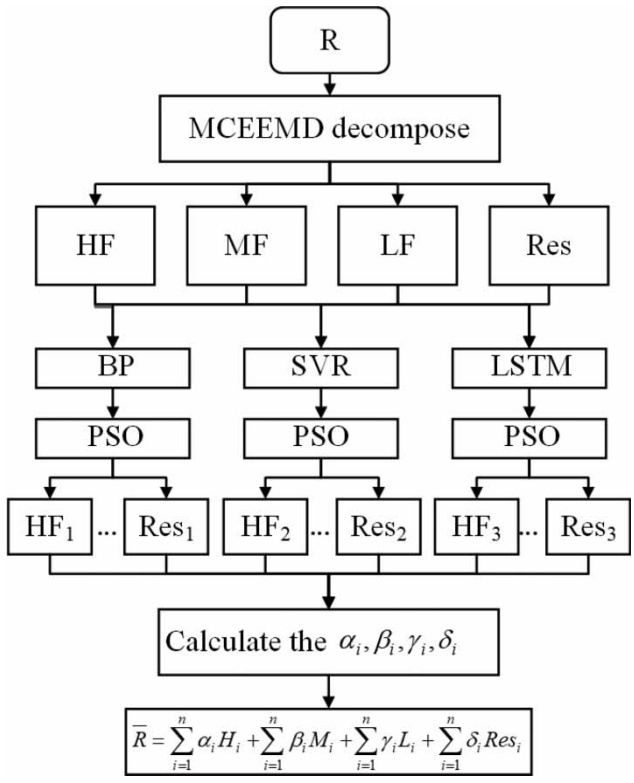


Figure 3 | The framework of the weighted integrated model.

5. Develop the weighted integrated model based on the weights and reconstruct the predicted results of the frequency terms to obtain the final prediction result $\overline{R}$.

Model evaluation

The Nash–Sutcliffe efficiency (NSE) and the qualified rate (QR) were used to evaluate the performance of the weighted integrated model (Nash & Sutcliffe 1970). According to the Standard for hydrological information and hydrological forecasting (2008), a qualified sample is one in which the error between the predicted and observed value does not exceed 20%, and a QR greater than 70% is considered to be a reliable model. Maximum absolute error (MaxAE), minimum absolute error (MinAE), mean absolute error (MAE), mean absolute percentage error (MAPE), and root mean square error (RMSE) were selected as the comparison indicators of the integrated model and other single prediction models (Willmott & Matsuura 2005; Myttenaere *et al.* 2016). The calculation formulas are as follows:

$$NSE = 1 - \frac{\sum_{i=1}^{n} [y_c(i) - y_0(i)]^2}{\sum_{i=1}^{n} [y_0(i) - \bar{y}_0]^2} \tag{18}$$

$$QR = \frac{m}{n} \times 100\% \tag{19}$$

$$MaxAE = MAX(|y_c(i) - y_o(i)|) \tag{20}$$

$$MinAE = MIN(|y_c(i) - y_o(i)|) \tag{21}$$

$$MAE = \frac{1}{m} \sum_{i=1}^{n} |y_c(i) - y_o(i)| \tag{22}$$

$$MAPE = \sum_{i=1}^{n} \left| \frac{y_o(i) - y_c(i)}{y_o(i)} \right| \times \frac{100\%}{n} \tag{23}$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} [y_o(i) - y_c(i)]^2} \tag{24}$$

where $y_c(i)$ is the predicted value; $y_0(i)$ is the measured value; $\bar{y}_0$ is the mean value of the measured value; n is the total number of predicted samples; and m is the total number of qualified predicted samples.

STUDY AREA AND DATA

Lijin hydrological station is located in Dongying, Shandong province, 104 km from the mouth of the Yellow River, which is located in the lowest reaches of the Yellow River. Gaocun hydrological station is an important control station for the Yellow River flowing into Shandong province, with a section 579.1 km from the estuary and a catchment area of 734,146 km^2. Aishan hydrological station is located in Liaocheng, Shandong province. The above three hydrological stations are responsible for providing water conditions for flood control and water resources dispatching in the lower reaches of the Yellow River, studying and exploring the changing rules of hydrological factors, and collecting hydrological data for river regulation and water and sand resource utilization in the lower reaches of the Yellow River. The data from these stations are complete and consistent with the law of hydrology. Figure 4 shows the location of these hydrological stations.

This study used monthly runoff data for a total of 780 months from January 1950 to December 2014. The training set used runoff data from January 1950 to December 2001, a total of 624 months. The validation set runoff data from January 2002 to December 2014, a total of 156 months, was used to verify results of the model. Table 1 shows the characteristics of monthly runoff data of the three hydrological stations. Coefficient of variation (CV) is a relative index to measure the dispersion degree of runoff data, skewness represents the degree of asymmetry in the distribution of runoff data, and kurtosis represents the peak shape characteristic of probability density distribution curve. Table 1 shows the following:

1. The CVs of each hydrological station in the lower reaches of the Yellow River are 0.86–1.07, which indicates that the change of monthly runoff is dramatic and the distribution is uneven in the year.
2. All skewness are greater than 0, and all are positive deviations, indicating that a small number of monthly runoff data are large.
3. All kurtosis are greater than the kurtosis of normal distribution and uniform distribution, indicating that the monthly runoff data differ greatly from the mean value and have many extreme values.

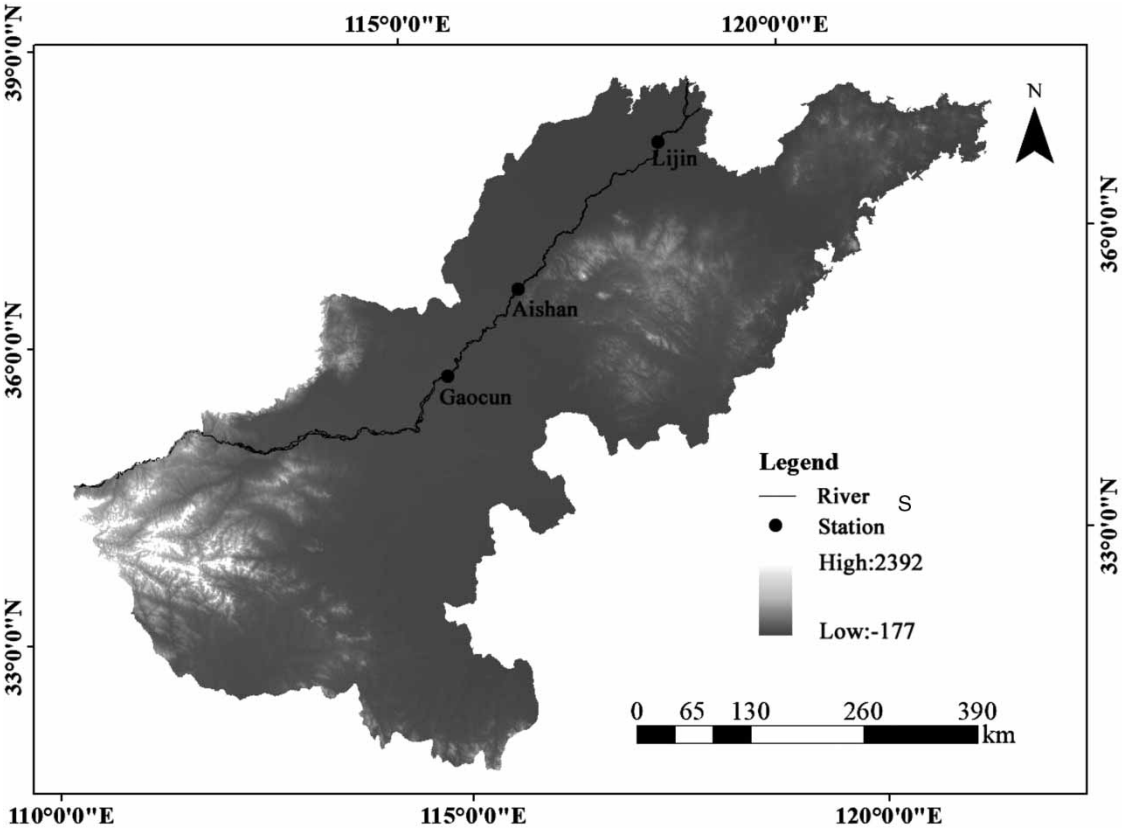


Figure 4 | The location of Lijin hydrological station, Gaocun hydrological station, and Aishan hydrological station.

Table 1 | Characteristics of monthly runoff data of three hydrological stations

Station	Average monthly runoff	CV	Skewness	Kurtosis
Lijin	25.65 (10^8 m^3)	1.07	2.05	5.11
Gaocun	29.01 (10^8 m^3)	0.86	2.10	5.02
Aishan	27.93 (10^8 m^3)	0.96	2.14	5.51

The variation process of monthly runoff is shown in Figure 5. It can be seen from Table 1 and Figure 5 that the monthly runoff had a large amplitude of variation and a weak periodicity, presenting a nonlinear and non-stationary characteristic state.

RESULTS

The original series decomposed by MCEEMD

MCEEMD was used to decompose the monthly runoff data from January 1950 to December 2014. The decomposed results included three frequency terms and one residual term. The components were reduced to four after the reconstruction, and the characteristics and objective laws of the original time series data were still retained. In addition, the high frequency term obtained by combining IMF1 and IMF2 eliminated the nonlinear characteristics of IMF1. The monthly runoff data after decomposition by MCEEMD are shown in Figure 6. As can be seen from Figure 6, although the reconstructed high frequency term retains some features of IMF1 and IMF2, it has reduced the nonlinear feature, and the features of the intermediate frequency term and low frequency term are similar to those of the IMF. It can be seen that HF has the highest frequency, the largest fluctuation, and the shortest wavelength. The frequency gradually decreases in MF and LF, the fluctuations gradually weaken, the wavelength gradually increases and the periodicity becomes stronger. The residual term represents the trend of monthly runoff at stations over time from 1950 to 2014. As can be seen in Figure 6, the monthly

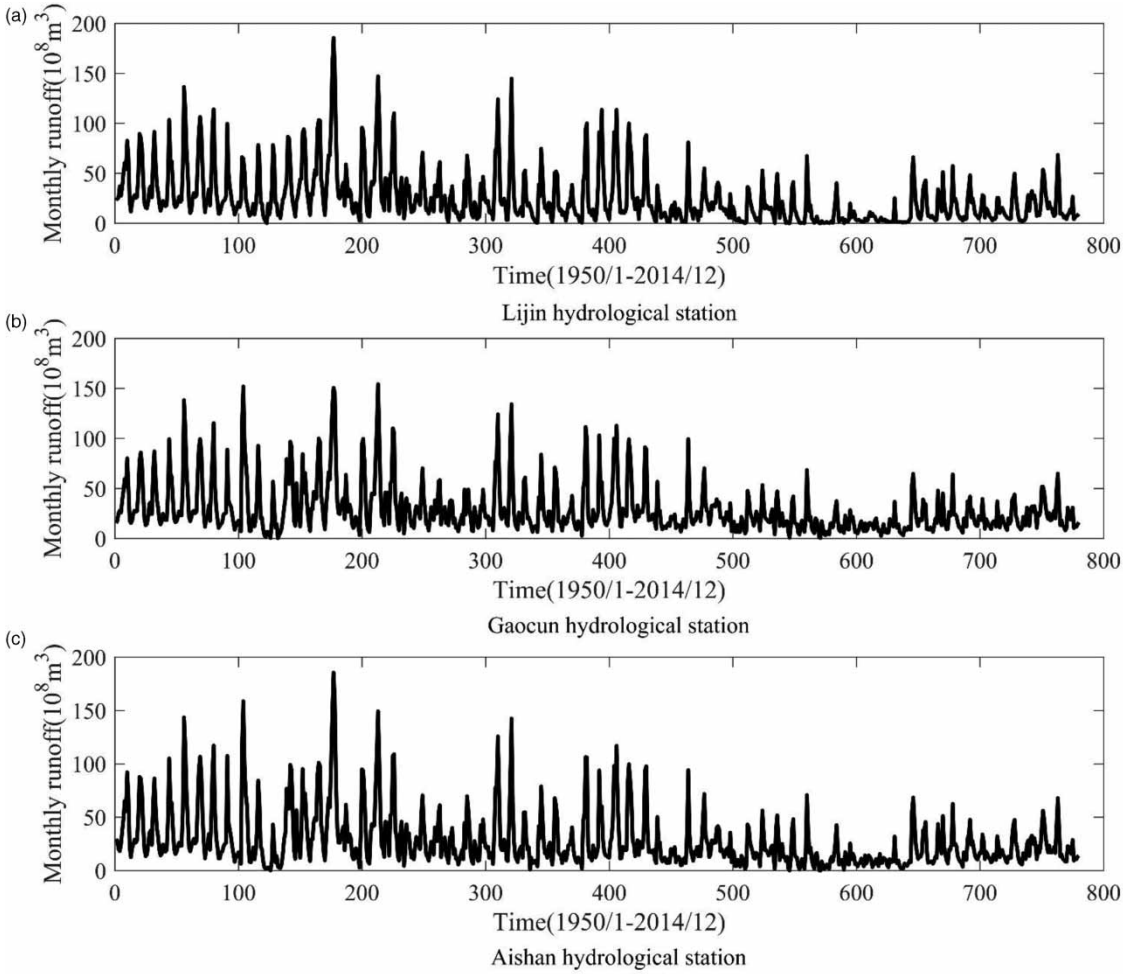


Figure 5 | The observed monthly runoff from January 1950 to December 2014 in (a) Lijin hydrological station, (b) Gaocun hydrological station, (c) Aishan hydrological station.

runoff shows a declining trend year to year, with a low decline rate during the early months and a higher decline rate in the later months.

Data processing

Since the monthly runoff time series data are unstable and nonlinear, when these data are directly input to the algorithm, the model training process will generate large numerical fluctuations. Therefore, the monthly runoff data need to be normalized before the training process. The normalized formula is as follows:

$$y_i' = \frac{y_i - \min(y_i)}{\max(y_i) - \min(y_i)} \tag{25}$$

After prediction, the data can be re-scaled following the contrary procedure of Equation (25).

Parameter selection by PSO

Proper parameter selection can directly determine the performance of the training process. The parameters of the predictive algorithm are usually selected manually. However, manual selection is too subjective. In addition, the numerous parameters of the algorithm can make a lot of unnecessary work. Therefore, algorithms in this study were optimized using PSO with global optimization capabilities.

After optimization, the radial basis function (RBF) kernel function of SVR was selected as the kernel function, the optimal penalty factor C of SVR was 8.3598, and the

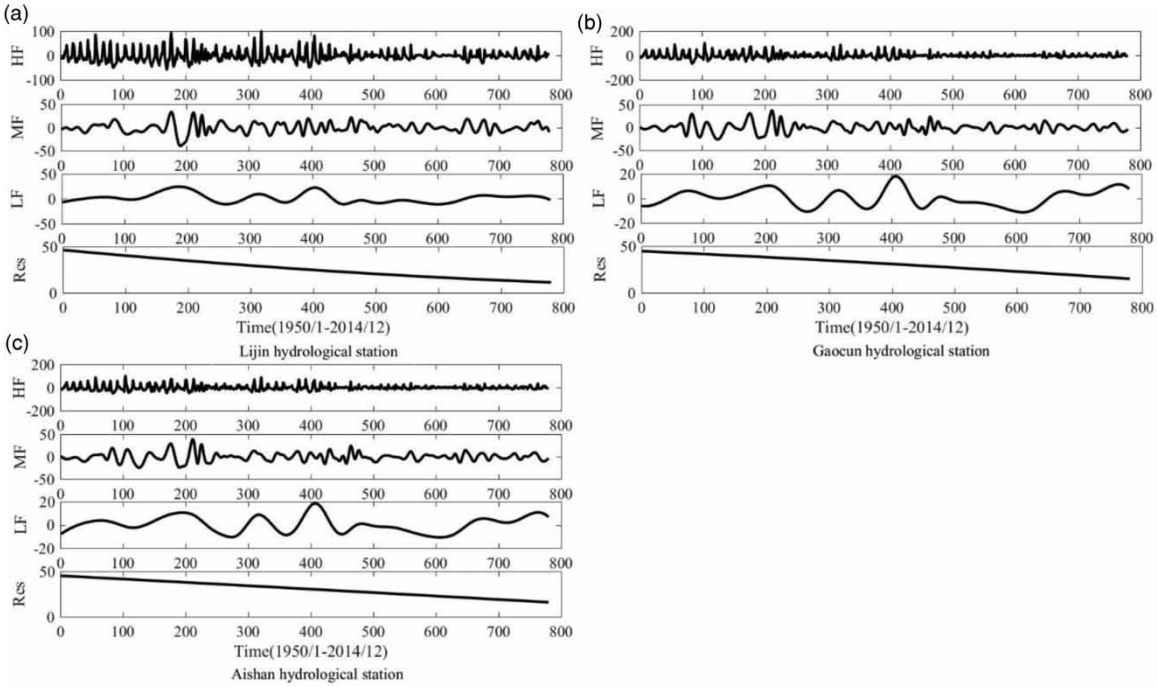


Figure 6 | The frequency terms of monthly runoff after MCEEMD decomposition from January 1950 to December 2014 in (a) Lijin hydrological station, (b) Gaocun hydrological station, (c) Aishan hydrological station.

parameter g of the optimal kernel function was 0.031413 for SVR. The input layer of BP was set to 7 and the output layer to 1. The activation function of LSTM was selected as rectified linear unit (ReLU), the units of the LSTM were selected as 20, the loss function of LSTM was selected as MAE, and the optimizer of LSTM was selected as adaptive moment estimation (Adam).

The activation function ReLU of LSTM is a non-saturated activation function. Compared with other activation functions, ReLU has the advantages of fast convergence and fast calculation speed (Ramachandran *et al.* 2017). The optimizer Adam, which is computationally efficient and has little memory requirements, combines first moment estimation (the mean of the gradients) and second moment estimation (the uncentralized variance of the gradients) (Kingma & Ba 2014).

Predicted results

According to Equation (17), the weights of each frequency term for different single models in the weighted integrated model were obtained by calculating average relative errors of the predicted results for the same frequency term of different models, and the results are shown in Table 2. The greater the weight, the greater the contribution of the single model to the integrated model.

The predicted values of the weighted integrated model were calculated according to the weights obtained in Table 2. Table 3 indicates the prediction performance of

Table 2 | The weight coefficients of frequency terms for each single model

Stations	Terms	MCEEMD-PSO-SVR	MCEEMD-PSO-BP	MCEEMD-PSO-LSTM
Lijin	HF	0.32	0.26	0.43
	MF	0.36	0.30	0.33
	LF	0.31	0.32	0.37
	Res	0.29	0.34	0.37
Gaocun	HF	0.36	0.24	0.40
	MF	0.40	0.22	0.38
	LF	0.34	0.28	0.37
	Res	0.30	0.35	0.35
Aishan	HF	0.23	0.22	0.55
	MF	0.35	0.29	0.36
	LF	0.29	0.35	0.36
	Res	0.33	0.36	0.31

Table 3 | The performance of the weighted integrated model on each hydrological station

Indicators	Training set		Verification set	
Stations	NSE	QR/%	NSE	QR/%
Lijin	0.98	87	0.93	78
Gaocun	0.99	86	0.95	79
Aishan	0.98	85	0.92	75

the weighted integrated model on each station in training set and verification set. It can be seen that the NSEs of each station during the training period and the verification period were all greater than 0.5, indicating that the predicted results were reliable. NSE was greater than 0.98 and QR was greater than 85% in each station during the training period. NSE was greater than 0.92 and QR was greater than 75% in each station during the verification period. The predicted results accord with the standard of hydrological forecast.

Since the data in this study were divided into training set and verification set, based on the following considerations, we did not pay much attention to the prediction results of the training set:

1. when establishing the prediction model, the performance of the verification set can represent the real application effect;
2. for the adaptive prediction model, we do not care about the prediction performance of the training set;
3. the output of the training set is not a real prediction, it is only an input to the model of the verification set.

Therefore, in the following sections, we just show the prediction results of the verification set. The errors of the weighted integrated model were compared with those of the MCEEMD-PSO-SVR model, the MCEEMD-PSO-BP model, and the MCEEMD-PSO-LSTM model, and the results are shown in Table 4. It can be obviously seen from Table 4 that the weighted integrated model has the best performance and it outperforms the other models in terms of all the error coefficients, acquiring the best MAE, MAPE, and RMSE. The MAE, MAPE, and RMSE of the integrated model in Lijin, Gaocun, and Aishan hydrological stations were reduced by 24.36%–30.04%, 18.54%–24.82%, 8.63%–16.37%; 29.31%–34.66%, 10.91%–35.21%, 15.73%–26.52%; and 16.17%–24.23%, 12.44%–13.55%, 13.08%–16.22%, respectively, compared with the other three models. In addition, the values of MaxAE and MinAE for the integrated model and the difference between them were the smallest of all models.

Figure 7 illustrates the runoff prediction of Lijin, Gaocun, and Aishan stations by the MCEEMD-PSO-SVR,

Table 4 | Comparison of error indicators of the single prediction models and the weighted integrated model in three hydrological stations

Stations	Indicators	MCEEMD-PSO-SVR	MCEEMD-PSO-BP	MCEEMD-PSO-LSTM	Integrated model
	MaxAE (10^8 m^3)	12.43	13.06	18.17	10.71
	MinAE (10^8 m^3)	0.05	0.12	0.09	0.01
Lijin	MAE (10^8 m^3)	2.41	2.53	2.34	1.77
	MAPE (%)	30.86	29.43	28.48	23.20
	RMSE (10^8 m^3)	3.37	3.42	3.13	2.86
	MaxAE (10^8 m^3)	13.93	17.89	9.28	7.82
	MinAE (10^8 m^3)	0.08	0.09	0.07	0.01
Gaocun	MAE (10^8 m^3)	2.51	2.42	2.32	1.64
	MAPE (%)	15.58	18.66	13.57	12.09
	RMSE (10^8 m^3)	3.21	3.28	2.86	2.41
	MaxAE (10^8 m^3)	14.98	23.00	11.05	8.44
	MinAE (10^8 m^3)	0.03	0.05	0.02	0.01
Aishan	MAE (10^8 m^3)	2.35	2.49	2.60	1.97
	MAPE (%)	16.42	16.32	16.53	14.29
	RMSE (10^8 m^3)	3.21	3.33	3.22	2.79

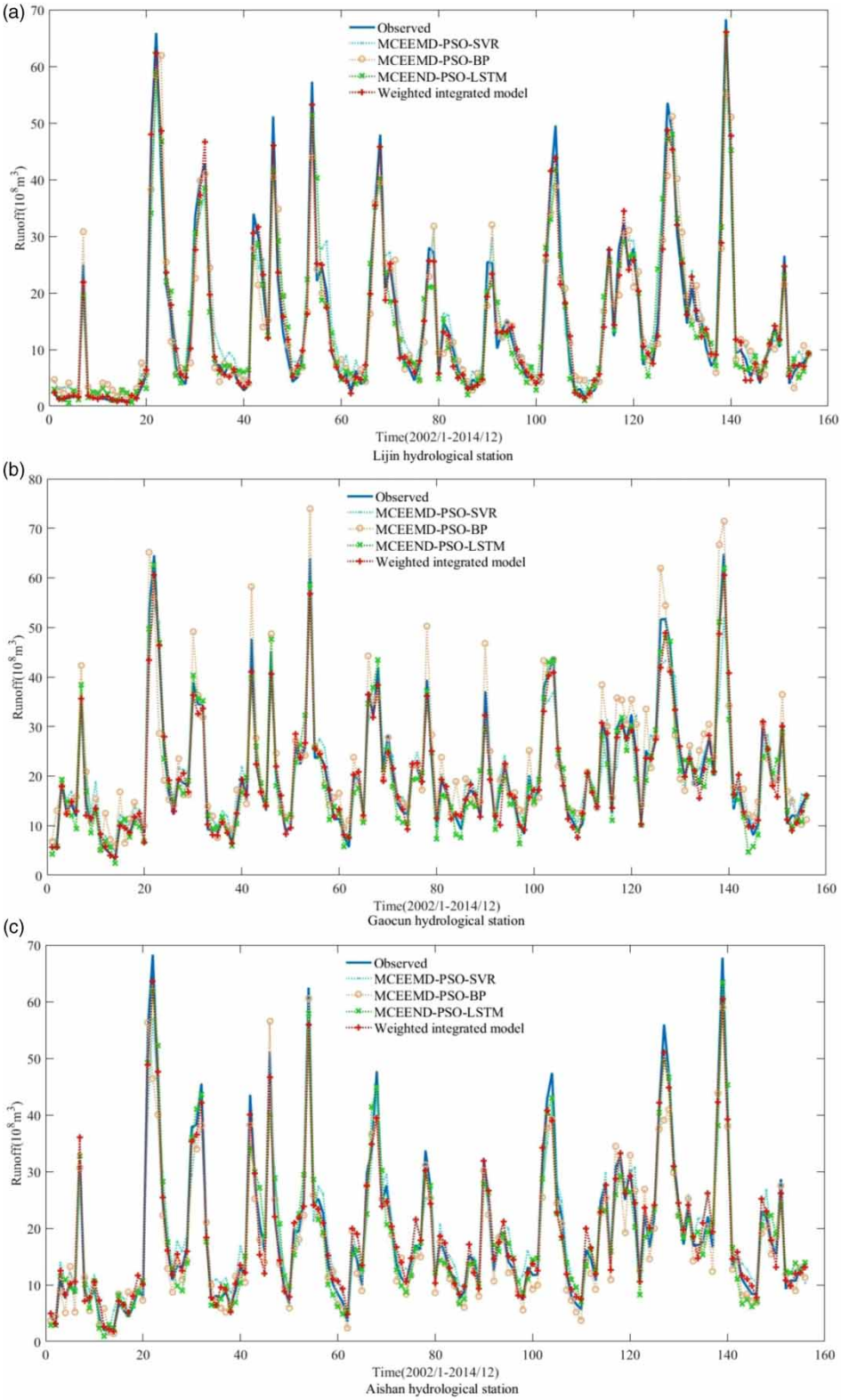


Figure 7 | Predicted and observed monthly runoff during verification period by single prediction models and the weighted integrated model in (a) Lijin hydrological station, (b) Gaocun hydrological station, (c) Aishan hydrological station.

MCEEMD-PSO-BP, MCEEMD-PSO-LSTM and the weighted integrated model in the verification set. Figure 8 shows the scatter plots of predicted value and observed value of the three stations using all models, respectively.

Some studies have found that atmospheric circulation anomaly factors have a strong correlation with extreme weather and climate anomalies (Wang 2009; Huang *et al.* 2015a, 2015b, 2017; Liu *et al.* 2017; Yang *et al.* 2016). Unfortunately, the present study mainly focused on the atmospheric circulation anomaly factors to the overall effect of the runoff, without considering the decomposition technique abnormal atmospheric circulation factors for size (Meng *et al.* 2017). Therefore, monthly scale data of Nino3.4, Arctic Oscillation (AO), Pacific Decadal Oscillation (PDO) and Atlantic Multidecadal Oscillation (AMO) were used as additions to each frequency term of the input data of the weighted integrated model to improve the accuracy and stability of runoff forecasting during some periods. In order to select optimal factors, the correlation coefficient method was used to analyze the runoff frequency terms decomposed by MCEEMD and the atmospheric circulation anomaly factors with different time delays. Results are shown in Figure 9, where the darker the shade represents the higher the correlation. The factors whose correlation coefficient is greater than 0.3 were selected as the additional term of input data. It can be seen from Figure 9(a) that in Lijin station, Nino3.4 has significant influence on the high frequency term and the residual term, while AMO has significant influence on the residual term. It can be seen from Figure 9(b) that in Gaocun station, Nino3.4 has significant influence on the high frequency term and medium frequency term, while PDO has significant influence on the medium frequency term and lower frequency term. It can be seen from Figure 9(c) that in Aishan station, Nino3.4 has significant influence on the high frequency term, lower frequency term, and residual term.

Atmospheric circulation anomaly factors that have significant influence on each frequency term of runoff in the above analysis were selected and used as additional eigenvectors to the weighted integrated model input data. In the weighted integrated model, the frequency terms and the atmospheric circulation anomaly factors were used as its input data, and the model first obtained the prediction of each frequency term and then reconstructed each frequency term according to the weights to obtain the final runoff prediction.

It can be seen from Table 5 that the weighted integrated model combined with the information of atmospheric circulation anomaly factors improved the runoff prediction accuracy of each station in different degrees. The NSE of each station increased by 1.05%–3.26%, the QR increased by −2.67% to 8.97%, the RMSE decreased by 2.10%–3.94%, and the difference between MaxAE and MinAE decreased by 8.78%–24.11%, among which the QR of Aishan station decreased by 2.67%, which may be due to the weak influence of the atmospheric circulation anomaly factor on Aishan station.

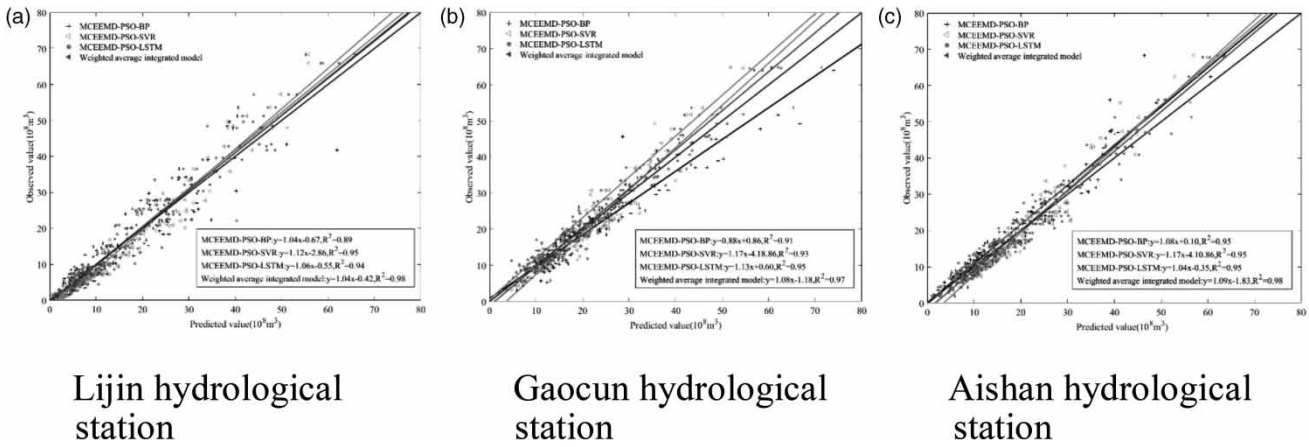


Figure 8 | The relationship of predicted and observed monthly runoff by single prediction models and the weighted integrated model in (a) Lijin hydrological station, (b) Gaocun hydrological station, (c) Aishan hydrological station.

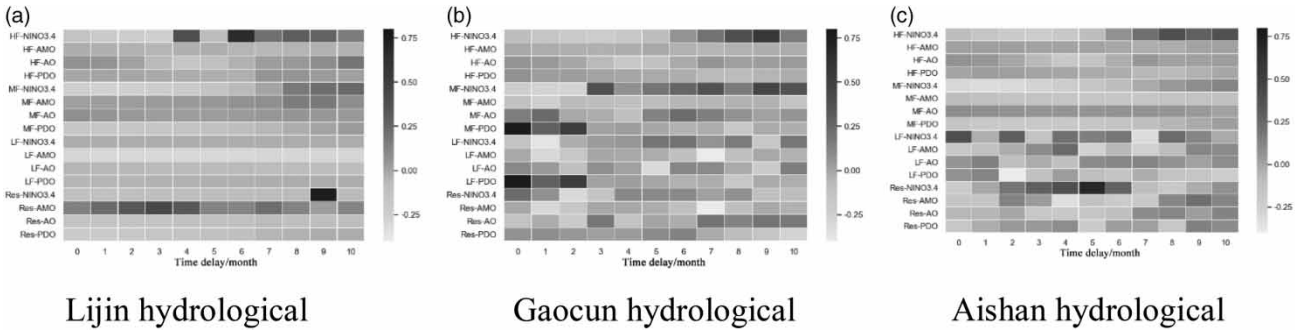


Figure 9 | Thermal diagram of the correlation coefficient between the frequency term of each station and the atmospheric circulation anomaly factor. (a) Lijin hydrological station, (b) Gaocun hydrological station, (c) Aishan hydrological station.

Table 5 | Predictive changes of atmospheric circulation anomaly factors before and after fusion

Stations	NSE	QR (%)	RMSE (10^8 m^3)	Difference (10^8 m^3)
Lijin	0.93 → 0.95	78 → 85	2.86 → 2.80	10.70 → 8.12
Gaocun	0.95 → 0.96	79 → 82	2.41 → 2.34	7.81 → 6.26
Aishan	0.92 → 0.95	75 → 73	2.79 → 2.68	8.43 → 7.69

DISCUSSION

According to the residual term decomposed by MCEEMD from the monthly runoff data of Lijin, Gaocun, and Aishan hydrological stations, the overall runoff monitored by stations showed a declining trend, especially in recent years. It indicates that the lower reaches of the Yellow River water shortage and other problems are increasingly serious, one important reason it is necessary to improve the accuracy of runoff prediction to promote the optimal allocation of water resources and maximize the benefit of water resources.

It appears from Table 4 that the MCEEMD-PSO-LSTM model is second only to the weighted integrated model based on MCEEMD in performance. In addition, it can be found from Table 2 that the MCEEMD-PSO-LSTM model has the largest weight in the weighted integrated model, thus it can be explained that the weight determines the importance of a single model in the weighted integrated model and also indicates the performance of a single model. It can be seen from Table 4 that, although the models' errors are different, the differences among MCEEMD-PSO-SVR, MCEEMD-PSO-BP, and MCEEMD-PSO-LSTM are not too large, which indicates that the three models are similar in the performance of runoff prediction. The integrated model requires that the models that comprise it be close in performance, and the discussion above reveals that the three single models that comprise the integrated model meet this requirement. Moreover, the use of the weight coefficient further reduces the difference between single models, so that the single model with good performance takes up a larger proportion. In this way, the advantages of the model are magnified and the disadvantages of the model are reduced, which is conducive to improving the accuracy of runoff prediction.

It can be seen in Figure 7(a) that the cycle and the trend of the prediction results of the weighted integrated model were completely consistent with the original monthly runoff data in Lijin hydrological station. Moreover, the weighted integrated model had the best predictive performance among all models. Similarly, it can be observed from Figure 7(b) and 7(c) that the weighted integrated model had the better performance compared with the other models. In flood season, the weighted integrated model had higher prediction accuracy than other models.

Figure 8 indicates that the weighted integrated model had the best performance for forecasting the monthly runoff, as the linear trend line of it was closest to the $y = x$ that goes through the point (0,0) compared with the MCEEMD-PSO-SVR, the MCEEMD-PSO-BP, and the MCEEMD-PSO-LSTM. Similarly, the above conclusions can also be observed from Figure 8(b) and 8(c). R^2 represents the explanatory power of the equation variable x to y, and the closer R^2 to 1, the better the model fits the data. It can be seen that the R^2 of the weighted integrated

model was the largest and closest to 1, which proves that the prediction performance of it was the best of all models.

It can be observed in scatter plots that the peak values of the weighted integrated model were mostly above the line of $y = x$, indicating that the forecasted peaks are lower than the observed ones. This phenomenon may result from the limitation of our model. However, the values, except the peak value, in our model were evenly distributed around the line $y = x$ compared with other models, indicating that its fitting result had a better stability. In addition, the minimum difference between MaxAE and MinAE also indicated that.

From Figure 9 it appears that the influences of abnormal atmospheric circulation factors on the runoff of each station are different. This is because in this paper, the locations of the three hydrological stations are from west to east, and the influence of different locations by the Asian summer monsoon are different, so the correlation between the runoff of hydrological stations and abnormal atmospheric circulation factors is different.

Atmospheric circulation anomaly factors are greatly affected by extreme weather and climate anomalies, and the violent change of runoff in flood season is related to extreme weather. Therefore, the integrated model has a stronger ability to simulate the flood season runoff and its extreme value after adding atmospheric circulation anomaly factors, which improves the stability of the model and the prediction accuracy.

For complicated and non-stationary monthly runoff, the single prediction model has low accuracy and poor prediction effect and is not suitable for medium- and long-term prediction. CEEMD reduces the mutual interference between different trend information, retains the objective law of the original data, and provides a suitable data basis for the prediction algorithm. However, if modeling and forecasting are carried out for each of the eight components obtained after the decomposition of CEEMD, a large amount of work and errors will be incurred. Therefore, this study proposed MCEEMD, which reconstructed the IMFs with similar fluctuation frequency and amplitude to obtain the frequency terms, which is then used as the input of the prediction algorithm. The weighted integrated model combines the advantages of each algorithm and determines the importance of each algorithm in the integration model according to the weight coefficient. By combining the MCEEMD with the weighted integrated model, runoff can be reduced into a stationary sequence to increase the performance of prediction model, and the defects of the single prediction algorithm were also reduced. At the same time, the addition of abnormal atmospheric circulation factors improves the stability and prediction accuracy of the model under extreme weather conditions.

CONCLUSIONS

In order to improve the accuracy of runoff prediction, a weighted integrated model based on MCEEMD was proposed, which was verified by using the monthly runoff data from three hydrological stations (Lijin, Gaocun, Aishan). The MCEEMD reduces the calculation problem caused by excessive components by establishing frequency terms, and also solves the problem that prediction algorithm has poor prediction effect on IMF1. The weighted integrated model uses MCEEMD to process training data, combines PSO-SVR, PSO-BP, and PSO-LSTM, and integrates the advantages of each single model, thus improving the prediction accuracy, stability, and fitting effect of the single model. Compared with single prediction models, the weighted integrated model has a better performance in runoff prediction and higher accuracy of prediction results. The prediction results of the weighted integrated model are in line with the hydrological prediction standard and have a high reliability. Compared with the single prediction algorithms, the MAE, MAPE, and RMSE of the weighted integrated model were reduced by more than 15.81%, 10.91%, and 13.42%, respectively. Therefore, the combination of the MCEEMD method and weighted integrated model is viable in the medium- and long-term runoff prediction. In addition, it can also provide guidance for flood control and drought relief.

Considering that runoff is affected by climate and extreme weather, atmospheric circulation anomaly factors with strong correlation with runoff were added and selected as additional items of input data of the integrated model in this study. The results show that the model combined with atmospheric circulation anomaly factors has higher

prediction accuracy and stability, especially in extreme weather and flood seasons.

ACKNOWLEDGEMENTS

Research of this article is granted by NSFC under grant no. U1604152 and the key research project of Ministry of Water Resources of the People's Republic of China. The authors of this paper would like to express their thanks to CMDC for providing the data. The authors declare no conflict of interest. (http://61.163.88.227:8006/hwsq.aspx, http://data.cma.cn/, https://www.ncdc.noaa.gov/).

DATA AVAILABILITY STATEMENT

All relevant data are available from an online repository or repositories. http://61.163.88.227:8006/hwsq.aspx, http://data.cma.cn/, https://www.ncdc.noaa.gov/.

REFERENCES

Chu, H., Wei, J., Li, T. & Jia, K. 2016 Application of support vector regression for mid- and long-term runoff forecasting in 'Yellow River Headwater' region. *Procedia Engineering* **154**, 1251–1257. https://doi.org/10.1016/j.proeng.2016.07.452

Cui, D. 2013a An improved Elman neural network and its application to runoff forecast. *Hydro-Science and Engineering* **2**, 71–77. https://doi.org/10.16198/j.cnki.1009-640x.2013.02.003

Cui, D. 2013b Application of hidden multilayer BP neural network model in runoff prediction. *Hydrology* **33**(1), 68–73.

Eberhart, R. & Kennedy, J. 2002 A new optimizer using particle swarm theory. *Proceedings of the Sixth International Symposium on Micro Machine and Human Science*, IEEE, Nagoya, Japan. https://doi.org/10.1109/MHS.1995.494215

Han, R., Dong, Z. C., Wang, X. W. & Ma, H. L. 2017 Application of weighted average integrated model in runoff forecasting. *Yellow River* **39**(6), 16–20, 23. https://doi.org/10.3969/j.issn.1000-1379.2017.06.004

Huang, N. E. & Wu, Z. H. 2008 A review on Hilbert-Huang transform: method and its applications to geophysical studies. *Reviews of Geophysics* **46**(2). https://doi.org/10.1029/2007RG000228

Huang, N. E., Shen, Z., Long, S. R., Wu, M. C., Shih, H. H., Zheng, Q. A., Yen, N. C., Tung, C. C. & Liu, H. H. 1998 The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. *Philosophical Transactions – Royal Society. Mathematical, Physical and Engineering Sciences* **454**, 903–995. https://doi.org/10.1098/rspa.1998.0193

Huang, S. Z., Chang, J. X., Huang, Q. & Chen, Y. T. 2015a Monthly streamflow prediction using modified EMD-based support vector machine. *Journal of Hydrology* **511**, 764–775. https://doi.org/10.1016/j.jhydrol.2014.01.062

Huang, S. Z., Huang, Q., Chang, J. X., Zhu, Y. L., Leng, G. Y. & Xing, L. 2015b Drought structure based on a nonparametric multivariate standardized drought index across the Yellow River basin, China. *Journal of Hydrology* **530**, 127–136. https://doi.org/10.1016/j.jhydrol.2015.09.042

Huang, S. Z., Li, P., Huang, Q. & Leng, G. Y. 2017 The propagation from meteorological to hydrological drought and its potential influence factors. *Journal of Hydrology* **547**, 184–195. https://doi.org/10.1016/j.jhydrol.2017.01.041

Jing, Y. P. & Zhang, X. 2012 Mid-long term river runoff forecast based on modified combination model. *Journal of Hydroelectric Engineering* **31**(6), 14–21, 31.

Kingma, D. & Ba, J. 2014 Adam: a method for stochastic optimization. *arXiv: 1412.6980*.

Liu, S., Huang, S. Z., Huang, Q., Xie, Y. Y., Leng, G. Y., Luan, J. K., Song, X. Y., Wei, X. & Li, X. Y. 2017 Identification of the non-stationarity of extreme precipitation events and correlations with large-scale ocean-atmospheric circulation patterns: a case study in the Wei River Basin, China. *Journal of Hydrology* **548**, 184–195. https://doi.org/10.1016/j.jhydrol.2017.03.012

Lu, D. & Zhou, H. C. 2014 Medium and long-term runoff forecasting based on mutual information and BP neural network. *Hydrology* **34**(4), 8–14, 67.

Meng, E. H., Huang, S. Z., Huang, Q., Liu, D. F. & Bai, T. 2017 Runoff prediction incorporating anomalous atmospheric circulation factors. *Journal of Hydroelectric Engineering* **36**(8), 34–42.

Myttenaere, A. D., Golden, B., Grand, B. L. & Rossi, F. 2016 Mean absolute percentage error for regression models. *Neurocomputing* **192**(5), 38–48. https://doi.org/10.1016/j.neucom.2015.12.114

Nash, J. E. & Sutcliffe, J. V. 1970 River flow forecasting through conceptual models part I – a discussion of principles. *Journal of Hydrology* **10**(3), 282–290. https://doi.org/10.1016/0022-1694(70)90255-6

Ouyang, R. L., Ren, L. L. & Zhou, C. H. 2010 Similarity search in hydrological time series. *Journal of Hohai University (Natural Sciences)* **38**(3), 241–245.

Ramachandran, P., Zoph, B. & Le, Q. V. 2017 *Searching for Activation Functions*. *arXiv:1710.05941*.

Sepp, H. & Jürgen, S. 1997 Long short-term memory. *Neural Computation* **9**, 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Standard for Hydrological Information and Hydrological Forecasting 2008 General Administration of Quality Supervision, Inspection and Quarantine of the People's Republic of China/ Standardization Administration, Beijing, China.

Tesfaye, Y. G., Meerschaert, M. M. & Anderson, P. L. 2006 Identification of periodic autoregressive moving average models and their application to the modeling of river flows. *Water Resources Research* **42**(1). https://doi.org/10.1029/2004WR003772

Vapnik, V. 1998 *Statistical Learning Theory*. Wiley, New York, USA.

Wang, Y. M. 2009 *Atlantic Multidecadal Oscillation (AMO) on Climate in The Asian Monsoon Region: Observations and Multi-Model Simulation. PhD Thesis*, Ocean University of China, Qingdao, China.

Willmott, C. J. & Matsuura, K. 2005 Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research* **30**(1), 79–82. https://doi.org/10.3354/cr030079

Xiang, Z., Yan, J. & Demir, I. 2020 A rainfall-runoff model with LSTM-based sequence-to-sequence learning. *Water Resources Research* **56**(1), e2019WR025326. https://doi.org/10.1029/2019WR025326

Yang, K. B., Zhu, J., Ge, Z. X. & Gong, C. H. 2016 Modeling and forecasting of flood season runoff in Datong station under the influence of ENSO event. *Hydropower Energy Science* **34**(5), 5–8. https://doi.org/CNKI:SUN:SDNY.0.2016-05-002

Yeh, J. R., Shieh, J. S. & Huang, N. E. 2010 Complementary ensemble empirical mode decomposition: a novel noise enhanced data analysis method. *Advances in Adaptive Data Analysis* **2**(2), 135–156. https://doi.org/10.1142/S1793536910000422

Zhang, P., Dai, Y. S., Zhang, H. Q., Wang, C. X. & Zhang, Y. H. 2019 Combining CEEMD and recursive least square for the extraction of time-varying seismic wavelets. *Reviews of Geophysics* **170**. https://:doi.org/10.1016/j.jappgeo.2019.103854

First received 12 July 2020; accepted in revised form 6 October 2020. Available online 18 November 2020