Water Supply

Water Supply Vol 22 No 8, 6807 doi: 10.2166/ws.2022.262

Annual runoff forecast based on a combined EEMD-ARIMA model

Xianqi Zhang[a,b,c], Yaohui Lu[a,*], Guoyu Zhu[d], Xilong Wu[a], Dong Zhao[a] and Bingsen Duan[a]

[a] Water Conservancy College, North China University of Water Resources and Electric Power, Zhengzhou 450046, China
[b] Collaborative Innovation Center of Water Resources Efficient Utilization and Protection Engineering, Zhengzhou 450046, China
[c] Technology Research Center of Water Conservancy and Marine Traffic Engineering, Zhengzhou, Henan Province 450046, China
[d] State Key Laboratory of Hydraulics and Mountain River Engineering, College of Water Resource and Hydropower, Sichuan University, Chengdu 610065, China
*Corresponding author. E-mail: 407556540@qq.com

ABSTRACT

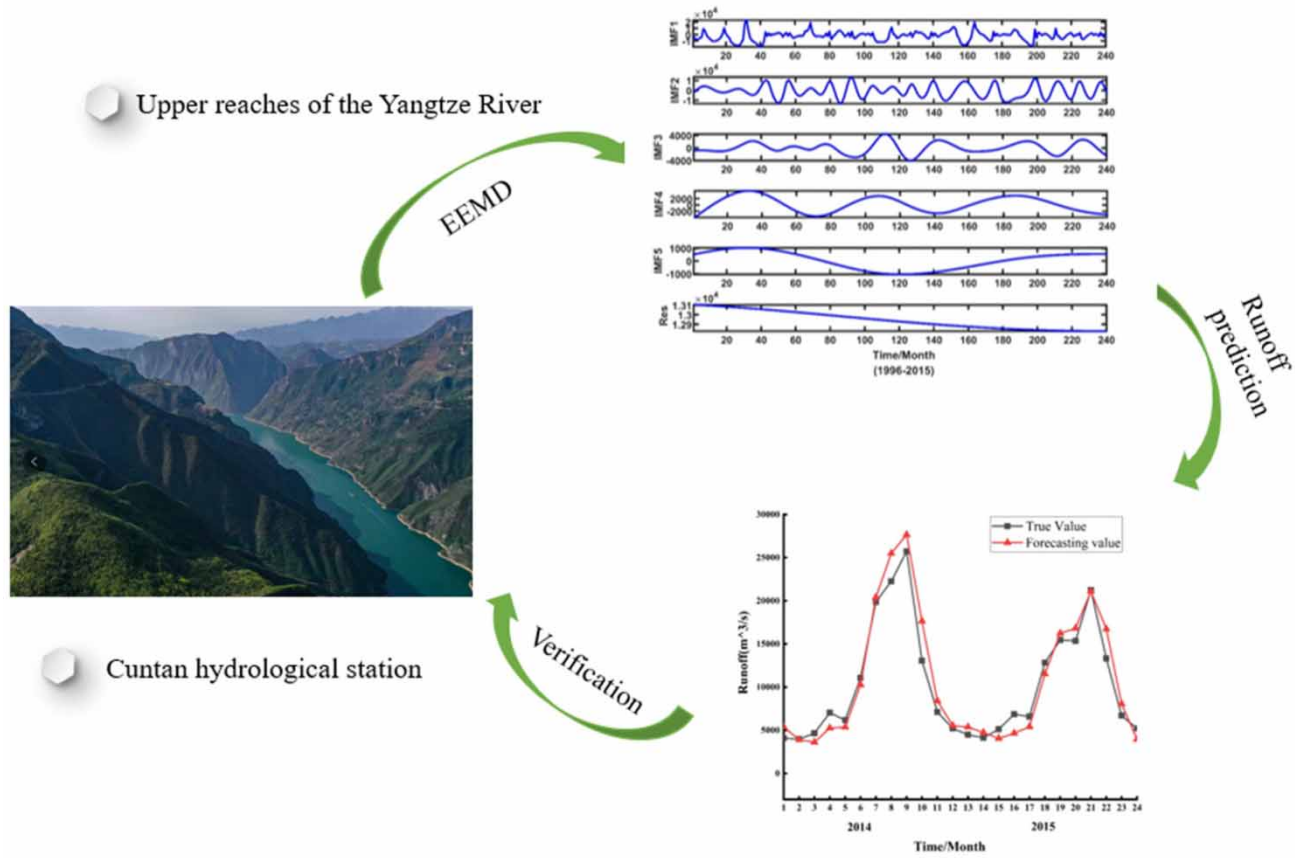
In order to improve the accuracy of hydrological models for runoff prediction and to solve the problem of disappearing time series data break or data fluctuation due to the extension of time series, monthly runoff data from 1996 to 2015 at the Cuntan hydrological station in the upper reaches of the Yangtze River are used as the basis to develop the annual runoff prediction. Prediction of the five Intrinsic Mode Function (IMF) components and one residual from the Ensemble Empirical Mode Decomposition (EEMD) decomposition used different models of Auto-Regressive Integrated Moving Average Model (ARIMA). Except for the small errors in the first two components, the rest of the ARIMA models are highly fitted. For prediction of runoff in 2014 and 2015 based on EEMD-ARIMA model, the relative errors are 6.26% and −1.17%. Compared with the single ARIMA model and Back Propagation (BP) model, the prediction effect is better, and the prediction method is simple and clear. It is shown that the combined EEMD-ARIMA model is an efficient and useful method for the prediction of annual runoff volume.

Key words: annual runoff, ARIMA, Cuntan hydrological station, EEMD, predictions

HIGHLIGHTS

- The study used 240 months of runoff data from Cuntan hydrological station in the upper reaches of the Yangtze River.
- An annual runoff prediction model based on EEMD-ARIMA hybrid method is proposed.
- The EEMD-ARIMA model outperforms the EMD-ARIMA, BP, and LSTM models.

GRAPHICAL ABSTRACT



1. INTRODUCTION

Runoff prediction plays an important role in the development of hydrology and water conservancy construction in China. Runoff affects the development of soil, plant growth and the formation of rivers and lakes, etc., and is of great importance in the national economy. Runoff is an important condition that constitutes the regional industrial and agricultural water supply and is a constraint on the scale of regional socio-economic development. Predicting annual runoff not only contributes to flood mitigation, water use and sediment management, but also provides valid scientific data for hydrologists and water resources engineers for water development. River runoff sequences are varied and no single method or model can be applied to all runoff predictions. For relatively stable runoff series in the medium and long term, the runoff series can be analyzed periodically by wave splitting method or predicted by Back Propagation (BP) neural network model. Foreign hydrologists have mainly used mathematical models for predicting annual runoff. For example, Thea & Gao (2004) combined a two-site statistical model of point and regional observations for annual runoff prediction in the Worth area; Sedki *et al.* (2009) used a neural network based on a real coded genetic algorithm to predict daily runoff in a semi-arid climate catchment in Morocco; Coulibaly *et al.* (2015) used a recurrent neural network based on a low-frequency climate change index to predict annual runoff in the northern region of Quebec. In addition, Samir Mitra *et al.* (2018) used an artificial neural network based runoff model combining the Feed-Forward Back Propagation (FFBP) algorithm and Levenberg-Marquardt (LM) algorithm to predict runoff in the Hoshangabad basin of the Narmada River. Domestic hydrologists initially used conceptual models for runoff prediction. For example, Lan *et al.* (2020) verified the applicability of the Soil and Water Assessment Tool (SWAT) model for the calculation of runoff into reservoirs in small watersheds in Zhejiang Province by constructing the SWAT model for Lake Man Reservoir, and Liu *et al.* (2019) constructed eight MIKE SHE models for the Yarkant River basin to understand the specific ways in which the driving data affect hydrological processes. These prediction methods and models are effective in prediction analysis, but the evolution of runoff is a complex and variable system, and its instability

and ambiguity are influenced by climate change and human activities. When dealing with such a complex set of nonlinear data, traditional statistical models such as neural networks cannot predict high-frequency mutation data well enough to achieve prediction results (Jin *et al.* 1999). Therefore, decomposing it into relatively smooth modal components first, and then using the corresponding method to build a combined prediction model by reducing the non-smoothness of the runoff series will effectively improve the accuracy of runoff prediction. Empirical Mode Decomposition (EEMD) can process the nonlinear complex sequence, keeping the information of the original sequence, decompose the smooth components of the signal at different levels and different feature scales, so as to obtain some series of eigenmode function (IMF) components of the signal containing the local features of the original signal at different time scales and reveal its intrinsic runoff period, which is very effective for processing complex nonlinear signals, but the decomposition process will produce the loss of IMF components (Ma *et al.* 2020). Ensemble Empirical Mode Decomposition (EEMD) is an improved algorithm based on EMD, which can decompose nonlinear unsteady runoff series from different time scales step by step with good local adaptivity and intuitiveness, and it adds Gaussian white noise sequences to the signal to compensate for the loss of IMF components by using its uniformly distributed characteristics. EEMD extracts the components and trends of the signal in each frequency domain by separating the high and low frequency scales, thus reducing the non-smoothness of the runoff series. ARIMA can solve the problem of random perturbation of runoff series by analyzing different frequency domains and smoothing the data. A The annual runoff is decomposed by using EEMD, and the obtained IMF component series are used as the input data of ARIMA model, and the combined EEMD-ARIMA model can be established to effectively predict the annual runoff of Cuntan hydrological station in the upper reaches of Yangtze River (Ren *et al.* 2015). Predicting annual runoff by using models and methods with better runoff forecasting accuracy is always a challenging task in hydrological forecast reporting research. In order to effectively formulate the water allocation and scheduling plan of the hydrological station, it is necessary to carry out medium and long-term forecasting of its annual runoff, and the prediction accuracy directly affects the scientificity and accuracy of the water scheduling of the hydrological station. Located in the upper reaches of the Yangtze River, the Cuntan Hydrological Station is an important hydrological control station, controlling more than half of the water volume in the upper reaches of the Yangtze River. Effective annual runoff forecasting for it will not only contribute to the rational use of water resources around the Yangtze River, but also provide safe flood control. Early warning brings higher social and economic security.

2. METHODOLOGY

2.1. EEMD model

Ensemble Empirical Mode Decomposition (EEMD) is an improved method based on Empirical Mode Decomposition (EMD) to compensate for the loss of the Intrinsic Mode Function (IMF) component by adding Gaussian white noise to the signal to reduce the polar difference between high and low frequencies and to make the frequency more smooth. This method is suitable for processing sequences of nonlinear, non-stationary signals (Zhang *et al.* 2018). Changes in runoff elements affect the entire hydrological and water resources system, with complex change patterns, it has the characteristics of mutability, randomness, and non-linearity. Using EMD decomposition, the obtained IMF components still need to be judged whether they meet the requirements, and are prone to the phenomenon of mode mixing or modal blending. For instance, a single IMF component signal contains different time scales, or the same time scale appears on different IMF components. EEMD decomposes the signal from the time-scale features of the data itself, and is capable of adaptively extracting the trend of the decomposed components of the signal, extracting the volatility of the signal from it. Decompose complex runoff series into IMF components at different time scales, and transform complex runoff evolution systems into multiple single variable predictions summed. In this way, the non-stationarity of the runoff series can be reduced to reveal the intrinsic runoff cycle (Zheng *et al.* 2021).

The EEMD decomposition steps are as follows.

(1) A Gaussian white noise sequence ($g(t)$) is added to the original runoff sequence ($f(t)$) to obtain an overall runoff sequence ($F(t)$).

$$F(t) = f(t) + g(t) \tag{1}$$

White noise has the property of uniform spectral distribution. After adding white noise, the signals in the sequence with different time scales will be automatically separated to the appropriate reference scale to avoid the appearance of implied scales.

(2) The overall runoff series $F(t)$containing Gaussian white noise series is decomposed by EMD to obtain the IMF component c_{ij} and the trend term $r_i(t)$.

$$F_i(t) = \sum_{j=1}^{n} c_{ij} + r_i(t) \tag{2}$$

(3) Repeat steps (1) and (2), adding different white noise sequences with equal square root mean each time to obtain k different sets of IMF components and residual components.
(4) The IMF components obtained each time are integrated and averaged as the final IMF group.

$$c_j = \frac{1}{k}\sum_{i=1}^{k} c_{ij}(t) \tag{3}$$

where k is the number of added white noise sequences.

2.2. ARIMA model

The Auto-Regressive Integrated Moving Average Model (ARIMA) is a method for analyzing time series, a well-known time series forecasting method proposed by Box and Jenkins in the early 1970s (Wang *et al*. 2011). The autoregressive model (AR) describes the relationship between current and historical values, using the variable's own historical time data to predict itself, and this regression model must satisfy the requirement of stability. The following equation for the q-order autoregressive process needs to be satisfied:

$$y_t = \mu + \sum_{i=1}^{p} \gamma_i y_{t-i} + \varepsilon_t \tag{4}$$

where y_t is the current value, μ is a constant term (indicating that there is no zero mean change in the series data), p is the autoregressive model order, γ_i is the autocorrelation coefficient, and ε_t is the error (white noise series).

The moving average model (MA) is concerned with the accumulation of error terms in the autoregressive model, and the moving average method is effective in eliminating random fluctuations in forecasts. The following equation for the q-order autoregressive process needs to be satisfied:

$$y_t = \mu + \varepsilon_t + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} \tag{5}$$

where q is the moving model order and θ_i is the error term combination parameter.

Autoregressive sliding average model (ARMA) combines autoregression with moving average. The following formula needs to be satisfied:

$$y_t = \mu + \sum_{i=1}^{p} \gamma_i y_{t-i} + \varepsilon_t + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} \tag{6}$$

In ARIMA (p, d, q), p is the autoregressive term, d is the number of differences made when the time series becomes stationary, and q is the number of moving average terms. p and q are generally determined by the trailing and truncated autocorrelation function (ACF) and partial autocorrelation function (PACF). A drag tail is a series that decays monotonically or oscillates at an exponential rate, while a truncated tail is a series that becomes very small from a point in time (Qu *et al*. 2015). The basic principle of model identification is to observe the autocorrelation and bias correlation function graph of the

model. If PACF truncates the tail and ACF trails then the time series fits the AR model (d, $q = 0$); if PACF trails and ACF truncates, the time series fits the MA model (p, $d = 0$); if both PACF and ACF drag the tail, the time series fits the ARIMA (p, d, q) model. ARIMA is based on stochastic theory, which models the perturbation terms and describes them approximately with certain mathematical models, so that the model integrates past values, initial values and error values to predict future values.

The basic principle of ARIMA is a model built by transforming a non-stationary time series into a stationary time series and then regressing the dependent variable on only its lagged values and the present and lagged values of the random error term. The operating principle of the model is shown in Figure 1.

2.3. Combined model based on EEMD-ARIMA

The EEMD method and ARIMA model were coupled to predict the annual runoff series. EEMD can decompose complex runoff series into stable components, providing a smooth premise for ARIMA prediction models. The structure of the combined EEMD-ARIMA forecasting model is shown in Figure 2.

The methodological steps of the combined EEMD-ARIMA forecasting model prediction are as follows:

(1) EEMD decomposition of annual runoff data using MATLAB to obtain the IMF components and trends of the runoff series.
(2) Verify that the decomposed IMF components and trend terms are smooth series; if smooth, then d takes 0, otherwise d is determined by the difference order, and p and q are determined by the autocorrelation function (ACF) and partial correlation function (PACF).
(3) The decomposed IMF components and trend terms are modeled with an ARIMA forecasting model, and the corresponding forecasts are made for each component separately.
(4) The predicted results of all components are summed to obtain the final result as the predicted value of annual runoff.

2.4. The data source

Cuntan hydrological station is located in Cuntan Town, Jiangbei District, Chongqing City, 7.5 km downstream of the confluence of the main stream and Jialing River in the upper reaches of the Yangtze River (Zhang & Bian 2009), with a watershed area of 866,600 square kilometers, and is an important inlet control station of the Yangtze River Three Gorges Water Conservancy Hub and an important water control station in the upper reaches of the Yangtze River. Cuntan hydrological station

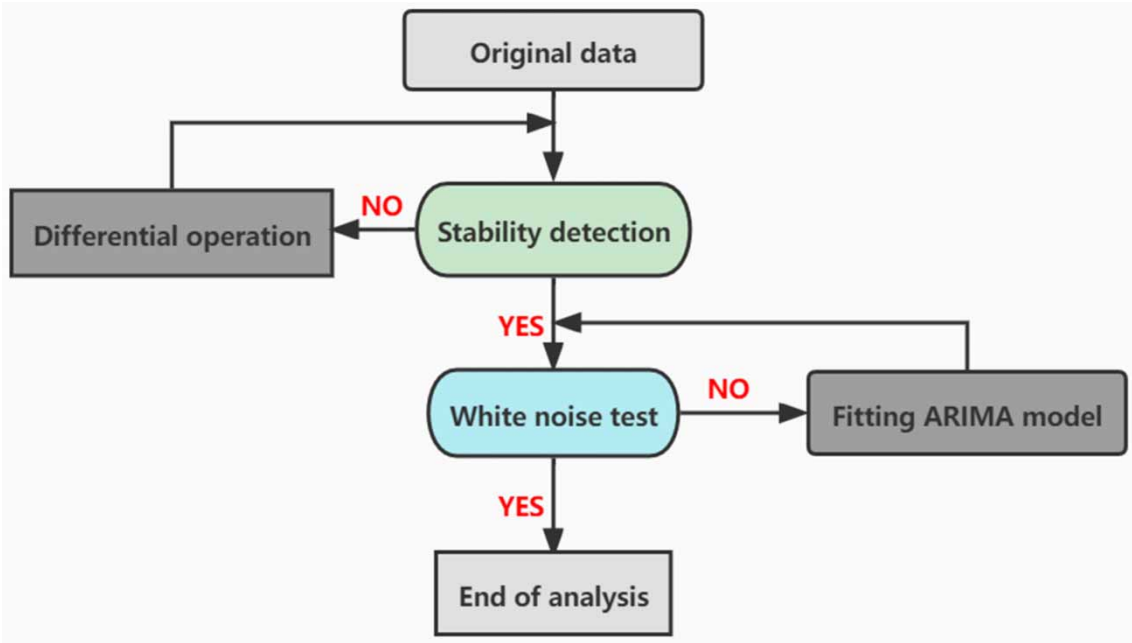


Figure 1 | ARIMA prediction model operating principle.

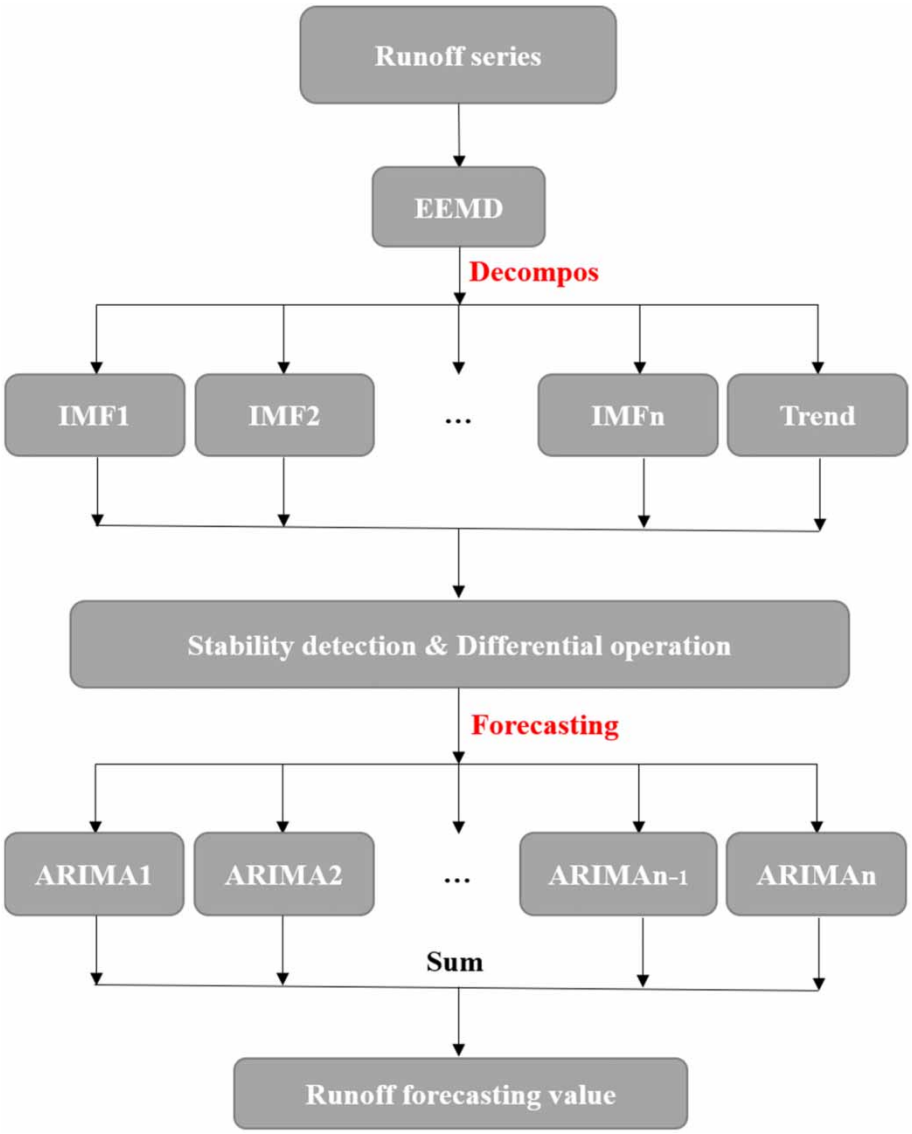


Figure 2 | Forecasting structure of the combined EEMD-ARIMA model.

was built in 1936, is a national hydrological station, and belongs to the Yangtze River Commission Hydrological Bureau management. The station measurement items are water level, water quality, flow, rainfall, etc., and measurement information is continuous and complete. Since the establishment of the station, the highest water supply discharge of the historical survey was 1870, the water level reached 196.25 m, the flow rate was 99,400 m^3/s, the average multi-year evaporation was 793.20 mm, the average annual precipitation was 1,078 mm, and the average annual runoff was about 420 mm deep. Cuntan hydrological station is an important hydrological control station in the upper reaches of Yangtze River, affecting four major rivers on it, Jinsha River, Min River, Tuo River and Jialing River, controlling about 60% or more of the water volume in the upper reaches of Yangtze River. Reasonable measurement and prediction of this hydrological station accumulates experience for future flood control forecasting, so as to better serve the flood control of Yangtze River and Three Gorges dispatching, and provide a better social and economic benefit (Guo *et al.* 2015).

In this paper, the runoff data of Cuntan hydrological station from 1996 to 2015 were used, in which the runoff data from 1996 to 2013 were used as the simulation training of the model, and the data from 2014 to 2015 were used for validation. The runoff data were obtained from the precipitation and air temperature (1961–2020) at the Cuntan Hydrological Station in the upper reaches of the Yangtze River provided by the National Meteorological Information Center of the China Meteorological Administration (http://data.cma.cn) and passed stricter data quality control. Taking into account the update and

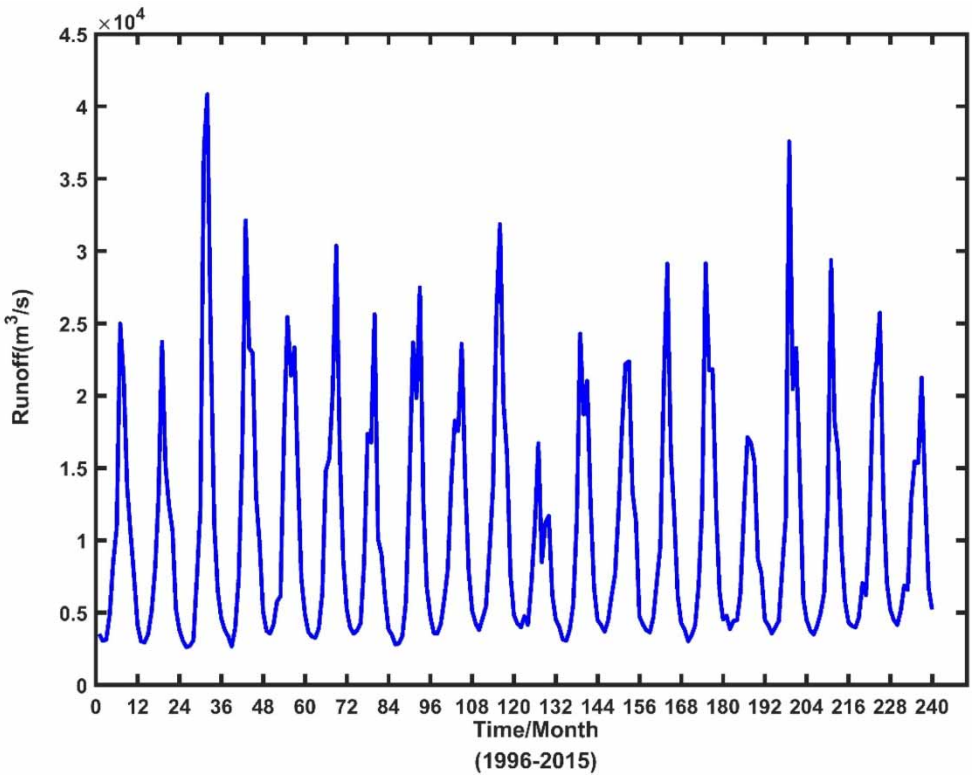


Figure 3 | Cuntan hydrological station 1996–2015 runoff change map.

completeness of the data, if the proportion of missing data exceeded 0.2%, it was excluded. For other stations with missing data, we preferentially select the data of the nearest stations, and perform linear regression every five years to fill in the gaps of missing data. The changes of runoff from 1996 to 2015 at Cuntan Hydrological Station are shown in Figure 3.

From Figure 3, it can be seen that the runoff from 1996 to 2015 at Cuntan hydrological station in the upper reaches of Yangtze River has certain volatility and inconsistent fluctuations, reflecting the instability of this runoff series, and EEMD has great advantages for non-stationary and non-linear time series data processing, so it is reasonable to use this decomposition method.

3. OBTAINED RESULTS

3.1. EEMD decomposition

The EEMD algorithm was used to decompose the runoff data from Cuntan Hydrological Station in the upper reaches of Yangtze River, and it was found that when the noise variance was taken as 0.3 and the noise count was 120, the decomposition effect was more satisfactory. After EEMD decomposition of this runoff series, five IMF components and one trend quantity were obtained as shown in Figure 4.

From the decomposition results, it can be seen that the first IMF component has the largest volatility with high frequency and short wavelength, and after the decomposition of EEMD, the other IMF components gradually decrease in amplitude, frequency and wavelength, making the volatility and non-stationarity of the time series greatly reduced. The trend term shows that the time series decreases year by year.

3.2. EEMD-ARIMA combined model forecast

The annual runoff from the Cuntan hydrological station in the upper reaches of the Yangtze River is decomposed by EEMD. Five IMF components and a trend term contribute differently to the prediction series, and the IMF components and the trend term can be regarded as the drivers of the annual runoff, so each component can be predicted, and the sum of the IMF components and the trend term prediction results is the prediction result of the annual runoff. Using the 2014–2015 runoff data as

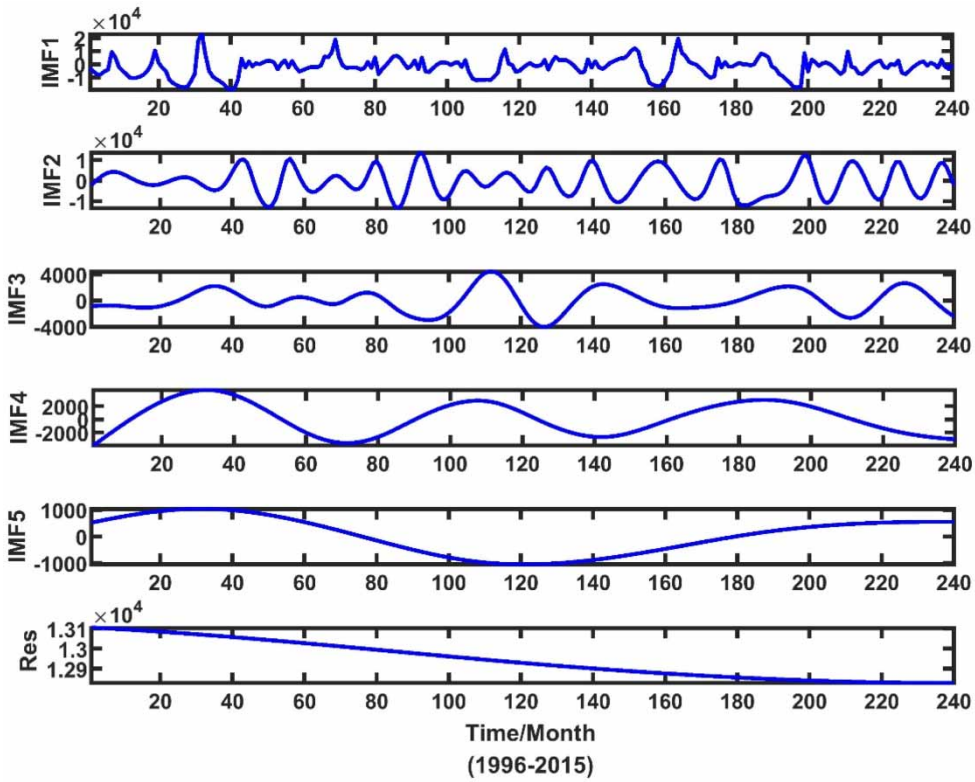


Figure 4 | Results of EEMD decomposition of runoff from 1996 to 2015 at Cuntan hydrological station.

a prediction sample for simulation, ARIMA predictions were made for each of the five IMF components and a trend term. It can be obtained that the ARIMA model selected for each component varies, and the autoregressive order p and the moving regressive order q differ, and the difference operation is also needed to determine the difference order d when the signal is not stable. Looking at the autocorrelation function plot and the partial autocorrelation function plot for each component, the model used for each component can be determined (Zhu & Zhang 2021). The IMF1 component has the smallest Akaike information criterion (AIC) and the largest R^2 (R-Squared, goodness of fit) when using the model ARIMA (1, 0, 2), and the prediction results are more accurate (Yang *et al.* 2014); IMF2 uses the ARIMA (1, 1, 2) model, and the original IMF2 signal is poorly smoothed and needs to be differenced once to make it stable. IMF3, IMF4, and IMF5 have PACF fifth-order truncated tails and ACF trailing tails, so this component time series is suitable for the AR model, using the ARIMA (5, 0, 0) model. Res differential is not significantly different from the original series, and autocorrelated trailing, biased autocorrelated first-order truncated tail, so the ARIMA (1, 0, 0) model is used. The predicted results are shown in Figure 5.

IMF component and trend forecast values and errors are shown in Table 1. From Table 1, it can be seen that IMF1 and IMF2 have large prediction errors, and IMF3, IMF4, IMF5, and trend terms have small prediction errors. The non-smoothness of the first two components is relatively high, and the prediction effect of the latter components and the trend term is significant, indicating its low non-smoothness Zhang *et al.* 2008. The total flow for each month is obtained by adding up the corresponding components, and then the total flow for each month is added up to obtain the total annual flow for 2014 and 2015, which naturally leads to the required predicted annual runoff. A comparison of the overall raw and forecast data for 2014 and 2015 is shown in Figure 6, and the IMF component and trend term forecasts and errors for both years are shown in Table 1.

From Table 1, we can see that the prediction results of IMF1 and IMF2 have relatively large errors, and the prediction results of the remaining components are in high agreement with the original series. After modeling the fitted model, the residuals of its prediction result series are subjected to white noise test, and they are all white noise series, which proves that the information extraction of the model is sufficient to achieve the prediction purpose (Farhan & Rajiv 2020). The

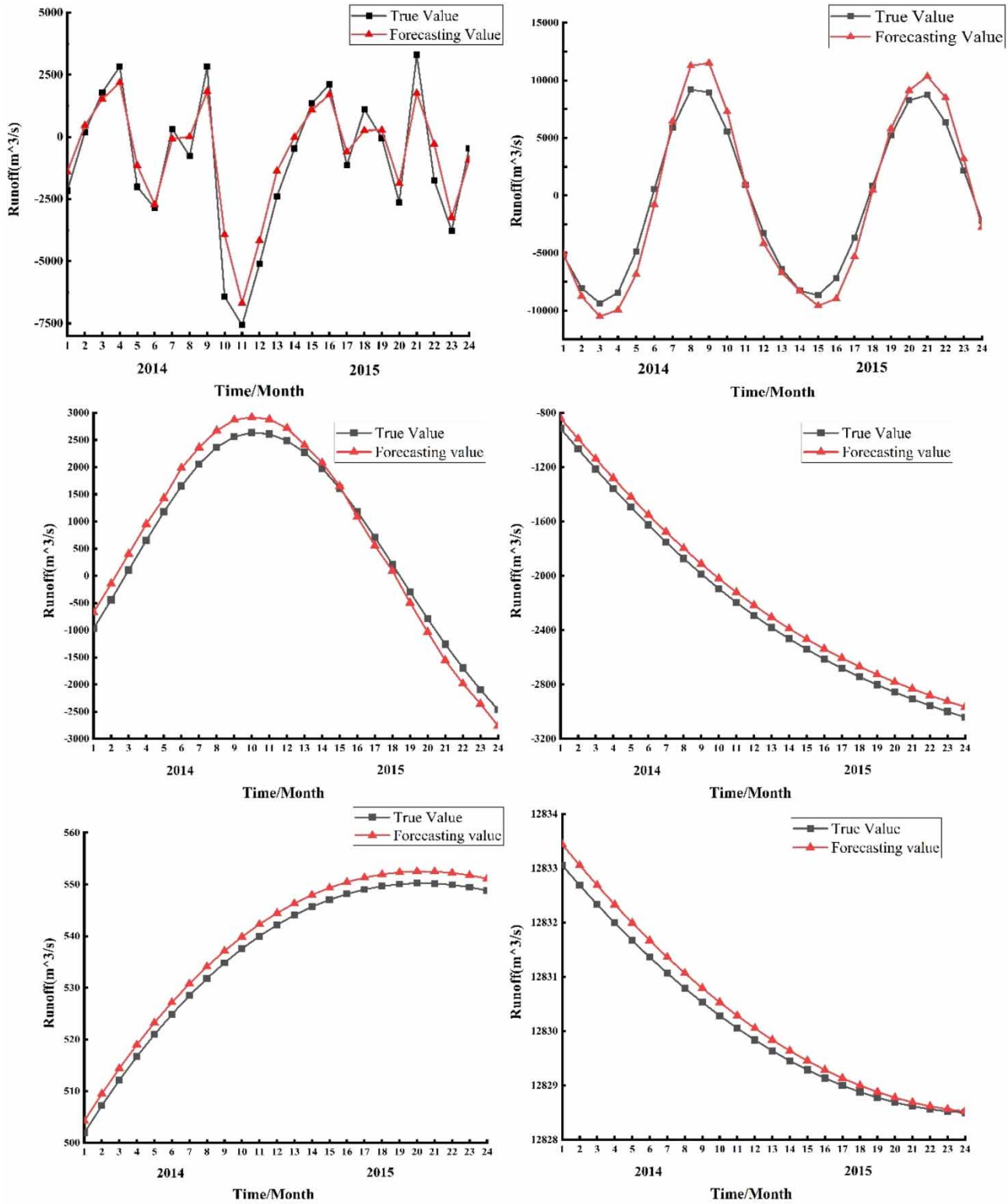


Figure 5 | ARIMA forecasting results for IMF components and trend terms.

predicted results of each component were summed to obtain the final annual runoff results, and the predicted results and errors are shown in Table 2.

Table 1 | IMF component and trend term forecasts and errors

IMF component	Time (Year)	Original data	Predictive value	Relative error (%)	ARIMA (p,d,q)
IMF1	2014	−497.06	−371.19	−25.32	(1, 0, 2)
	2015	−211.36	−83.96	−60.28	(1, 0, 2)
IMF2	2014	−211.63	−225.57	6.59	(1, 1, 2)
	2015	−126.24	−111.58	−11.61	(1, 1, 2)
IMF3	2014	443.79	536.14	20.81	(5, 0, 0)
	2015	−17.06	−61.16	72.11	(5, 0, 0)
IMF4	2014	−522.18	−498.53	−4.53	(5, 0, 0)
	2015	−866.96	−843.31	−2.73	(5, 0, 0)
IMF5	2014	165.53	166.25	−0.43	(5, 0, 0)
	2015	172.98	173.71	−0.42	(5, 0, 0)
Res	2014	4,046.48	4,046.58	0	(1, 0, 0)
	2015	4,045.73	4,045.76	0	(1, 0, 0)

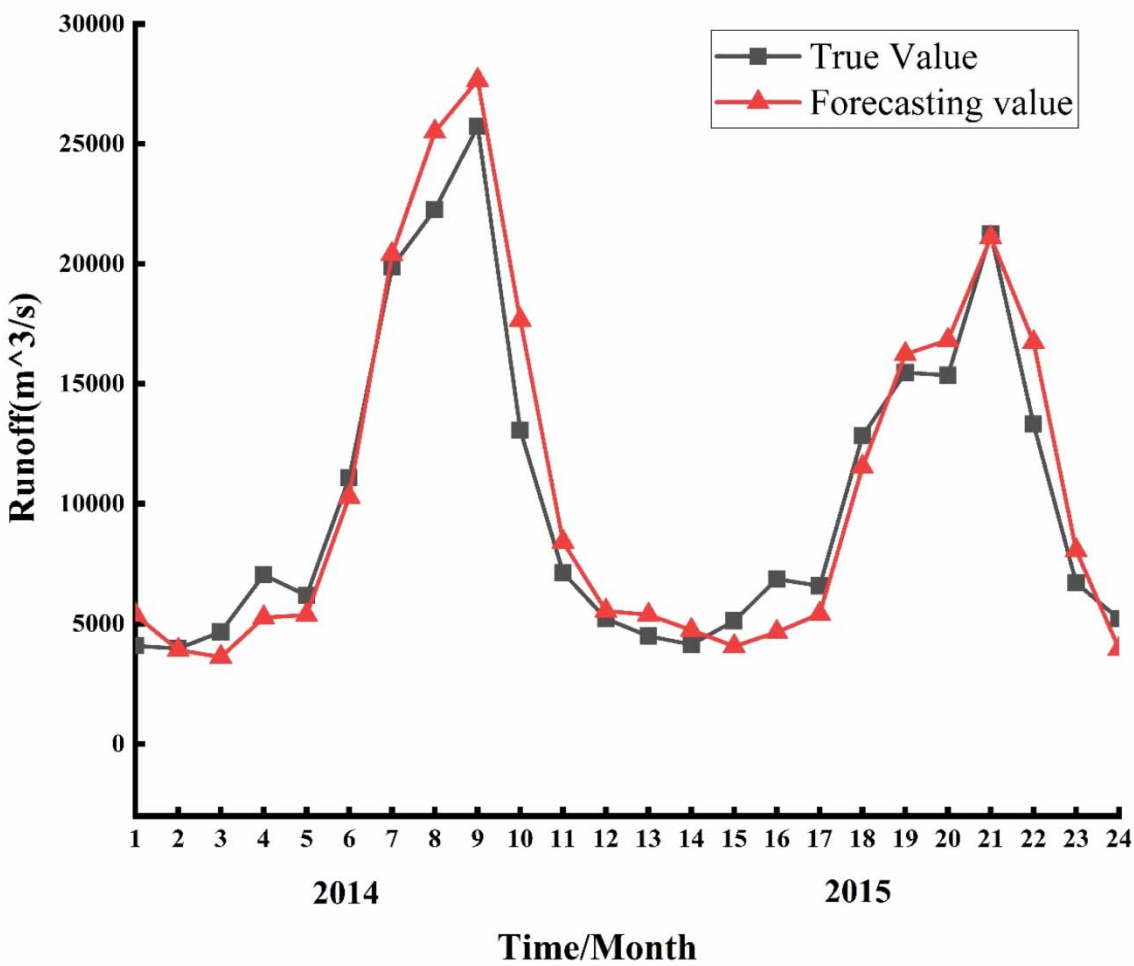


Figure 6 | ARIMA model prediction results for runoff in 2014 and 2015.

From the above results, it can be seen that the combined EEMD-ARIMA model predicts the runoff volume very well, and the predicted trend is basically consistent with the original data. The stability and trends of the IMF components are different, and then the contribution to runoff is also different. Although the prediction error of IMF1 is relatively high, the proportion of

Table 2 | Annual runoff prediction results and errors

Time (year)	True value (10^8 m^3)	Prediction value (10^8 m^3)	Relative error of prediction (%)
2014	3,424.91	3,653.68	6.26
2015	3,083.44	3,119.51	−1.17

IMF1 in.the runoff series is small, and the contribution of IMF1 to runoff prediction is small, so its prediction error does not affect the overall runoff prediction error. The prediction of 2014 and 2015 runoff by using monthly runoff data from 1996 to 2015 is feasible and accurate with relative errors of 6.26% and −1.17%.

4. DISCUSSION

In order to compare the prediction effects of the models, we selected three models for prediction respectively, including single ARIMA model, EMD-LSTM combined model and BP model. Long Short-Term Memory (LSTM) is a special Recurrent Neural Network (RNN). This method of processing time series not only solves the problem of artificially prolonged time tasks that RNN is difficult to solve, but also solves the problem of gradient disappearance that RNN is prone to. It is widely used in forecasting research on medium and long time series. BP can learn and store a large number of input-output pattern mappings without exposing the mathematical equations describing such mappings in advance. Its learning rule is to use the fastest descent method to continuously adjust the weights and thresholds of the network through backpropagation to minimize the sum of squared errors of the network. It is also used in time series forecasting research. The prediction results and relative errors are shown in Tables 3 and 4.

It can be seen from Table 3: under the BP prediction model, the prediction effect is not ideal, and the data is more than 20% higher than the real data, and the prediction results are the worst among the four models, the most obvious one is 2015, which is significantly higher than the original data and has the largest error. The prediction effect of the single model ARIMA is better than that of the BP model, but due to the instability of the original time series, it cannot meet the requirements of stationarity, resulting in a relative error of more than 10%, which cannot reach the prediction accuracy, and the overall error is large and deviates from the true value. Under the combined prediction model of EMD-LSTM, the prediction result is closer to the real value than the BP model and ARIMA model, and the change trend of the result is basically the same as the measured

Table 3 | Results and errors of different prediction methods

Algorithms	Time (year)	Relative error
EEMD-ARIMA	2014	6.26%
	2015	−1.17%
ARIMA	2014	15.42%
	2015	14.24%
EMD-LSTM	2014	−8.54%
	2015	7.61%
BP	2014	23.07%
	2015	29.67%

Table 4 | Error comparison table between the combined EEMD-ARIMA model and other models

Algorithms	MAE m	MAPE	NS
EEMD-ARMA	0.16	0.05	0.92
ARIMA	0.73	0.23	0.75
EMD-LATM	0.47	0.15	0.88
BP	0.81	0.25	0.71

sequence. However, due to the phenomenon of mode mixing or mode aliasing in the EMD decomposition, the final relative error exceeds 7%, which does not meet the prediction requirements. The EEMD-ARIMA combined model has the best prediction effect, which is highly consistent with the change trend of the original series, as shown in Figure 6, the data has basically fitted the original sequence, and some individual data is too small. It shows that the decomposition of EEMD cannot only restrain the modal aliasing phenomenon of EMD, but also better adapt to the complex frequency information contained in the sequence. The combination with ARIMA is also feasible, the accuracy is improved, and the frequency is high after EEMD decomposition. The series is more stationary, which is beneficial to the final prediction result.

To measure the prediction accuracy of the EEMD-ARIMA coupled model, the mean relative error (MAPE), Nash-Sutcliffe efficiency coefficient (NSE), mean absolute error (MAE), and root mean square error (RMSE) between the predicted value and the actual runoff data were used as an evaluation criterion. The specific formula is as follows:

$$\mathrm{MAPE} = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100\% \tag{7}$$

$$NSE = 1 - \frac{\sum_{t=1}^{N}(y_t - \bar{y}_t)^2}{\sum_{t=1}^{N}(y_t - \mu_t)^2} \tag{8}$$

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}\left|y_i - \hat{y}_i\right| \tag{9}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \tag{10}$$

In the formula: y_i is the actual value of runoff at time i; $\hat{y}_i$ is the predicted value of runoff at time i; N is the total length of the time series.

It can be seen from Table 4: The NS of each model has reached more than 0.7 at the time of prediction, and the pass rate is high. With the optimization of the single model to the combined model, the MAE and MAPE indicators are getting smaller and smaller, and the NS is closer to 1. The predicted effect is highly close to the true value.

It was found that the EEMD-ARIMA prediction model had the smallest relative error and the prediction results were better than those predicted using a single prediction algorithm. EEMD decomposition can effectively decompose the original signal into IMF components of different time scales. Since the original series has some seasonality due to the prediction using monthly flow data, using this method can effectively reduce the non-smoothness of this runoff series. ARIMA models can be modeled based on the series obtained from the decomposition to find highly fitted ARIMA prediction models. After analysis, it is found that the source of error of this algorithm is the short time series and the large base of the original series. As the Yangtze River is the world's largest hydroelectric power, for the Cuntan Hydrological Station in the upper reaches of the Yangtze River, using monthly runoff data to predict annual runoff data makes the seasonality more prominent, so to achieve better prediction results it is necessary to lengthen the time series and use annual runoff series for prediction.

5. CONCLUSION

The annual runoff time series of Cuntan Hydrological Station in the upper reaches of Yangtze River has high stochasticity and instability; this paper combines the EEMD decomposition model and ARIMA prediction model to establish a combined EEMD-ARIMA model, and the following conclusions were drawn:

(1) EEMD reduces the non-smoothness of the original time series and meets the requirement that the ARIMA forecasting model must be predicated on a smooth series. The results after the prediction of the components using different ARIMA model modeling were tested for white noise series, and their residuals were all white noise series, which proved that the model information extraction was adequate. The relative error of prediction results does not exceed 7%, and the method is better than a single prediction model and has good prediction effect and high accuracy.

(2) The runoff time series is decomposed by EEMD, and the signal is decomposed into five IMF components and one residual component, whose predicted value is equal to the sum of the six components. Although the relative errors of IMF1 and IMF2 predictions are large, their contributions to the overall signal are small and do not affect the overall prediction results much.

(3) The final summation of monthly flow data to predict annual runoff data expands the effect caused by seasonality in it, and to achieve better predictions it is necessary to use annual runoff data and expand the length of the time series. However, this has limitations in the absence of data support, which requires exploring better methods for in-depth research.

AVAILABILITY OF DATA AND MATERIALS

Data and materials are available from the corresponding author upon request.

AUTHOR CONTRIBUTION

All authors contributed to the study conception and design. Writing and editing: Xianqi Zhang and Yaohui Lu. Preliminary data collection: Dong Zhao and Bingsen Duan. Chart editing: Xilong Wu and Guoyu Zhu. All authors read and approved the final manuscript.

FUNDING

This work was supported by the Key Scientific Research Project of Colleges and Universities in Henan Province (CN) [grant numbers 17A570004].

DATA AVAILABILITY STATEMENT

All relevant data are included in the paper or its Supplementary Information.

CONFLICT OF INTEREST

The authors declare there is no conflict.

REFERENCES

Coulibaly, P., Anctil, F., Rasmussen, P. & Bernard, B. 2015 Are current neural networks approach using indices of low-frequency climatic variability to predict regional annual runoff. *CGHU Special Issue of Hydrological Processes* **14** (15), 2755–2777.

Farhan, M. K. & Rajiv, G. 2020 ARIMA and NAR based prediction model for time series analysis of COVID-19 cases in India. *Journal of Safety Science and Resilience* **1** (1), 12–18.

Guo, S. L., Guo, J. L., Hou, Y. K., Hou, L. H. & Hong, X. J. 2015 Projecting future runoff changes in the Yangtze River basin based on Budyko's hypothesis. *Advances in Water Science* **26** (02), 151–160.

Jin, J. L., Yang, X. H. & Ding, J. 1999 A neural network-based model for annual runoff prediction. *People's Yangtze River* (S1), 58–59 + 62.

Lan, X. C., Wan, Y. S. & Zhang, Z. Q. 2020 Application of SWAT model to runoff calculation in small watersheds in Zhejiang. *People's Pearl River* **41** (12), 27–31 + 52.

Liu, J., Liu, X. W., Liu, T. & Qian, B. 2019 Influence of driving data on different result elements in watershed hydrological simulation. *Journal of Natural Resources* **34** (11), 2481–2490.

Ma, F. H., Jin, Y. C. & Sun, C. Y. 2020 Short-time prediction model for metro passenger flow based on EMD-optimized NAR. *Journal of Applied Sciences* **38** (06), 936–943.

Qu, L. L., Qi, L. Y. & Gao, S. Z. 2015 Application of time series analysis in runoff prediction. *Anhui Agricultural Science* **43** (21), 23–24 + 103.

Ren, B., Hu, Q. W. & Ren, Q. Z. 2015 Research on annual runoff prediction based on EMD-ARMA. *Soil and Water Conservation Applied Technology* **2015** (02), 25–26.

Samir Mitra, J. N., Zhu, Q. P., An, X. W., Mei, Y. & Chen, D. Z. 2018 Research on the prediction of hydrological time series based on PSO-KELM model. *China Rural Water and Hydropower* **429** (07), 21–24.

Sedki, A., Ouazar, D. & Mazoudi, E. E. 2009 Evolving neural network using real coded genetic algorithm for daily rainfall–runoff prediction. *Expert Systems with Applications* **36** (3), 4523–4527.

Thea, T. W. & Gao, J. B. 2004 Detecting dynamical change sin time series using the permutation entropy. *Physical Review. E, Statistical, Nonlinear, and Soft Matter Physics* **70** (4 Pt 2), 046217.

Wang, Y., Gu, H. Y. & Xu, W. K. 2011 Prediction of annual runoff in river based on ARMA model. *Journal of Harbin University of Commerce (Natural Sciences Edition)* **27** (03), 338.

Yang, G., Meng, J. & Wang, S. 2014 Pruning algorithm of classificaiton and regression decision tree based on Akaike information criteria. *Journal of Computer Applications* **34** (S2), 147–150.

Zhang, C. N. & Bian, W. M. 2009 *Analysis for the Wavelet Cycle of Recent 50-Years Flow of Cuntan Gauge*. Jilin Water Resources.

Zhang, J. X., Ma, X. Y., Zhao, W. H., Hao, J. J. & Qu, J. N. 2008 Application of the life cycle-Markov combination model for river annual runoff prediction. *Journal of Hydroelectric Engineering* **27** (6), 32–36.

Zhang, X. Q., Song, C. & Hu, D. K. 2018 *Research on Groundwater Depth Prediction Model of Irrigation District Based on EEMD and Elman Neural Network*. Water Saving Irrigation.

Zheng, F. F., Wang, W. S. & Zhang, L. T. 2021 Application of improved EEMD-NNBR coupled model in annual runoff prediction. *People's Pearl River* **42** (02), 1–6.

Zhu, Y. M. & Zhang, Z. H. 2021 Portfolio forecasting of bilateral trade volume between China and the United States based on CEEMD-ARIMA. *China Collective Economy* **2021** (34), 92–95.

First received 15 April 2022; accepted in revised form 4 July 2022. Available online 12 July 2022