scientific reports

Check for updates

OPEN

Runoff prediction of lower Yellow River based on CEEMDAN–LSSVM–GM(1,1) model

Shaolei Guo[1], Yihao Wen[1], Xianqi Zhang[1,2,3]✉ & Haiyang Chen[1]

Accurate medium and long-term runoff forecasts play a vital role in guiding the rational exploitation of water resources and improving the overall efficiency of water resources use. Machine learning is becoming a common trend in time series forecasting research. Least squares support vector machine (LSSVM) and grey model (GM(1,1)) have received much attention in predicting rainfall and runoff in the last two years. "Decomposition-forecasting" has become one of the most important methods for forecasting time series data. Complete ensemble empirical mode decomposition with adaptive noise (CEEMDAN) decomposition method has powerful advantages in dealing with nonlinear data. Least squares support vector machine (LSSVM) has strong nonlinear fitting ability and good robustness. Gray model (GM(1,1)) can solve the problems of little historical data and low serial integrity and reliability. Based on their respective advantages, a combined CEEMDAN–LSSVM–GM(1,1) model was developed and applied to the runoff prediction of the lower Yellow River. To verify the reliability of the model, the prediction results were compared with the single LSSVM model, the CEEMDAN–LSSVM model and the CEEMDAN–support vector machines (SVM)–GM(1,1). The results show that the combined CEEMDAN–LSSVM–GM(1,1) model has a high accuracy and the prediction results are better than other models, which provides an effective prediction method for regional medium and long-term runoff prediction and has good application prospects.

Runoff prediction is an important element in hydrological forecasting research, and its prediction results can provide the basis for flood and drought prevention, reservoir scheduling and hydroelectric power generation[1,2]. Because the time interval between medium and long-term forecasts is too long and there are too many uncertainties, the medium and long-term runoff series are highly nonlinear and random, and the forecast results are not ideal[3]. How to establish a runoff prediction model with higher forecast accuracy is extremely important for the optimal allocation of water resources in the basin and regional development planning.

Currently, medium and long-term runoff prediction mainly includes process-driven model and data-driven model[4]. Process-driven models rely on the scientific theories of hydrology, hydraulics and erosion dynamics[5], consider the physical mechanisms within the water cycle system in all aspects, and integrate factors such as land use, soil type, meteorological changes, and water quality advantages and disadvantages to simulate the runoff production process, with the SWAT model[6], the DHSVM model[7], and Xin'anjiang model[8] as representatives. Since the rainfall runoff process is influenced by various factors such as topography, rainfall distribution, soil properties, land use, and climate change, process-driven models require a large amount of data for modeling, and insufficient data will have an impact on the successful establishment of the model, while data-driven models require little data information and have a fast development time, so the data-driven approach is still mainly used for medium and long-term runoff prediction[9,10]. Data-driven models target the optimal relationship between data and use mathematical methods to establish a relationship between the input data and the output target of the model. Traditional data-driven models include multiple regression models, time series models, mathematical statistical models, etc. With the development of computer theory, modern data-driven models make more use of neural networks, fuzzy mathematics, gray systems and support vector machines for hydrological data prediction[11]. Typically, data-driven models can be further classified into three categories: statistical-based

[1]Water Conservancy College, North China University of Water Resources and Electric Power, Zhengzhou 450046, China. [2]Collaborative Innovation Center of Water Resources Efficient Utilization and Protection Engineering, Zhengzhou 450046, China. [3]Technology Research Center of Water Conservancy and Marine Traffic Engineering, Zhengzhou 450046, Henan Province, China. ✉email: 1368928033@qq.com

predictive models, machine learning models, and combinatorial models. Thomas et al.[12] proposed an autoregressive model (AR). Carlson et al.[13] applied an autoregressive moving average (ARMA) model to annual runoff predictions. Elshorbagy et al.[14] applied a multiple linear regression model (MLR) for daily runoff prediction. The above statistical modeling methods are based on linear regression theory, and the prediction accuracy still needs to be improved when dealing with complex runoff information, and more powerful models are needed for runoff prediction to deal with non-linear and complex runoff simulations[15]. Cortes et al.[16] proposed a support vector machine model (SVM) based on the Vapnik–Chervonenkis dimensionality theory and structural risk minimization theory in statistical theory. Liao et al.[17] attempted to apply SVM models to the field of runoff prediction and compared them with the threshold regression TR model to confirm the superiority of SVM models in runoff prediction. Li et al.[18] proposed a support vector machine model based on the principle of least squares (LSSVM), and experimentally confirmed that the runoff prediction results of the LSSVM model can still maintain high accuracy when the sample data are small. Shabri et al.[19] applied the cross-validation method and grid search method to the LSSVM model for the selection of parameters in the runoff prediction process, which greatly reduced the modeling time of the LSSVM model.

However, due to the high nonlinearity of runoff series, relying on raw data to build machine learning models may not meet the prediction needs. Based on the research on machine learning models, scholars at home and abroad have conducted extensive research on improving the accuracy of runoff prediction by pre-processing time series. The prediction process can be simplified by pre-noise reduction or decomposition of the data, and the non-linear and non-smooth characteristics in the hydrological series can be analyzed in advance before the prediction, which can effectively improve the efficiency of the later prediction. Mallat[20] proposed an easy to implement and less computational multi scale analysis algorithm, which processes the raw data into multiple layers of smoother low-frequency components and high-frequency components, which can improve the prediction accuracy of the raw time series data. Huang et al.[21] proposed the empirical mode decomposition (EMD), which is a method for the analytical processing of nonlinear and non-smooth signals. The original complex nonlinear signal is decomposed by EMD into several intrinsic mode functions (IMF) and a residual. And the IMF obtained by the EMD method may have the problem of mode mixing. Wu et al.[22] proposed the ensemble empirical mode decomposition (EEMD) method, which adds Gaussian white noise to the original signal during the decomposition process and can effectively suppress mode mixing. Complete ensemble empirical mode decomposition with adaptive noise (CEEMDAN) decomposition of adaptive noise adds a finite number of adaptive white noises at each stage, reducing the number of iterations and improving reconstruction accuracy compared to EEMD[23]. Dragomiretskiy et al.[24] proposed the Variational mode decomposition (VMD) method. The method is an adaptive, fully non-recursive approach to modal variation and signal processing. Effective separation of intrinsic modal functions (IMF) and frequency domain division of signals can be achieved. The effective decomposition components of the given signal are obtained, and finally the optimal solution of the variational problem is obtained. Raj Huan et al.[25] proposed an EEMD–LSSVM model for predicting dissolved oxygen data based on time series and compared the prediction results with a single LSSVM model, and the results showed that the EEMD–LSSVM model has high prediction accuracy and generalization ability. Huang et al.[26] applied the EMD–GM(1,1) model to predict landslide deformation and compared it with the traditional GM(1,1) model, and found that EMD–GM(1,1) had higher prediction accuracy. Jamei et al.[27] used the Multivariate Variational Mode Decomposition (MVMD) decomposition method with the LSSVM model applied to predict the daily wave energy in coastal areas, this research can help authorities in the field of renewable and sustainable energy for better planning and development. Jamei et al.[28] used a novel decomposition method, namely time varying filter-based empirical mode decomposition (TVF–EMD), before predicting daily flood levels at two sites in the Clarence River, Australia. Raj et al.[29] studied the sea level prediction problem for small island countries such as Kiribati and Tuvalu, and he used a new method of data decomposition, namely, continuous variational modal decomposition (SVMD).

The water resources of the Yellow River basin are facing the problems of large dynamic changes in river runoff hydrology, uneven seasonal distribution, and large sand content of rivers. The source of the Yellow River has a decreasing trend of historical flood runoff under the influence of climate change and human activities, and the middle and lower reaches of the Yellow River are in the temperate monsoon zone with large seasonal variation of precipitation, especially under the influence of global warming, which has increased the dynamic variability of natural runoff of the Yellow River. Therefore, it is particularly necessary to study the prediction of runoff in the Yellow River basin, especially in the densely populated areas of the middle and lower reaches of the Yellow River. Timely and accurate prediction of natural runoff in the middle and lower reaches of the Yellow River will provide a basis for promoting the rational and optimal allocation of water resources in the Yellow River area, maximizing the demand for water for agriculture, industry and domestic use, and promoting the virtuous cycle of the regional ecosystem, as well as providing a basis for rational allocation and development of water scheduling implementation plans. Based on the good decomposition effect of CEEMDAN and the combination effect with LSSVM model and GM(1,1), a CEEMDAN–LSSVM–GM(1,1) model is constructed in this paper to predict monthly runoff from four hydrological stations in the lower Yellow River, and the prediction results are compared with those of several different models to illustrate the effectiveness of the model and provide a new combined machine learning model solution to the runoff prediction problem.

Methods

CEEMDAN. The Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN) method is an improved EMD algorithm, which overcomes the modal confusion problem of the original EMD method and can more accurately extract from the nonlinear sequences the relatively smooth Intrinsic Mode Functions (IMF) and Residuals (Res) of these components of the nonlinearity decreases layer by layer[30], which

can clearly reflect the fluctuation characteristics of different cycles and provide convenience for the analysis and prediction of complex sequences. Figure 1 shows the working principle of CEEMDAN.

The runoff time series is denoted by $S(t)$ and $V^i(t)$ is a Gaussian white noise series with standard normal distribution added in the ith trial, so the ith signal series can be expressed as Eq. (1):

$$S^i(t) = S(t) + \varepsilon_0 V^i(t) \quad i \in \{1, 2, \ldots, M\} \tag{1}$$

where ε_0 is the noise factor, M is the number of integrations, generally between 10 and 20.

Define the operator $emd_i(\,)$ as the modal component of the ith stage generated by applying the EMD algorithm, and the ith modal component obtained after decomposition with the CEEMDAN algorithm is noted as IMF_i.

The specific steps of the algorithm are as follows:

(1) A series of adaptive Gaussian white noise $\varepsilon_0 V^i(t) \quad i \in \{1, 2, \ldots, M\}$ is first added to the original time series $S(t)$, M trials are performed for the ith signal $S^i(t)$, and the new time series $S^i(t)$ is decomposed using the EMD algorithm, and the first IMF component $IMF_1^i(t)$ obtained from the decomposition is averaged to obtain the first CEEMDAN modal component.

$$IMF_1(t) = \frac{1}{M}\sum_{i=1}^{M} IMF_1^i(t) \tag{2}$$

(2) The first component $IMF_1(t)$ of the decomposition is removed from the original time series $S(t)$ to obtain the first residual series $R_1(t)$.

$$R_1(t) = S(t) - IMF_1(n) \tag{3}$$

(3) In the same way as in step (1), a series of adaptive Gaussian white noise $\varepsilon_1 V^i(t) \quad i \in \{1, 2, \ldots, M\}$ is added to the first residual sequence, and the EMD decomposition of the sequence $R_1(t) + \varepsilon_1 emd_1\left(V^i(t)\right)$ is continued to obtain the second IMF modal component.

$$IMF_2(t) = \frac{1}{M}\sum_{i=1}^{M} emd_1\left(R_1(t) + \varepsilon_1 emd_1\left(V^i(t)\right)\right) \tag{4}$$

(4) Repeat steps (1) and (2) above to calculate the kth residual signal and $k+1$ modal components for each of the remaining stages to obtain the remaining IMF modal components, where K is the total number of IMF modes.

$$R_k(t) = R_{k-1}(t) - IMF_k(t) \quad (k = 2, 3, \ldots K) \tag{5}$$

$$IMF_{(k+1)}(t) = \frac{1}{M}\sum_{i=1}^{M} emd_1\left(R_k(t) + \varepsilon_k emd_k\left(v^i(t)\right)\right) \tag{6}$$

(5) The termination condition of the decomposition is that the residual sequence cannot be decomposed further if there are at most 2 residual extreme value points. The remaining residual sequence that cannot be decomposed further is called the residual signal, and its expression is:

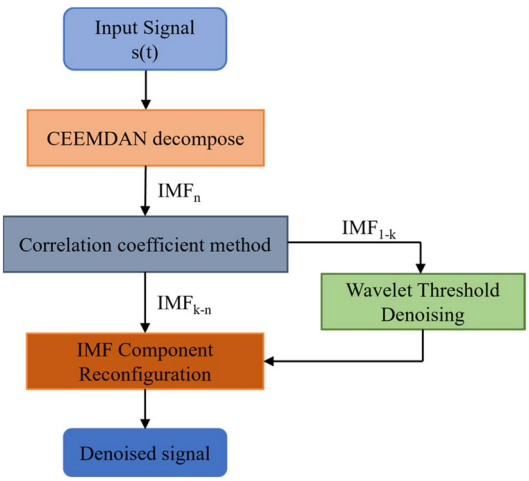


Figure 1. CEEMDAN operating principle.

$$r(t) = S(t) - \sum_{k=1}^{K} IMF_k(t) \tag{7}$$

(6) Therefore, the original time series $S(t)$ can be expressed by Eq. (8) after CEEMDAN decomposition:

$$S(n) = r(n) + \sum_{k=1}^{K} IMF_k(n) \tag{8}$$

Based on the above process we can see that the CEEMDAN decomposition is able to perform an accurate reconstruction of the original signal data. The problem of mode mixing in EMD decomposition is avoided in the whole implementation of the algorithm, and at the same time effectively reduces the number of iterations compared to the EEMD decomposition to increase the reconstruction accuracy and improve its computational efficiency, which is more suitable for the analysis of non-linear signals.

LSSVM. Least squares support vector machine (LSSVM) is a special kind of support vector machine, the basic idea is to map nonlinear data to linear regression in high-dimensional space[31], the main algorithm is to use least squares to transform the inequality constraints of the actual problem into a problem of solving a set of linear equations, which simplifies the calculation. With the research and development, it has been widely used in the field of hydrology, such as hydrological series prediction model, basin annual and monthly runoff prediction, etc., and has achieved better results.

The derivation process for the LSSVM algorithm is as follows:

(1) Given a training sample set of $\mathrm{P} = \left\{\left(x_k, y_k\right), k = 1, 2, \ldots N\right\}$, where $x_\mathrm{k} \in R^n$, $\mathrm{y_k} \in R$. The LSSVM algorithm non-linear regression function is:

$$f(x) = \omega^T \varphi(x) + b \tag{9}$$

where b is the deviation value; ω is the weight vector.

(2) By means of a non-linear transformation, the optimal hyperplane solution in higher dimensions can be transformed into the following form.

$$\begin{cases} \min J(\omega, b, \xi) = \frac{1}{2}\omega^T\omega + \frac{c}{2}\sum\limits_{i=1}^{N}\xi_k^2 \\ s.t. y_k = \omega^T\varphi(x_k) + b + e_k \end{cases} \tag{10}$$

where J is the loss function, c is the penalty factor, ξ_k is the error, and $\varphi(x_k)$ is a non-linear function.

(3) Constructing Lagrangian functions.:

$$L(\omega, b, e, \alpha) = J(\omega, b, e) - \sum_{k=1}^{N} \alpha_k \left\{\omega^T \phi(x_k) + b + \xi_k - y_k\right\} \tag{11}$$

(4) Then, according to the optimization condition, find the partial derivatives of ω, b, e, α respectively, and make the partial derivatives equal to 0. We have:

$$\begin{cases} \frac{\partial L}{\partial \omega} = 0 \Rightarrow \omega = \sum\limits_{i=1}^{n} \alpha_i y_i \varphi(x_i) \\ \frac{\partial L}{\partial b} = 0 \Rightarrow \sum\limits_{i=1}^{n} \alpha_i = 0 \\ \frac{\partial L}{\partial e_i} = 0 \Rightarrow \alpha_i = \gamma e_i \\ \frac{\partial L}{\partial \alpha_i} = 0 \Rightarrow \omega^T \varphi(x_i) + b + e_i - y_i = 0 \end{cases} \tag{12}$$

(5) By eliminating ω and ξ_i, the following linear system is obtained by simplification:

$$\begin{bmatrix} 0 & I^T \\ I & \Omega + I/C \end{bmatrix} \begin{bmatrix} b \\ a \end{bmatrix} = \begin{bmatrix} 0 \\ y \end{bmatrix} \tag{13}$$

where $y = \left[y_1, \ldots, y_n\right]^\mathrm{T}, \mathrm{I} = [1, \ldots, 1]^\mathrm{T}, a = [a_1, \ldots, a_n]^\mathrm{T}; \Omega_{M\times M}$ is the kernel matrix and the radial basis function RBF is the kernel function, then:

$$\Omega_{ij} = K\left(x_i, x_j\right) = \varphi(x_i)^T \varphi\left(x_j\right) \tag{14}$$

(6) After finding a and b from Eq. (13), the expression for the LSSVM non-linear regression function is obtained as

$$y = \omega^T \phi(x_k) + b = \sum_{k=1}^{N} a_k K(x, x_k) + b \tag{15}$$

Compared with SVM, LSSVM has smaller computational complexity and therefore faster processing speed; when dealing with dynamic problems, LSSVM can be extended into an autoregressive form; strong non-linear fitting ability, sparsity, generalization ability, good robustness and the ability to find optimal solutions quickly.

GM(1,1). The grey model (GM) is a system that includes partly unknown and partly known information[32], and grey theory is a mathematical method for solving systems with incomplete information by replacing the original random process with a grey process that has a temporal pattern and a limited range of variation, thus transforming the initial data, where no pattern can be found, into data that is easy to study with regular variation. The theory can therefore solve the problem of uncertainty in situations where there is insufficient information and too little data. The modelling steps of the model are as follows: Let $X^{(0)}$ be the original time series:

$$X^{(0)} = \left[X^{(0)}(1), X^{(0)}(2), \ldots, X^{(0)}(n)\right], \quad i = 1, 2, \ldots, n \tag{16}$$

(1) The original time series is generated cumulatively once to obtain the generated series $X^{(1)}$. This weakens the randomness and enhances the regularity, and the generated series will be close to the regularity of the exponential relationship.

$$X^{(1)} = \left[X^{(1)}(1), X^{(1)}(2), \ldots, X^{(1)}(n)\right] \tag{17}$$

where

$$X^{(1)}(k) = \sum_{i=1}^{k} X^{(0)}(i), k = 1, 2, \ldots n \tag{18}$$

(2) The cumulative sequence $X^{(1)}$ is generated by making the nearest neighbor mean according to Eq. (18) to obtain the sequence $Z^{(1)}$

$$Z^{(1)}(k) = \frac{1}{2}X^{(1)}(k) + \frac{1}{2}X^{(1)}(k-1), k = 1, 2, \ldots n \tag{19}$$

$$Z^{(1)} = \left[Z^{(1)}(1), Z^{(1)}(2), \ldots Z^{(1)}(n)\right] \tag{20}$$

(3) The differential equation for GM(1,1) is established from the cumulative generating sequence $X^{(1)}$.

$$\begin{cases} \frac{dX^{(1)}}{dk} + aX^{(1)} = b \\ X^{(1)}(1) = X^{(0)}(1) \end{cases} \tag{21}$$

where a is the development coefficient and b is the amount of grey action. From the solution method in the theory of ordinary differential equations, the analytical solution of the equation is found as:

$$\hat{X}^{(1)}(k+1) = \left[X^{(0)}(1) - \frac{b}{a}\right]e^{-ak} + \frac{b}{a}, \quad (k = 1, 2, \ldots, n) \tag{22}$$

(4) Find the values of a, b by the principle of least squares.

$$\overline{a} = (a, b)^T = \left(B^T B\right)^{-1} B^T Y \tag{23}$$

where

$$Y = \begin{bmatrix} X^{(0)}(2) \\ X^{(0)}(3) \\ \vdots \\ X^{(0)}(n) \end{bmatrix} \tag{24}$$

$$B = \begin{bmatrix} -Z^{(1)}(2) & 1 \\ -Z^{(1)}(3) & 1 \\ \vdots & \vdots \\ -Z^{(1)}(n) & 1 \end{bmatrix} \tag{25}$$

(5) Reducing Eq. (22) to a grey forecasting model for the original time series $X^{(0)}$.

$$\hat{X}^{(0)}(k+1) = \left(1 - e^{a}\right)\left[X^{(0)}(1) - \frac{b}{a}\right]e^{-ak}, (k = 1, 2, \ldots, n) \tag{26}$$

The grey model is a differential equation model created by generating new data from the original data series. Since the solution of its differential equation is in exponential form, it is more accurate in predicting variables with exponential growth or decreasing trends. Although the model is suitable for modelling sequences that are monotonic and vary exponentially, the non-linear and non-smooth nature of the monthly runoff series makes it so that the generated series after the accumulation of the original runoff series does not necessarily obey the exponential variation law, so this section is based on the residuals obtained from the CEEMDAN decomposition for forecasting, which is feasible with this model as the residuals has a monotonic downward trend.

Model development. The partitioning of time series datasets is often judged according to the rule of thumb; normally, the training part of the dataset should carry more than 60% of the overall and the validation part should be more than 20% of the overall, and many researchers have used different partitioning situations: Kumar et al.[33] used 70% of the data to train RNN and LSTM models for the "all-India" monthly average precipitation data to build the model; Liu et al.[34] used 80% of the data as the training set to train the model for wind speed prediction and the remaining 20% portion as the test set. We prevent overfitting by increasing the training data of the model. In this study, 90% of the data were taken to train the model and 10% of the data were used to test the model performance.

CEEMDAN–LSSVM–GM(1,1) model. In this paper, the monthly runoff data of four hydrological stations in the lower reaches of the Yellow River: Huayuankou Station, Gaocun Station, Aishan Station and Lijin Station from 1965 to 2014 were selected for the study, and the runoff characteristics and change patterns were analyzed and examined, and on this basis, a runoff prediction model was established using least squares support vector machine and grey theory, and the flow chart is shown in Fig. 2.

1. Monthly historical runoff data of four hydrological stations in the lower Yellow River from 1965 to 2014 were decomposed using CEEMDAN to obtain several high-frequency IMF components and a low-frequency residual.
2. A total of 550 months of data from January 1965 to October 2010 were used for training the model, and 50 months of data from November 2010 to December 2014 were used for validation.
3. The high-frequency IMF components obtained from the CEEMDAN decomposition are predicted using the LSSVM model, and then the low-frequency residuals are predicted using the GM(1,1) model.
4. Weighted summation of the predictions of the IMF components and residuals calculated in step (3) to obtain the final monthly runoff time series prediction results.

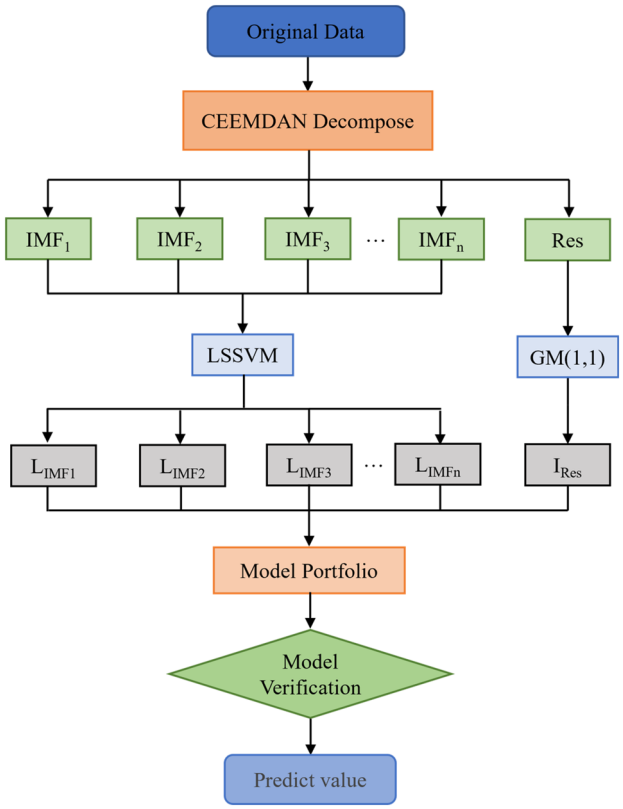


Figure 2. Flow chart of CEEMDAN–LSSVM–GM(1,1) model.

Model evaluation indicators. In order to verify the prediction results of the CEEMDAN–LSSVM–GM(1,1) model for the monthly runoff at four hydrological stations in the lower reaches of the Yellow River, the following four evaluation indicators were used to evaluate the prediction results, with Nash–Sutcliffe efficiency coefficient (NSE), mean absolute error (MAE), mean absolute percentage error (MAPE), and root mean squared error (RMSE) as quantitative evaluation criteria, and the calculation equations are as follows:

$$NSE = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}, i = 1, 2, \ldots, n \tag{27}$$

$$MAE = \frac{1}{N} \sum_{i=1}^{n} \left| (y_i - \hat{y}_i) \right|, \quad i = 1, 2, \ldots, n \tag{28}$$

$$MAPE = \frac{1}{N} \sum_{i=1}^{N} \left| \frac{y_i - \hat{y}_i}{y_i} \right| \times 100\%, \quad i = 1, 2, \ldots, n \tag{29}$$

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2}, \quad i = 1, 2, \ldots, n \tag{30}$$

where $\hat{y}_i$ and y_i represent the predicted runoff series and its corresponding original runoff series respectively, $\bar{y}$ is the monthly average runoff volume of the original runoff series, and n is the number of runoff series values.

Case study

Study area. The Yellow River basin is located in the mid-latitude zone, with a range of 95° 53′–119° 05′ E, 32° 10′– 42° 50′ N. The Yellow River originates from the Bayankara Mountains on the Qinghai-Tibet Plateau in Qinghai Province, and spans a large area from east to west, flowing from west to east through Qinghai, Sichuan, Gansu, Ningxia, Inner Mongolia, Shaanxi, Shanxi, Henan and Shandong. The Yellow River has a total length of 5464 km, with an east–west direction of about 1900 km and a north–south direction of about 1100 km, covering a total area of 752,000 km^2 and a large geographical span. Three-quarters of the Yellow River basin is in the arid and semi-arid zone, with a predominantly continental monsoon climate. The average annual evaporation in the Yellow River basin ranges from 700 to 1800 mm, with high evaporation. There are many tributaries in the Yellow River basin and the spatial and temporal distribution of runoff is uneven and seasonally variable[35]. The Huayuankou hydrological station is located in Zhengzhou City, Henan Province, while the Gaocun hydrological station, Aishan hydrological station and Lijin hydrological station are located in Heze City, Liaocheng City and Dongying City, Shandong Province respectively. These four hydrological stations are responsible for the important tasks of water resources utilisation in the lower reaches of the Yellow River, regional water resources development and the investigation of hydrological and water resources change patterns, and have well preserved hydrological data. Runoff data from all hydrological stations in the study area are available in the Water Information System of the Yellow River Network (www.yrcc.gov.cn). The geographical location map of the study area is shown in Fig. 3, it is created using ArcMap 10.2, URL:www.arcgis.com, and the runoff series from 1965 to 2014 for the four hydrological stations are shown in Fig. 4.

Data sources. Linear regression analysis and moving average were selected to analyze the trend of monthly runoff at the four hydrological stations from 1965 to 2014. Figure 4 shows the process of runoff at each station. It can be seen that the monthly runoff at the four hydrological stations in the lower reaches of the Yellow River is highly non-linear and non-stationary, with the extreme values of monthly runoff occurring mostly during the flood season, and showing a high degree of time-variability and complexity. The magnitude and variability of monthly runoff at the Huayuankou hydrological station is higher than at the other three stations due to its special geographical location. The linear trend line and the periodic moving average curve in the graph represent the trend of monthly runoff. This indicates that the runoff shortage problem in the lower Yellow River has become increasingly serious, and that more accurate prediction models are needed to provide reliable and stable predictions of runoff.

Mutation detection of hydrological data helps in hydrological data analysis and hydrological data prediction. Reliable mutation detection can analyze the stage change characteristics of hydrological data and find out the factors that affect the hydrological prediction effect. It is of great guiding significance for in-depth analysis of hydrological data change characteristics and improvement of hydrological data prediction effect[36]. Given a significant level $\alpha = 0.05$, the critical value $u_{0.05} = \pm 1.96$, and the results of the Mann–Kendall test were obtained as shown in Fig. 5. The UF values corresponding to the intersection of the UF and UB statistics at the four hydrological stations are all less than 0, indicating that these intersection points are the abrupt change points for the reduction of monthly mean runoff, with the intersection of the UF and UB statistics at the Huayuankou hydrological station lying within the critical interval, indicating a significant abrupt change at this point, and the intersection of the UF and UB statistics at the remaining three stations all lying outside the critical interval, indicating that no significant abrupt change in runoff has occurred at these stations.

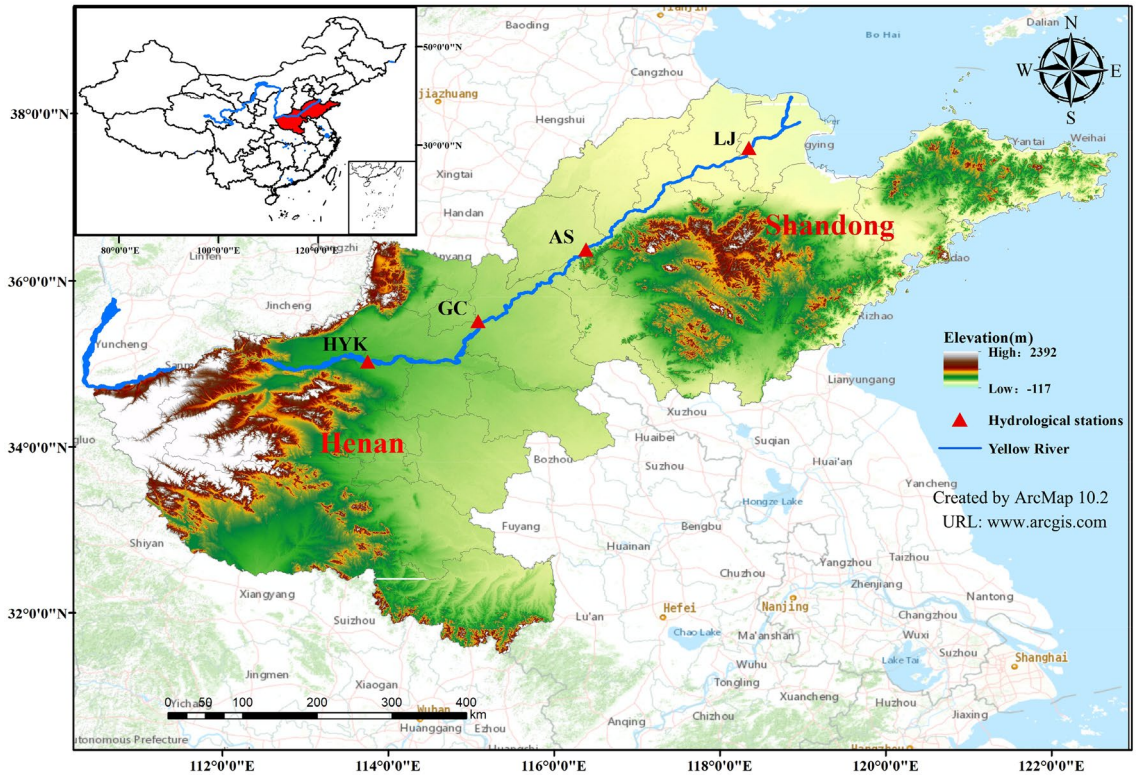
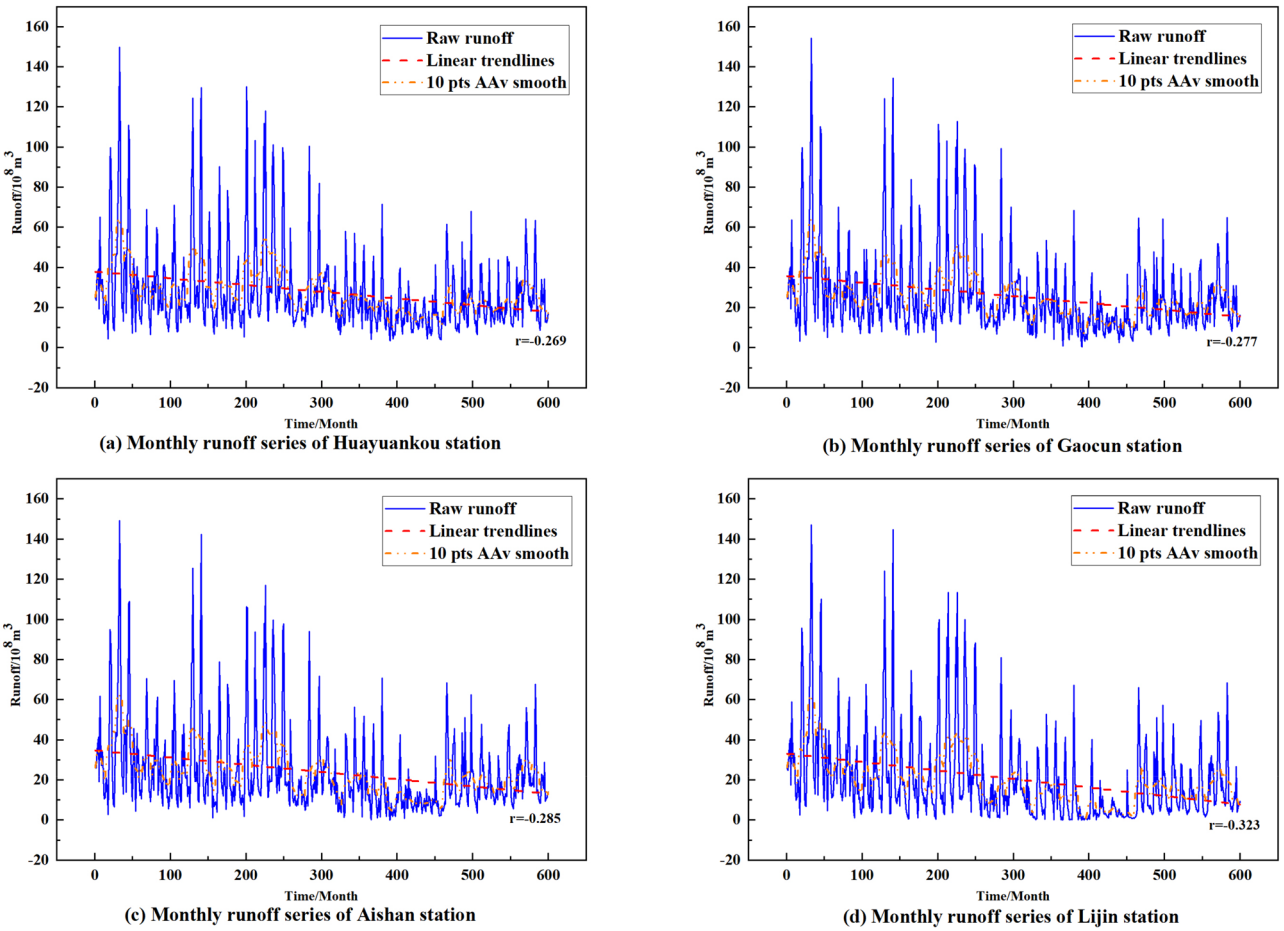


Figure 3. Hydrographic station distribution map.



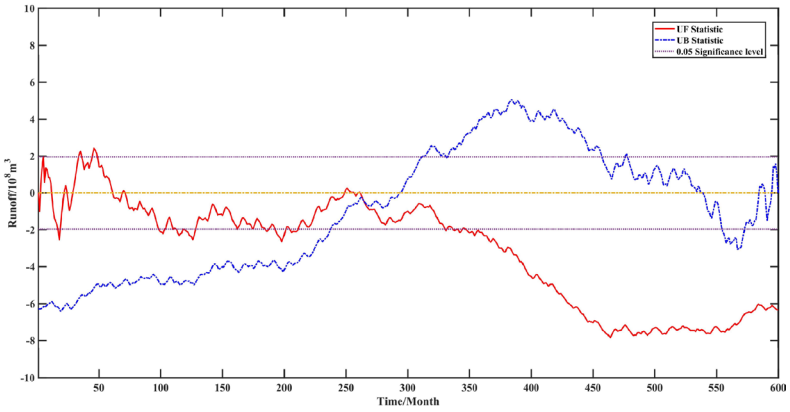
(a) Monthly runoff series of Huayuankou station

(b) Monthly runoff series of Gaocun station

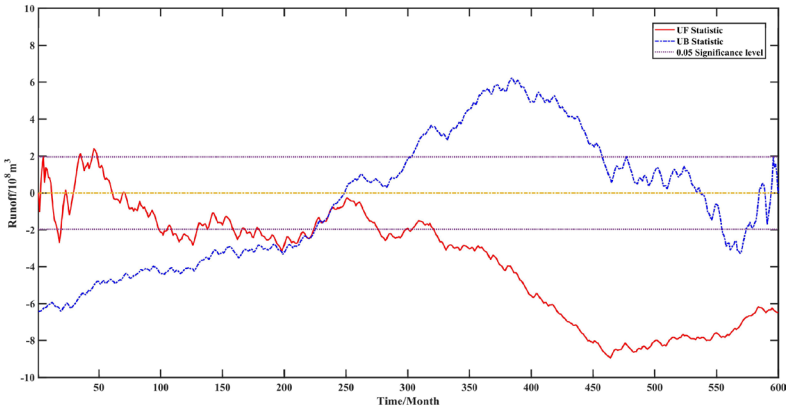
(c) Monthly runoff series of Aishan station

(d) Monthly runoff series of Lijin station

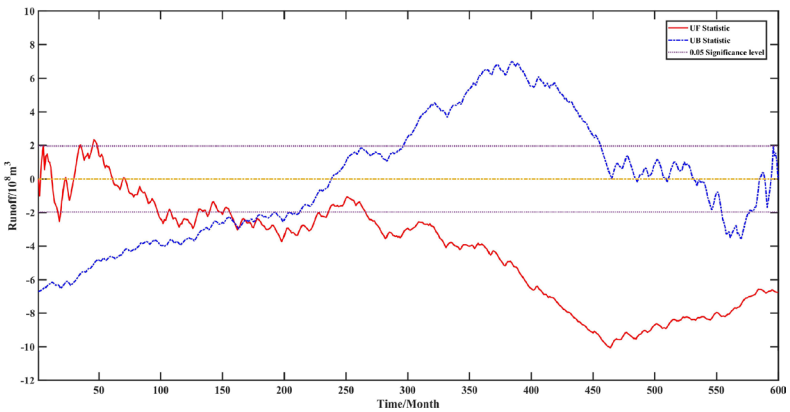
Figure 4. The course of monthly runoff at each hydrological station 1965–2014.



(a)Huayunkou station

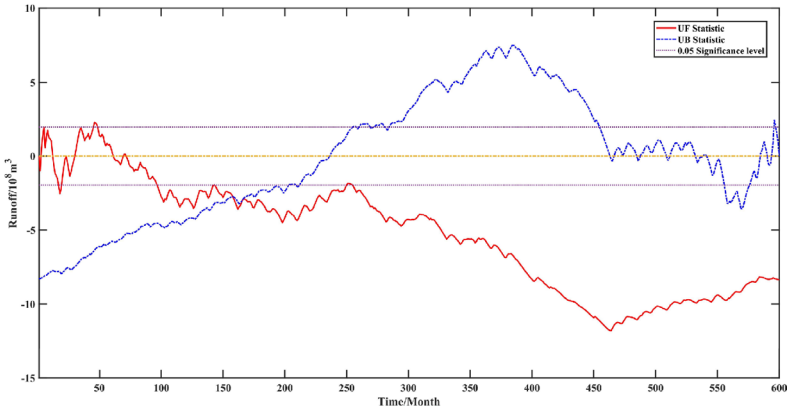


(b)Gaocun station



(c)Aishan station

Figure 5. Mann–Kendall test for runoff from January 1965 to December 2014 at each station.



(d)Lijin station

Figure 5. (continued)

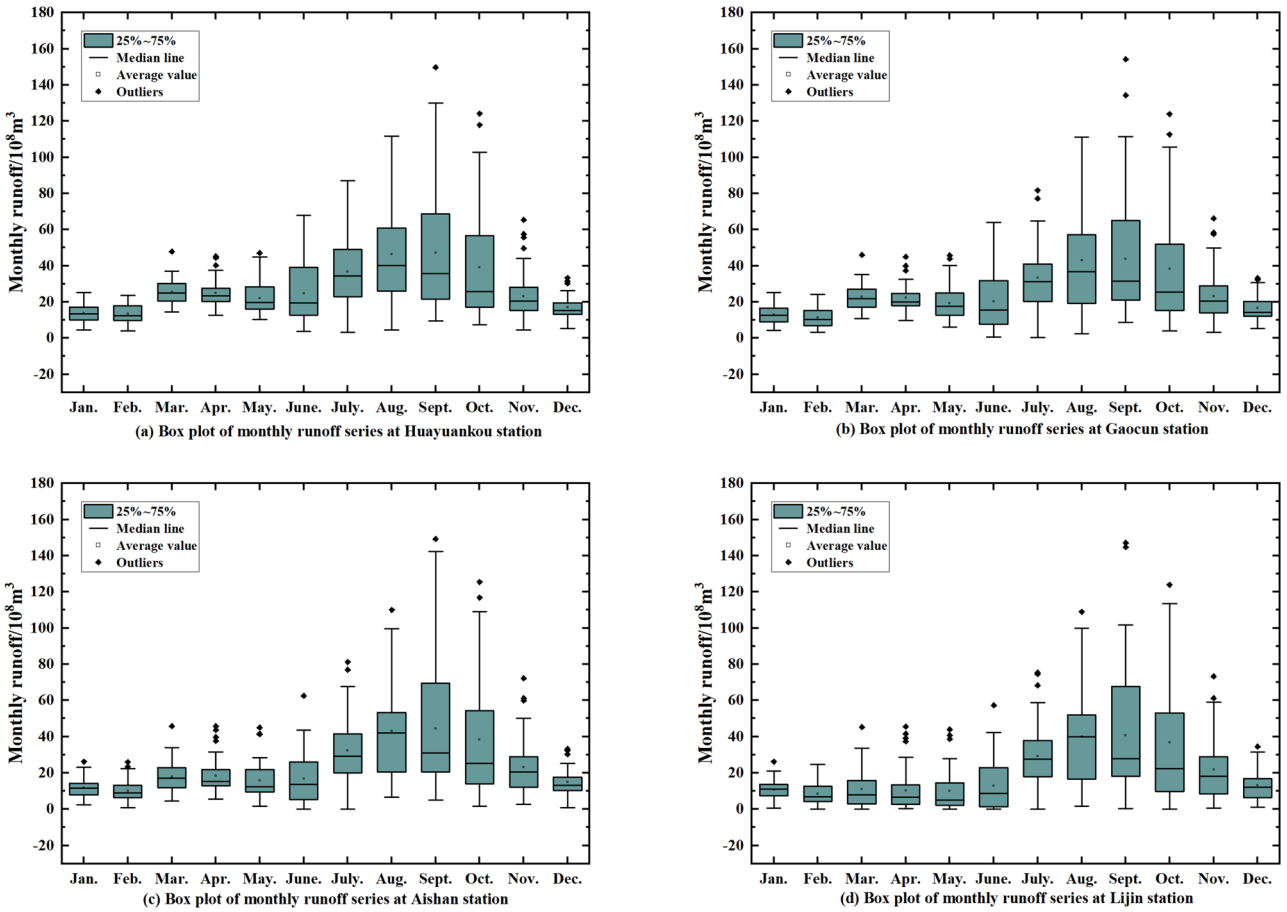


Figure 6. Yellow River downstream runoff box diagram.

In this paper, a total of 600 months of hydrological data from January 1965 to December 2014 were collected from four hydrological stations in the lower Yellow River, of which 550 months of runoff data from January 1965 to December 2009 were used as the training set data, and 50 months of runoff data from January 2010 to December 2014 were used as the test set to validate the model effect. Box plots were used to detect anomalies in the runoff data. Figure 6 shows the results of the box plot identification of the monthly runoff data of the lower Yellow River over the years. It can be seen from the figure that the box corresponding to June–October in the lower Yellow River is longer, which indicates that the runoff volume fluctuates more drastically in these months, among which, there are more anomalous values in July. The reason for the above phenomenon may be that this period is the flood season of the Yellow River basin, which is greatly influenced by rainfall as well

as certain extreme climatic factors, and the runoff volume accumulates in a short period of time, showing an irregular change in magnitude. Therefore, normalisation of the decomposed monthly runoff data is considered to reduce the volatility of the raw runoff data and enhance the stability of the model prediction. The normalisation formula is as follows:

$$y_i' = \frac{y_i - \min(y_i)}{\max(y_i) - \min(y_i)} \tag{31}$$

Results and discussion

Results. The results obtained after decomposing the monthly runoff data from the four hydrological stations using CEEMDAN are shown in Fig. 7. It can be seen that the IMF1 frequency is the highest, the amplitude is the largest, and the wavelength is the shortest, while the periodicity is the smallest at the same time, at the Huayuankou station. IMF1 to the rest of the frequency gradually decreases, the amplitude decreases and the periodicity becomes stronger. The stability of IMF2–IMF7 gradually increases, representing different time-scale components of the original runoff, while retaining some periodic features and some trends of the original runoff. The amplitude of IMF1 at Gaocun station is the largest, and IMF2–IMF7 gradually stabilize. The amplitude of IMF1 at Aishan station is also the largest, and IMF2–IMF7 fluctuates more from 1965 to 1982 and gradually decreases from 1983 to 2014. The frequency of IMF1 is the highest in Lijin station. While the fluctuations of IMF2–IMF7 from 1965 to 1990 are more obvious, the fluctuations from 1991 to 2014 are more stable. Res is the residual of the original runoff, which represents the overall trend of the original runoff series and is an important criterion to judge the change pattern of runoff. From the change curve of Res in Fig. 7, it can be seen that the runoff from Huayuankou and Gaocun stations gradually increased from 1965 to 1982 and decreased from 1983 to 2014; the runoff from Aishan and Lijin stations gradually increased from 1965 to 1974 and decreased from 1975 to 2014. The above observations show that the general trend of runoff in the lower Yellow River area has been increasing and then decreasing in the last 60 years.

After preprocessing the raw runoff time series, the LSSVM and GM(1,1) models were used to predict the high-frequency IMF components and low-frequency trend terms, respectively. The parameters of the LSSVM model are set as follows: the radial basis function (RBF) is chosen as the kernel function, and its expression is:$K(x_i, x_j) = \exp\left(-g \parallel x_i - x_j \parallel^2\right)$, The size of the penalty factor c and the kernel function parameter g affect the prediction accuracy of the model, and the LSSVM parameters are optimized by the grid search method, i.e., the search range is set for the penalty factor c and the kernel function parameter g respectively, and the parameters are optimized within the specified interval, and the penalty factor c is finally determined to be 150 and the kernel function parameter g is 1.5. The parameters of the GM(1,1) model are set as follows: development The coefficient a is taken as 0.3, and the gray action quantity b is taken as 3.2.

LSSVM was used to simulate the prediction of IMF1–IMF7 data from four hydrological stations obtained by CEEMDAN decomposition, and GM(1,1) was also used to simulate the prediction of Res obtained by CEEMDAN decomposition, and the prediction results were summed to obtain the monthly runoff prediction data of four hydrological stations, in which 550 months of data from January 1965 to October 2010 were The data from November 2010 to December 2014 were used for training and validation, and the monthly runoff predictions for the four hydrological stations in the lower Yellow River are shown in Fig. 8.

In order to verify the finiteness, accuracy and robustness of the CEEMDAN–LSSVM–GM(1,1) model for monthly runoff prediction, the prediction results of the LSSVM, CEEMDAN–LSSVM, CEEMDAN–SVM–GM(1,1) and CEEMDAN–LSSVM–GM(1,1) models were compared and analyzed in this paper, As shown in Fig. 9, the prediction accuracy of CEEMDAN–LSSVM–GM(1,1) model is the highest, the prediction accuracy of LSSVM model is poor, and the prediction results of the other two models are basically consistent with the original data.

Discussion. The Nash–Sutcliffe efficiency coefficient (NSE), mean absolute error (MAE), mean absolute percentage error (MAPE), root mean squared error (RMSE) were calculated based on the prediction results of the four models. The results are shown in Table 1. The CEEMDAN–LSSVM–GM(1,1) model has the highest prediction accuracy, in which the NSE of Huayuankou station is 0.9521, MAE is 2.807×10^7 m^3, MAPE is 1.20%, RMSE is 3.348×10^7 m^3, NSE of Gaocun station is 0.9345, MAE is 5.434×10^7 m^3, MAPE is 2.02%, RMSE is 5.638×10^7 m^3, NSE of Aishan station is 0.9334, MAE is 6.221×10^7 m^3, MAPE is 1.60%, RMSE is 4.874×10^7 m^3, NSE of Lijin station is 0.9214, MAE is 7.442×10^7 m^3, MAPE is 1.53%, RMSE is 4.687×10^7 m^3. The Nash efficiency coefficients are all above 0.9, the mean absolute error $\leq 7.442 \times 10^7$ m^3 and the average absolute percentage error is around 2%. Zhang et al. (2020) used modified ensemble empirical mode decomposition (MEEMD)-autoregressive integrated moving average (ARIMA), to predict the runoff from 2010 to 2014 at Huayuankou hydrological station with a relative error of, and an average absolute percentage error of 6.04%[37]. Zhang et al.[38] used the CEEMDAN-autoregressive moving average (ARMA) model to predict the runoff from 1960 to 2017 at Tang Naihai hydrological station in the Yellow River source area, and the prediction results showed a Nash efficiency coefficient of 0.786 and an average absolute percentage error of 8.78%, indicating that the CEEMDAN–LSSVM–GM(1,1) model has high prediction accuracy and good quality, and the credibility of the model is high.

The evaluation metrics of the prediction results of the four models are shown in Fig. 10. Compared to the LSSVM model, the prediction results of the CEEMDAN–LSSVM–GM(1,1) model showed an improvement of 15.68% in NSE, 87.94% in MAE, 90.87% in MAPE and 92.12% in RMSE; compared to the CEEMDAN–LSSVM

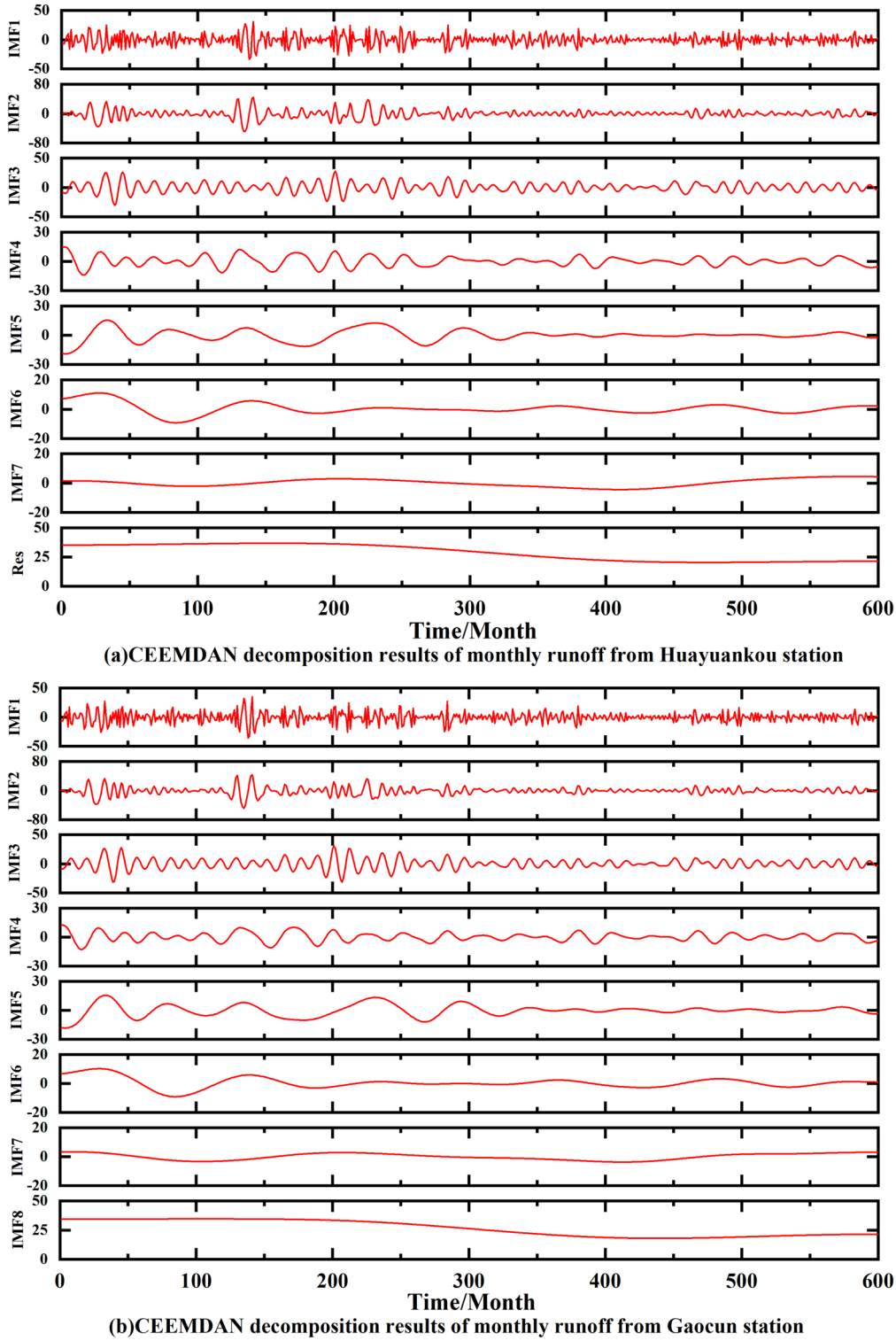
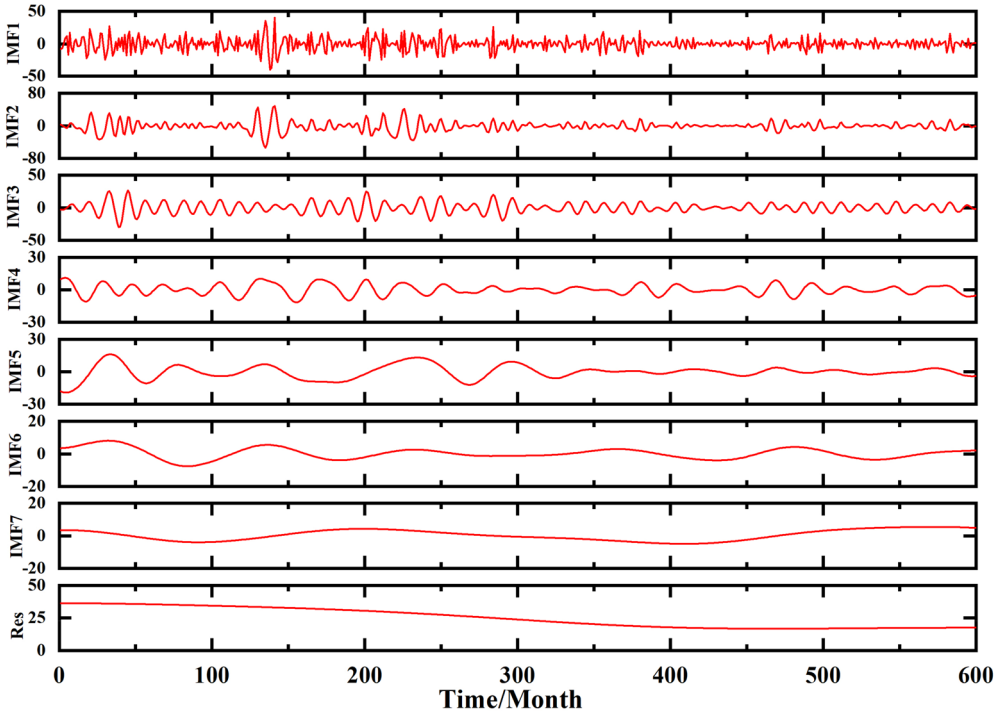


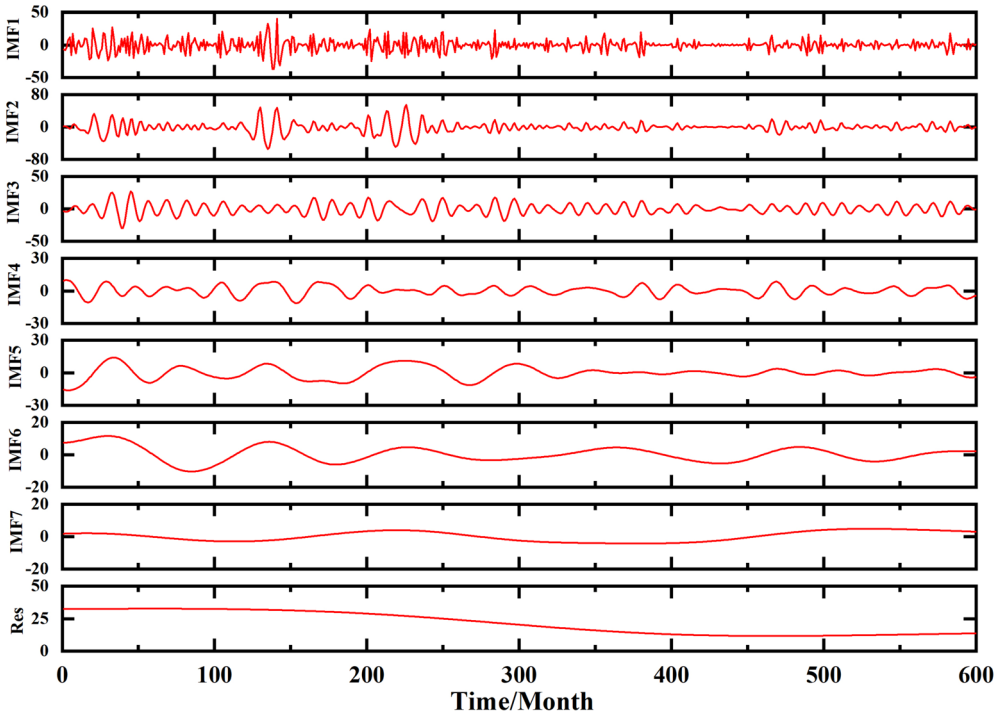
Figure 7. Monthly runoff data from the lower Yellow River stations using CEEMDAN decomposition.

model, NSE improved by 7.84%, MAE reduced by 78.97%, MAPE reduced by 85.34% and RMSE reduced by 85.78%; compared to CEEMDAN–SVM–GM(1,1) model, NSE improved by 2.81%, MAE reduced by 56.32%, MAPE reduced by 71.04% and RMSE by 69.71%.

Figure 11 shows the scatter plots of predicted versus observed values for each model at each station in the lower Yellow River and the corresponding linear trend lines for the scatter plots. The CEEMDAN–LSSVM–GM(1,1) model has a linear trend line closest to $y = x$ and therefore has optimal runoff simulation and prediction capabilities.



(c)CEEMDAN decomposition results of monthly runoff from Aishan station



(d)CEEMDAN decomposition results of monthly runoff from Lijin station

Figure 7. (continued)

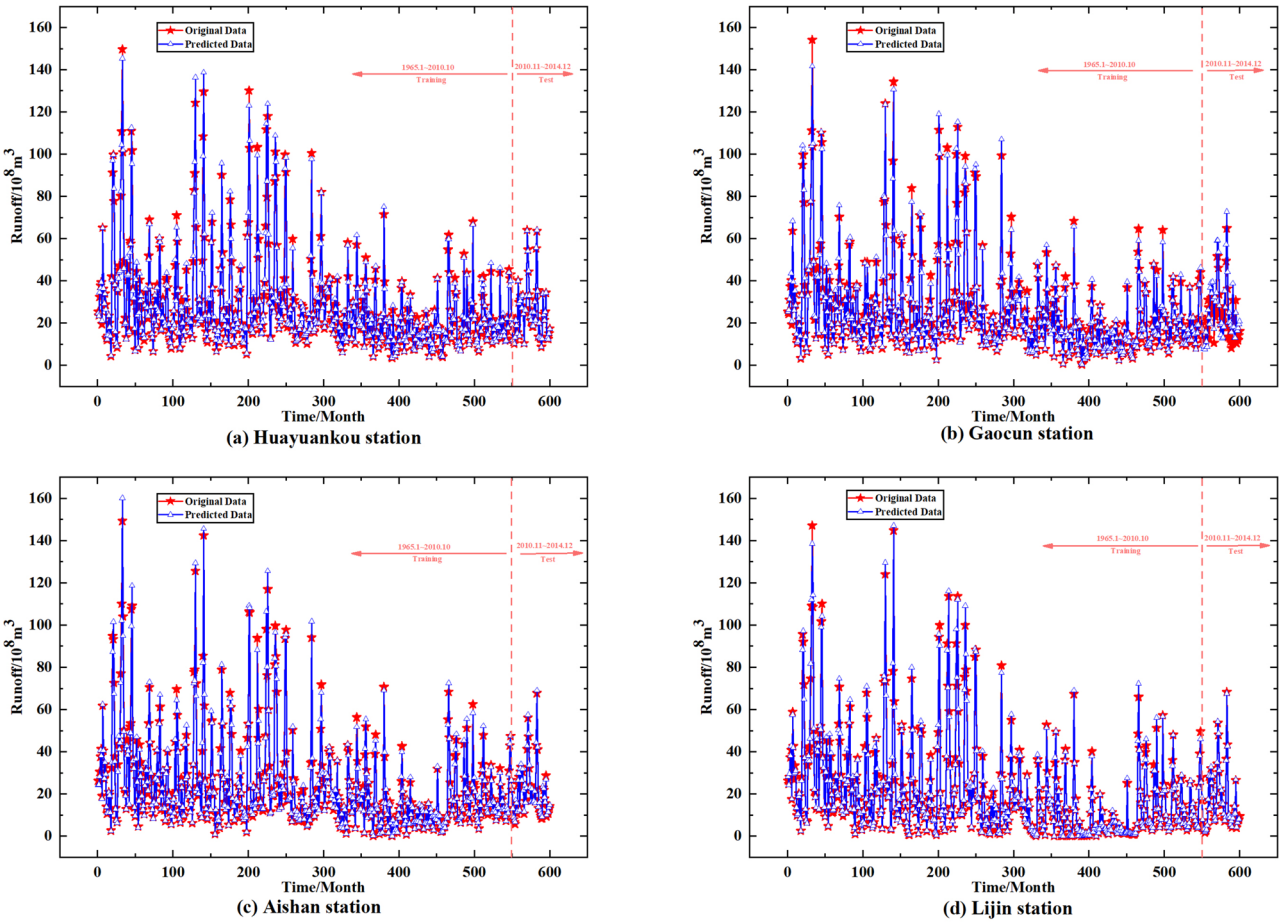


Figure 8. CEEMDAN–LSSVM–GM(1,1) model predicts monthly runoff results for stations under the Yellow River.

Conclusion

1. The results of the monthly runoff prediction for four hydrological stations in the lower Yellow River show that the CEEMDAN–LSSVM–GM(1,1) model proposed in this paper has good accuracy and robustness. The highest prediction accuracy is at Huayuankou station, the highest Nash efficiency coefficient is at Huayuankou station with 0.9521, MAE is 2.807×10^7 m^3, MAPE is 1.20%, RMSE is 3.348×10^7 m^3, the lowest Nash efficiency coefficient is at Lijin station with 0.9214, the highest MAE is at Lijin station with 7.442×10^7 m^3, MAPE The largest is Gaocun station with 2.02% and the largest RMSE is Gaocun station with 5.638×10^7 m^3. Its prediction accuracy is higher than that of the LSSVM model, the CEEMDAN–LSSVM model, and the CEEMDAN–SVM–GM(1,1) model. This indicates that the CEEMDAN–LSSVM–GM(1,1) model is feasible for monthly runoff prediction and can be effectively used for time series analysis in hydrology and related fields to guide the rational development and improve the utilization of water resources.
2. The CEEMDAN–LSSVM–GM(1,1) model proposed in this paper can reduce prediction errors, improve data fitting ability and model stability to a large extent through data pre-processing-decomposition-noise reduction-prediction, and can be used as one of the means to enrich and improve the decomposition of medium and long-term runoff prediction.
3. Although the CEEMDAN–LSSVM–GM(1,1) model has a promising application with its effective decomposition algorithm and stable and fast prediction capability. Due to the problem of the model, the lag brought by physical mechanisms such as precipitation on runoff cannot be considered, and the input can only be runoff time series, which is a shortcoming of the model and a focus of further research in the future.

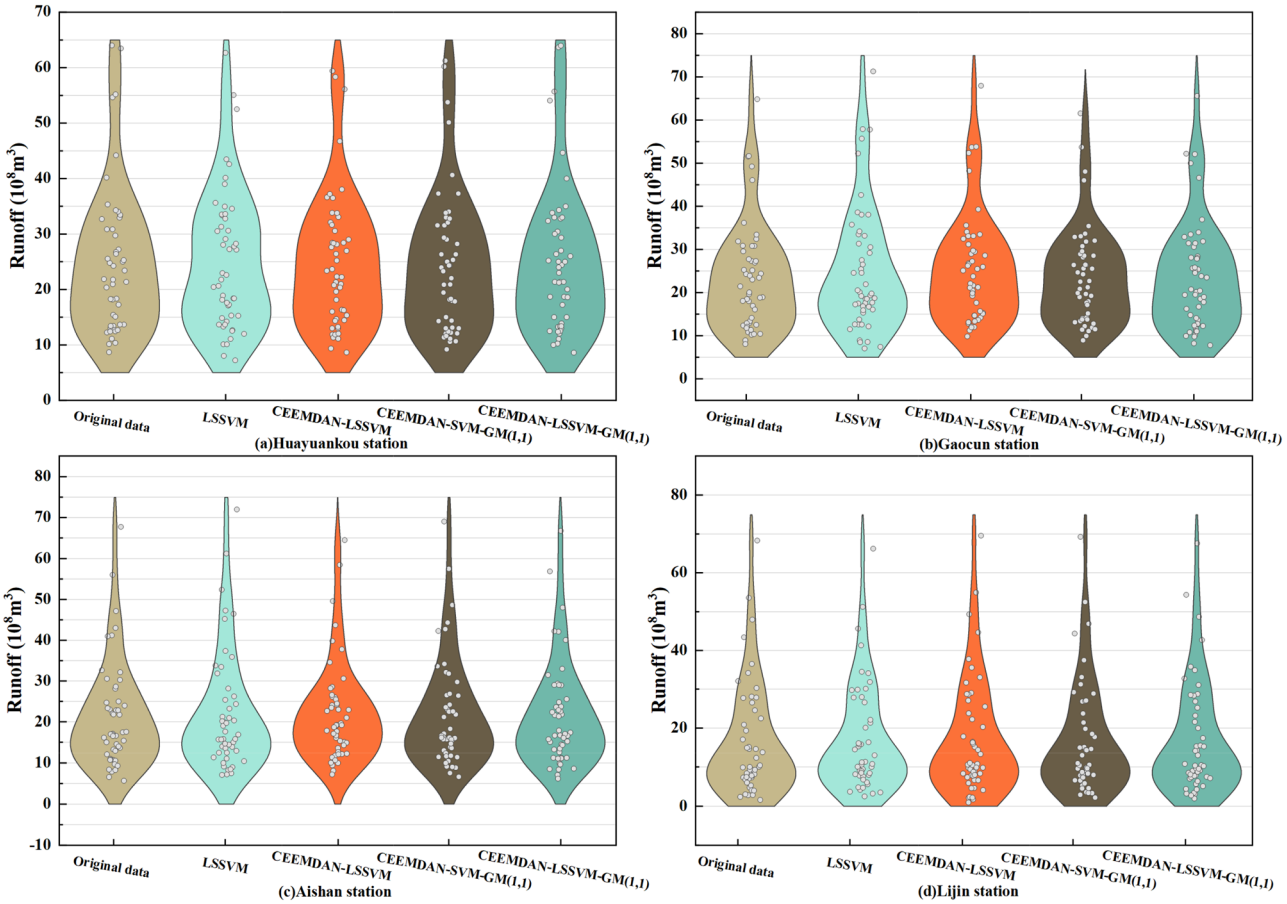


Figure 9. Comparison of the prediction results of the four models.

Station	Model	NSE	MAE/10^8 m^3	MAPE/%	RMSE/10^8 m^3
Huayuankou	LSSVM	0.8194	4.2318	17.63	5.6552
	CEEMDAN–LSSVM	0.8825	2.6074	10.67	3.2780
	CEEMDAN–SVM–GM(1,1)	0.9206	1.3158	5.40	1.6254
	CEEMDAN–LSSVM–GM(1,1)	0.9521	0.2807	1.20	0.3348
Gaocun	LSSVM	0.7936	4.5262	18.36	5.9261
	CEEMDAN–LSSVM	0.8554	2.0248	10.82	3.0628
	CEEMDAN–SVM–GM(1,1)	0.9026	1.0531	5.06	1.5228
	CEEMDAN–LSSVM–GM(1,1)	0.9345	0.5434	2.02	0.5638
Aishan	LSSVM	0.7832	4.3806	16.75	6.1321
	CEEMDAN–LSSVM	0.8467	2.5610	10.51	3.2416
	CEEMDAN–SVM–GM(1,1)	0.9143	1.2525	5.24	1.4852
	CEEMDAN–LSSVM–GM(1,1)	0.9334	0.6221	1.60	0.4874
Lijin	LSSVM	0.8382	5.0214	16.84	5.8113
	CEEMDAN–LSSVM	0.8847	3.2225	11.32	3.4562
	CEEMDAN–SVM–GM(1,1)	0.9016	1.3935	6.23	1.4895
	CEEMDAN–LSSVM–GM(1,1)	0.9214	0.7442	1.53	0.4687

Table 1. Evaluation of the prediction results of each model.

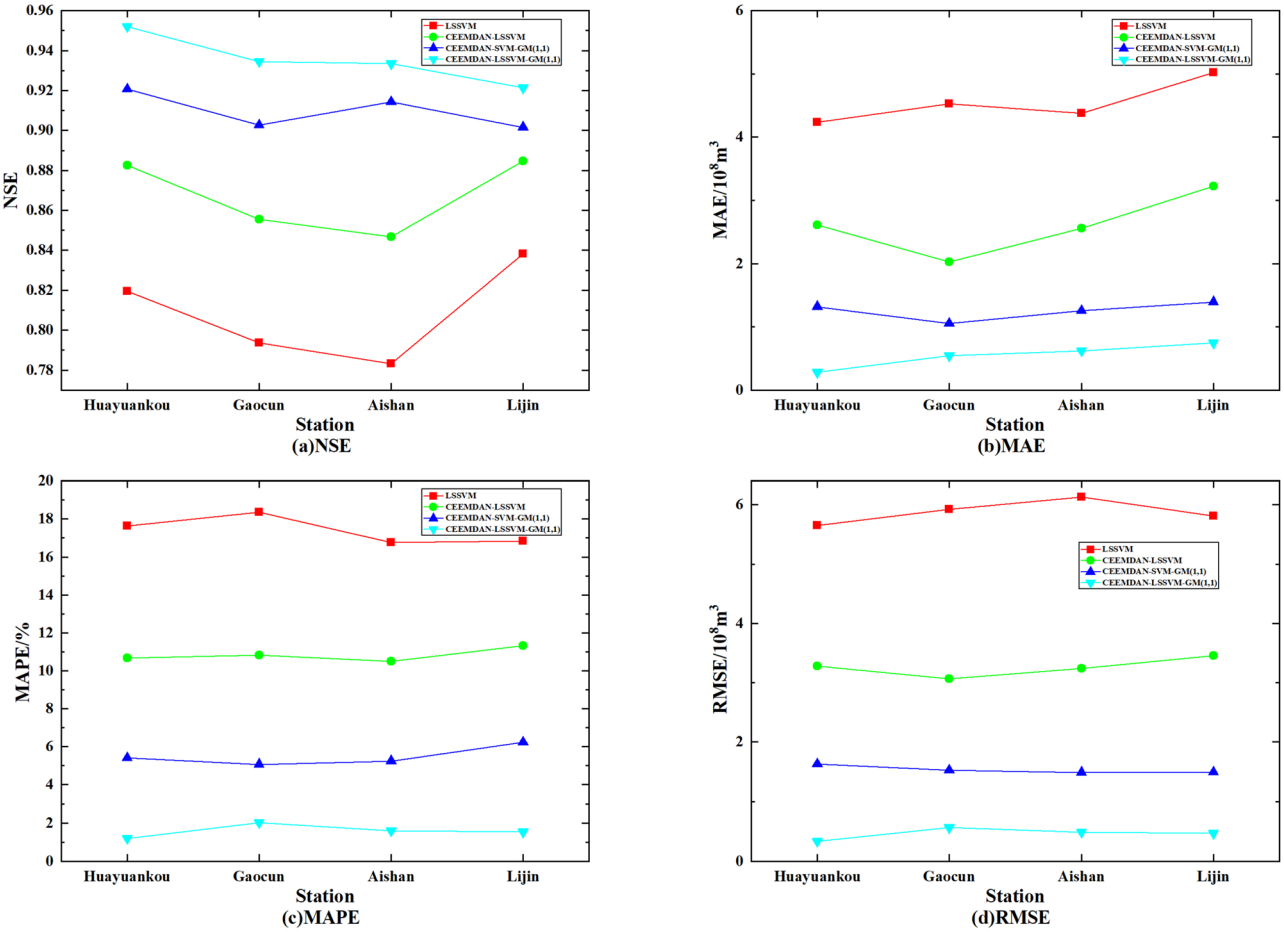


Figure 10. Comparison of four model evaluation indicators.

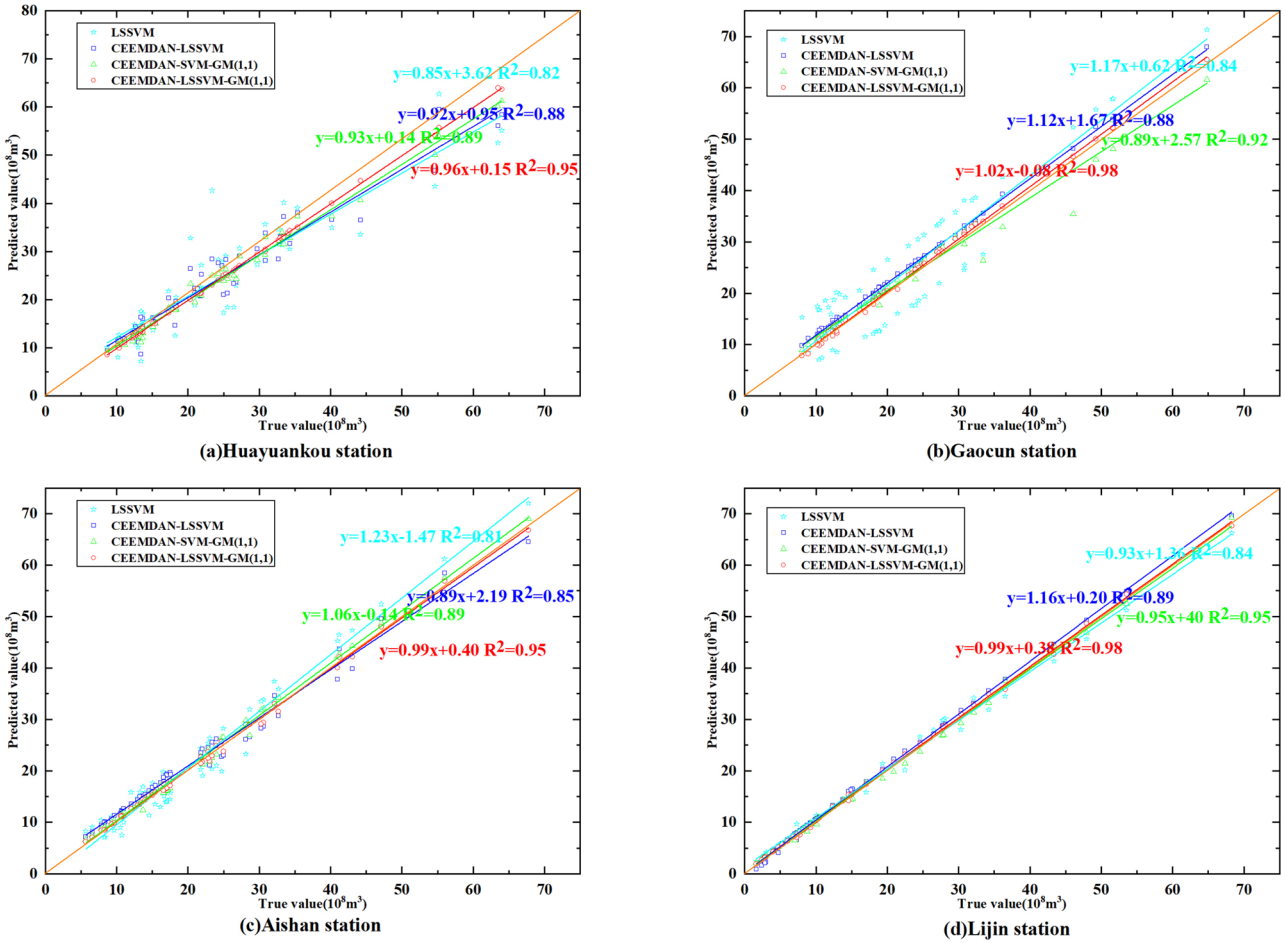


Figure 11. Scatter chart of four model prediction results.

Data availability

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

Received: 9 November 2022; Accepted: 23 January 2023
Published online: 27 January 2023

References

1. Jin, H., Chen, X. & Zhong, R. Runoff forecast and analysis of the probability of dry and wet transition in the Hanjiang River Basin. *Stoch. Env. Res. Risk Assess.* **36**(5), 1485–1502 (2022).
2. Tang, G. L., Zhou, H. C., Li, N. N. & Wang, Y. J. An optimal reservoir scheduling model considering runoff forecasting and its uncertainty. *J. Water Res.* **42**(6), 641–647 (2011).
3. Song, P. *et al.* Annual runoff forecasting based on multi-model information fusion and residual error correction in the Ganjiang River Basin. *Water* **12**(8), 2086 (2020).
4. Tan, Q. F., Wang, X., Wang, H. & Lei, X. H. A comparison of ANN, ANFIS and AR models for daily runoff time series prediction. *South North Water Divers. Water Resour. Sci. Technol.* **14**(6), 12–17 (2016).
5. Sun, W. *et al.* Hybrid short-term runoff prediction model based on optimal variational mode decomposition, improved Harris hawks algorithm and long short-term memory network. *Environ. Res. Commun.* **4**(4), 045001 (2022).
6. Srinivasan, R., Ramanarayanan, T. S., Arnold, J. G. & Bednarz, S. T. Large area hydrologic modeling and assessment part II: Model application 1. *J. Am. Water Resour. Assoc.* **34**(1), 91–101 (1998).
7. Liang, X., Lettenmaier, D. P., Wood, E. F. & Burges, S. J. A simple hydrologically based model of land surface water and energy fluxes for general circulation models. *J. Geophys. Res. Atmos.* **99**(D7), 14415–14428 (1994).
8. Pan, Y. W., Zhang, H. N., Xia, D. Z. & Shi, C. A comparative study of Xin'an River model and DHSVM in small and medium-sized basins. *Hydropower* **41**(4), 15–18 (2015).
9. Wang, W. C., Chau, K. W., Cheng, C. T. & Qiu, L. A comparison of performance of several artificial intelligence methods for forecasting monthly discharge time series. *J. Hydrol.* **374**(3–4), 294–306 (2009).
10. Piechota, T. C., Chiew, F. H., Dracup, J. A. & McMahon, T. A. Seasonal streamflow forecasting in eastern Australia and the El Niño-Southern Oscillation. *Water Resour. Res.* **34**(11), 3035–3044 (1998).
11. Han, R., Zengchuan, D., Xuewei, W. & Hongliang, Ma. Application of weighted average integration model in runoff prediction. *People's Yellow River* **39**(6), 16–20 (2017).
12. Thomas Harold, A. Mathematical synthesis of streamflow sequences for the analysis of river basin by simulation. *Des. Water Resour. Syst.* 459–493 (1962).

13. Carlson, R. F., MacCormick, A. J. A. & Watts, D. G. Application of linear random models to four annual streamflow series. *Water Resour. Res.* **6**(4), 1070–1078 (1970).
14. Elshorbagy, A., Corzo, G., Srinivasulu, S. & Solomatine, D. P. Experimental investigation of the predictive capabilities of data driven modeling techniques in hydrology—Part 1: Concepts and methodology. *Hydrol. Earth Syst. Sci.* **14**(10), 1931–1941 (2010).
15. Liang, H., Huang, S., Meng, E. & Huang, Q. Runoff prediction based on multiple hybrid models. *J. Hydraul. Eng.* **51**(1), 112–125 (2020).
16. Cortes, C. & Vapnik, V. Support-vector networks. *Mach. Learn.* **20**(3), 273–297 (1995).
17. Liao, J., Wang, W. S., Li, Y. Q. & Huang, W. J. Support vector machines and their application to runoff prediction. *J. Sichuan Univ. Eng. Sci. Edn.* **38**(6), 24–28 (2006).
18. Li, J., Wang, L., Ma, G. W. & Wu, K. Application of LS–SVM in runoff prediction, China. *Rural Water Conserv Hydropower* **5**, 8–10 (2008).
19. Shabri, A. & Suhartono,. Streamflow forecasting using least-squares support vector machines. *Hydrol. Sci. J.* **57**(7), 1275–1293 (2012).
20. Mallat, S. G. A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Trans. Pattern Anal. Mach. Intell.* **11**(7), 674–693 (1989).
21. Huang, N. E. *et al.* The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. *Proc. Roy. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* **454**(1971), 903–995 (1998).
22. Wu, Z. & Huang, N. E. Ensemble empirical mode decomposition: A noise-assisted data analysis method. *Adv. Adapt. Data Anal.* **1**(01), 1–41 (2009).
23. Torres, M. E., Colominas, M. A., Schlotthauer, G., & Flandrin, P. A complete ensemble empirical mode decomposition with adaptive noise. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* 4144–4147. IEEE (2011).
24. Dragomiretskiy, K. & Zosso, D. Variational mode decomposition. *IEEE Trans. Signal Process.* **62**(3), 531–544 (2013).
25. Huan, J., Cao, W. & Qin, Y. Prediction of dissolved oxygen in aquaculture based on EEMD and LSSVM optimized by the Bayesian evidence framework. *Comput. Electron. Agric.* **150**, 257–265 (2018).
26. Huang, C., Cao, Y. & Zhou, L. Application of optimized GM(1,1) model based on EMD in landslide deformation prediction. *Comput. Appl. Math.* **40**(8), 1–21 (2021).
27. Jamei, M. *et al.* Designing a multi-stage expert system for daily ocean wave energy forecasting: A multivariate data decomposition-based approach. *Appl. Energy* **326**, 119925 (2022).
28. Jamei, M. *et al.* Forecasting daily flood water level using hybrid advanced machine learning based time-varying filtered empirical mode decomposition approach. *Water Resour. Manag.* **36**(12), 4637–4676 (2022).
29. Raj, N. Prediction of sea level with vertical land movement correction using deep learning. *Mathematics* **10**(23), 4533 (2022).
30. Rezaie-Balf, M., Naganna, S. R., Kisi, O. & El-Shafie, A. Enhancing streamflow forecasting using the augmenting ensemble procedure coupled machine learning models: Case study of Aswan High Dam. *Hydrol. Sci. J.* **64**(13), 1629–1646 (2019).
31. Nourani, V., Baghanam, A. H. & Gokcekus, H. Data-driven ensemble model to statistically downscale rainfall using nonlinear predictor screening approach. *J. Hydrol.* **565**, 538–551 (2018).
32. Wu, L., Liu, S., Yao, L., Yan, S. & Liu, D. Grey system model with the fractional order accumulation. *Commun. Nonlinear Sci. Numer. Simul.* **18**(7), 1775–1785 (2013).
33. Kumar, D., Singh, A., Samui, P. & Jha, R. K. Forecasting monthly precipitation using sequential modelling. *Hydrol. Sci. J.* **64**(6), 690–700 (2019).
34. Liu, M. D., Ding, L. & Bai, Y. L. Application of hybrid model based on empirical mode decomposition, novel recurrent neural networks and the ARIMA to wind speed prediction. *Energy Convers. Manag.* **233**, 113917 (2021).
35. Zhang, Q., Xu, C. Y., Chen, Y. D. & Ren, L. Comparison of evapotranspiration variations between the Yellow River and Pearl River basin, China. *Stoch. Env. Res. Risk Assess.* **25**(2), 139–150 (2011).
36. Kisi, O. & Ay, M. Comparison of Mann–Kendall and innovative trend method for water quality parameters of the Kizilirmak River, Turkey. *J. Hydrol.* **513**, 362–375 (2014).
37. Zhang, X., Tuo, W. & Song, C. Application of MEEMD–ARIMA combining model for annual runoff prediction in the Lower Yellow River. *J. Water Clim. Change* **11**(3), 865–876 (2020).
38. Zhang, J., Xiao, H. & Fang, H. Component-based reconstruction prediction of runoff at multi-time scales in the source area of the Yellow River based on the ARMA model. *Water Resour. Manag.* **36**(1), 433–448 (2022).

Author contributions

All authors contributed to the study conception and design. writing and editing: S.G. and Y.W.; chart editing: H.C.; preliminary data collection: X.Z. 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].

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to X.Z.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.