Application of the decomposition-prediction-reconstruction framework to medium- and long-term runoff forecasting

Yi Ji, Hong-Tao Dong, Zhen-Xiang Xing, Ming-Xin Sun, Qiang Fu and Dong Liu

ABSTRACT

Medium- and long-term runoff forecasting has always been a problem, especially in the wet season. Forecasting performance can be improved using complementary ensemble empirical mode decomposition (CEEMD) to produce clearer signals as model inputs. In the forecasting models based on CEEMD, the entire time series is decomposed into several sub-series, each sub-series is divided into training and validation datasets and forecasted by some common models, such as least squares support vector machine (LSSVM), and finally an ensemble forecasting result is obtained by summing the forecasted results of each sub-series. This model was applied to forecast the inflow runoff of the Shitouxia Reservoir (STX Reservoir). The forecasting results show that the Nash efficiency coefficient of the LSSVM model is 0.815, and the Nash efficiency coefficient of the CEEMD-LSSVM model is 0.954, an increase of 13.9%. The root mean square error value is reduced from 20.654 to 10.235, a decrease of 50.4%. The runoff forecasting performance can be effectively improved by applying the CEEMD-LSSVM model. When analyzing the annual runoff forecasting results month by month, it was found that the forecasting results for November to April were unsatisfactory compared results from the nearest neighbor bootstrapping regressive (NNBR) model, which was more suitable for the dry season, but the forecasting results for May to October improved significantly. This also proves that the CEEMD-LSSVM model has a great advantage in the forecasting of inflow runoff during the wet season. In the optimized operation of reservoirs, the forecasting result of inflow runoff in the wet season is more important than in the dry season. Therefore, when forecasting annual runoff month by month, the CEEMD-LSSVM model is recommended for the wet season combined with the NNBR model for the dry season.

Key words | CEEMD-LSSVM, hybrid approach, medium- and long-term runoff forecasting, NNBR

Yi Ji
Hong-Tao Dong
Zhen-Xiang Xing (corresponding author)
Ming-Xin Sun
Qiang Fu
Dong Liu
School of Water Conservancy and Civil Engineering,
Northeast Agricultural University,
Harbin 150030,
China
E-mail: *zxxing@neau.edu.cn*

HIGHLIGHTS

- CEEMD is suitable for non-linear and non-stationary time series.
- A comprehensive evaluation system was used to evaluate the accuracy of the different models.
- The performance of the models can be improved by using the CEEMD method.
- Different models are used for forecasting in the dry season and the wet season respectively.

doi: 10.2166/ws.2020.337

INTRODUCTION

Hydrological forecasting is of significant importance for planning and managing water resources. If the forecasting lead time is longer than the maximal confluence time of the basin plus 3 days but shorter than 1 year, it is classed as medium- and long-term hydrological forecasting (Tang *et al.* 2008). Medium- and long-term hydrological forecasting is a powerful means of making full use of water resources and realizing optimal reservoir scheduling. It is an important basis for correct decision-making in reservoir operation management.

Runoff forecasting has attracted wide attention in the last few decades. Physical models are usually used for runoff prediction. In this method, the runoff generation process is simulated by equations with specific boundary conditions. But this kind of model needs a lot of accurate historical rainfall runoff data to calibrate the model parameters. In practice, it is difficult to ensure the accuracy of the data and meet the requirement of the sample size, which often leads to unsatisfactory calibration results of model parameters and poor forecasting performance of the model.

Many non-linear data-driven models, such as artificial neural networks (ANN) (Sivapragasam & Vasudevan 2010; Humphrey *et al.* 2016; Shiri & Kisi 2010; Sudheer *et al.* 2010), adaptive neuro-fuzzy inference system (ANFIS) (Nayak *et al.* 2004; Ashrafi *et al.* 2017), genetic programming (GP) (Kisi & Cimen 2011; Danandeh Mehr 2013; Ravansalar *et al.* 2017), support vector machines (SVM) (Asefa *et al.* 2006; Kisi & Cimen 2011; Huang *et al.* 2014), and nearest neighbor bootstrapping regressive models (NNBR) (Ye & Cheng-You 2011) have been proposed for medium- and long-term runoff forecasting. However, In the middle and long-term runoff forecasting, which model shows the best performance is not yet clear. ANN has strong non-linear mapping capabilities, but it also has problems such as slow learning speed, overfitting, and dimensionality disasters. Therefore, when processing complex hydrological data, its forecasting performance is not satisfactory. SVM is a small sample statistical learning model based on the Vapnik–Chervonenkis (VC) dimensionality theory and the principle of structural risk minimization; it can effectively avoid dimensionality disasters, has high simulation accuracy, and can theoretically achieve global optimization (Vapnik 2000). Least square support vector machine (LSSVM) is an improved SVM model. Its convergence speed is faster than traditional SVM and the model forecasting accuracy is better. Some studies have shown that clearer sub-signals can be generated as model inputs through signal decomposition technology, thus improving the forecasting performance of the model (Tan *et al.* 2018). Partal (2007) proposed a wavelet-neuro-fuzzy model to simulate precipitation, and applied the periodic expression ability of wavelet transform technology to improve the forecasting accuracy of the model. Since wavelet transform is suitable for processing non-stationary data mathematical tools, and the input data is required to be linear, the mother wavelet also requires a pre-set basis function (Niu *et al.* 2016), which limits its application. Hydrological time series usually have highly complex non-stationary characteristics, and the adjacent states are mostly non-linear relations (Huang *et al.* 2009).

Therefore, empirical mode decomposition (EMD) is used in the field of hydrological data analysis because it is suitable for processing complex non-linear and non-stationary time series (Karthikeyan & Nagesh Kumar 2013). In addition, EMD is based on the principle of local scale separation and does not require a predetermined basis function, which is adaptive and intuitive (Sang *et al.* 2012). The entire series is decomposed by EMD into several sub-series called intrinsic mode function (IMF) and a residue. However, there are various signal oscillation modes in the IMF component after EMD decomposition because it is discontinuous and there is a local intermittent component. In response to this phenomenon, ensemble empirical mode decomposition (EEMD) adds a certain white noise to the original signal, and the uniform distribution of Gaussian white noise in the frequency range makes the IMF component after EMD decomposition continuous on the time scale, thus overcoming the modal aliasing defects caused by intermittent components (Huang *et al.* 1998; Huang & Wu 2008). While the white noise added in the EEMD method overcomes the pattern aliasing problem in the

EMD method, it also brings a reconstruction error into the decomposed IMF component. The reconstruction error can only be reduced in the EEMD and cannot be eliminated. In order to overcome this shortcoming, a polymer EMD method, complementary ensemble empirical mode decomposition (CEEMD), is proposed (Yeh *et al.* 2010). Therefore, the IMF is finally determined according to the CEEMD method. The reconstruction error can be offset when the components are averaged. As an improved algorithm of EMD, CEEMD (Li *et al.* 2014) not only effectively solves the modal aliasing problem of EMD, but also preserves the advantages of EMD processing non-stationary signals, such as adaptiveness, two filtering characteristics and so on.

This study summarizes the limitations of other researchers' conclusions and the advantages of CEEMD, which is more suitable for non-linear and non-stationary data, and proposes the CEEMD-LSSVM model. The CEEMD-LSSVM model, based on the decomposition-prediction-reconstruction pattern can be applied to predict the non-linear and non-stationary runoff time series well. The input is not the runoff time series, but the signal decomposed by CEEMD. It can be used to verify whether the forecasting model can be applied to runoff forecasting. Taking the Shitouxia (STX) Reservoir as an example, the CEEMD model is used to decompose the original time series into several sub-series, and the LSSVM model is used to forecast each sub-series, then ensemble forecast is obtained by summing the forecasting results of each sub-series. Finally, the forecasting results are compared with those obtained by LSSVM and NNBR models.

METHODOLOGY

CEEMD

The EMD method is used for decomposing complex signals into single-frequency signals. EMD is an empirical, intuitive, direct and self-adaptive data processing method for non-linear and non-stationary time series (Huang *et al.* 1998). CEEMD (Torres *et al.* 2011) is an enhancement of EMD and EEMD (Wu & Huang 2009). The decomposed signals can be arranged according to the frequency from high to low. The white noise added into the EEMD method overcomes the pattern aliasing problem in the EMD method; it also brings the reconstruction error into the decomposed IMF component. CEEMD adds two Gaussian white noises at the same time in the process of reconstructing the signal. The amplitudes are the same but the phases are opposite. Therefore, the IMF is finally determined according to the CEEMD method. The reconstruction error can be offset when the components are averaged. The specific process of CEEMD decomposition is as follows.

Step 1. Set the maximum aggregation number I and white noise amplitude, and initialize it so that $i = 1$.

Step 2. Generate white noise $n_i(t)$, reconstruct a pair of new signals according to Equation (1):

$$\begin{cases} x_i^+(t) = x_{i-1}(t) + n_i(t) \\ x_i^-(t) = x_{m(t)i-1}(t) - n_i(t) \end{cases} \tag{1}$$

Step 3. Connect all local maxima and minima by a cubic spline interpolation, and generate an upper and lower envelope $e_{max}(t)$, $e_{min}(t)$.

Step 4. Compute the envelope mean using Equation (2):

$$m(t) = (e_{\max}(t) + e_{\min}(t))/2 \tag{2}$$

Step 5. Calculate the difference between $x(t)$ and $m(t)$ as $c(t)$:

$$c(t) = x(t) - m(t) \tag{3}$$

Step 6. Check whether or not $c(t)$ is an IMF according to the two conditions mentioned above. If $c(t)$ is an IMF, go to Step 6; otherwise, let $x(t) = c(t)$, and repeat Steps 3–5 until $c(t)$ is an IMF.

Step 7. Calculate the residue $r(t) = x(t)$-$c(t)$. If the residue $r(t)$ becomes a monotonic function or at most has one local extreme point, the whole decomposition is completed. Otherwise, let $x(t) = r(t)$. Repeat Step 3–6, to obtain q IMF components:

$$\begin{aligned} x_i^+(t) &= \sum_{j=1}^{q} c_{j,i}^+(t) \\ x_i^-(t) &= \sum_{j=1}^{q} c_{j,i}^-(t) \end{aligned} \tag{4}$$

Step 8. Determine if the maximum number of iterations is reached. If $i < I$, let $i = i + 1$, loop Steps 2–7, which requires adding white noise to the initial signal, and that the white noise added each time is different.

Step 9. Find the average value of the IMF after the EMD decomposition, and obtain:

$$c_j = \frac{1}{2I}\sum_{i=1}^{1}(c_{j,i}^{+} + c_{j,i}^{-}) \tag{5}$$

where c_j is the jth IMF.

LSSVM

LSSVM originated from SVM and is a powerful methodology for solving problems in non-linear classification, function estimation and regression. This method has been applied in pattern recognition, signal processing and non-linear regression estimation. LSSVM was proposed by Suykens and Vandewalle in 1999 and has been employed in chaotic time series prediction (Suykens & Vandewalle 1999). It uses a set of linear equations for training while SVM uses a quadratic optimization problem, which is the major difference between the two. Compared with SVM, LSSVM uses the least squares linear system as the loss function, and the solution process becomes a set of equations. The solution speed is faster. The model works as follows:

For a given data set $\{x_i, y_i\}_{i=1}^{N}$, the mathematical model of LSSVM can be described as:

$$y_i = \omega^{\tau}\varphi(x_i) + b + e_i,\ i = 1, \cdots, n \tag{6}$$

where x_i is the ith m-dimensional input, y_i is the ith real-valued output, n is the number of samples, φ is the kernel-space mapping function, ω^{τ} is the weight vector, b is the deviation, and $e_i \in R$ is the error variable.

Solving coefficients based on Lagrangian function, finally construct the LSSVM regression function as:

$$y(x) = \sum_{i=1}^{n}\alpha_i K(x, x_i) + b \tag{7}$$

where $y(x)$ is the forecast object, x_i is the support vector obtained through training, x is the forecasting sample, α is the Lagrangian coefficient obtained through training, b is the deviation amount, and $K(x,x_t)$ is the kernel function. This study chooses Gaussian radial basis function (RBF) as the kernel function (Maity *et al.* 2010) Its expression is as follows:

$$K(x, x_t) = \exp\left\{-\frac{\|x - x_i\|^2}{2\sigma^2}\right\} \tag{8}$$

where σ is the of the kernel width. The structure of LSSVM is shown in Figure 1.

CEEMD-LSSVM

Because the inflow runoff has non-linear and non-stationary characteristics, the traditional signal analysis method extracts the characteristic information of the signal from the time domain or the frequency domain, which makes the information have limitations and easily causes information loss. CEEMD is particularly well suited for processing complex non-linear systems due to its good adaptability. In order to cope with the non-stationary problem of runoff, a CEEMD-LSSVM hybrid model is built, where CEEMD is used to decompose the original time series into

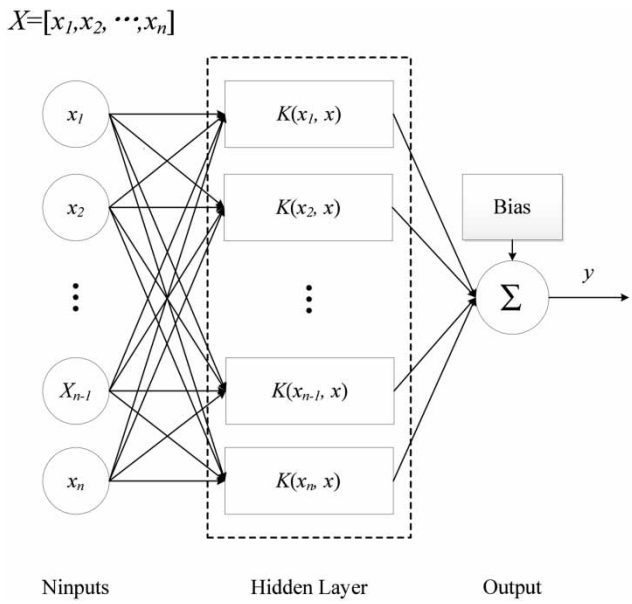

Figure 1 | The structure of LSSVM.

several sub-series, and LSSVM is used to build a forecasting model for each sub-series. Sub-series obtained by CEEMD are relatively stationary, and can provide information about the original data structure and its periodicity. Therefore, the performance of the forecasting models are expected to be improved by giving useful information on various resolution levels.

Forecasting adopts the 'decomposition-prediction-reconstruction' based on CEEMD, and includes these steps: Firstly, decompose runoff time series into a collection of IMFs and a residue using CEEMD; and each sub-series is divided into training samples and validation samples. Then, the CEEMD-LSSVM forecasting model of sub-series is established based on the training samples, and the decomposition signals of the training samples are forecast; finally, the forecasting results of the validation samples are reconstructed, and the accuracy of the model is analyzed. The framework of CEEMD-LSSVM is shown in Figure 2.

In order to verify the prediction performance of the CEEMD-LSSVM model in dry season and flood season, two forecasting models, LSSVM and NNBR (Lall & Sharma 1996), are applied for comparative prediction. The input of the NNBR model is monthly runoff in the dry season. The forecasting results of the three are then analyzed and the accuracy of the forecasting methods at different seasons are compared.

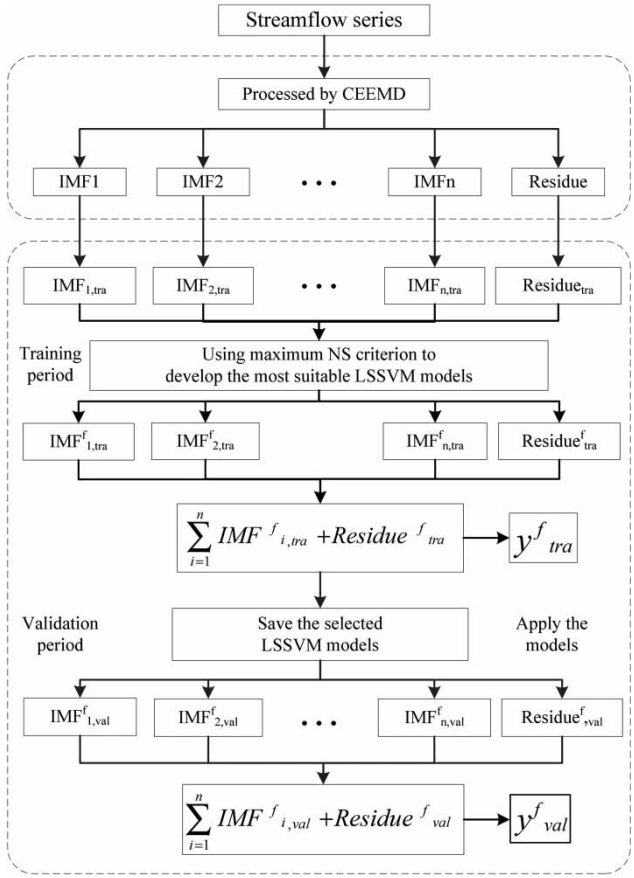

Figure 2 | The flowchart of the experiment.

CASE STUDY

Study area and data

Qinghai Province is located in arid area. In order to relieve the pressure of water supply in the Xining area, the provincial capital of Qinghai, an inter-basin water transfer system, the Datong river-to-Huangshui river, was built. It is an important project to solve the water supply problem in Xining city (Figure 3). The STX Reservoir is located on the main stream of the Datong River. The catchment area of the Datong River is 15,000 km^2 and the length of the main stream is 560.7 km. The STX Reservoir is the source reservoir of the project, so medium- and long-term runoff forecasting is of practical significance for the joint regulation of these reservoirs and rational allocation and utilization of water resources.

A longer training period should be selected in order to ensure the effect of the training period, so this study selected data for the STX Reservoir from 1956 to 2000 (540 time periods) as the training period, and 2001–2010 (120 time periods) as the validation period. Applying CEEMD to decompose the runoff time series of the STX Reservoir, it was decomposed into nine IMFs and one residue according to the frequency. As shown in Figure 4, as the component order increases, IMF_1-IMF_9 gradually show regularization and smoothing, the periodicity gets stronger and stronger, and the residue slowly decreases.

Model verification

The Nash–Sutcliffe efficiency coefficient (NS), root mean square error (RMSE), correlation coefficient (R), mean

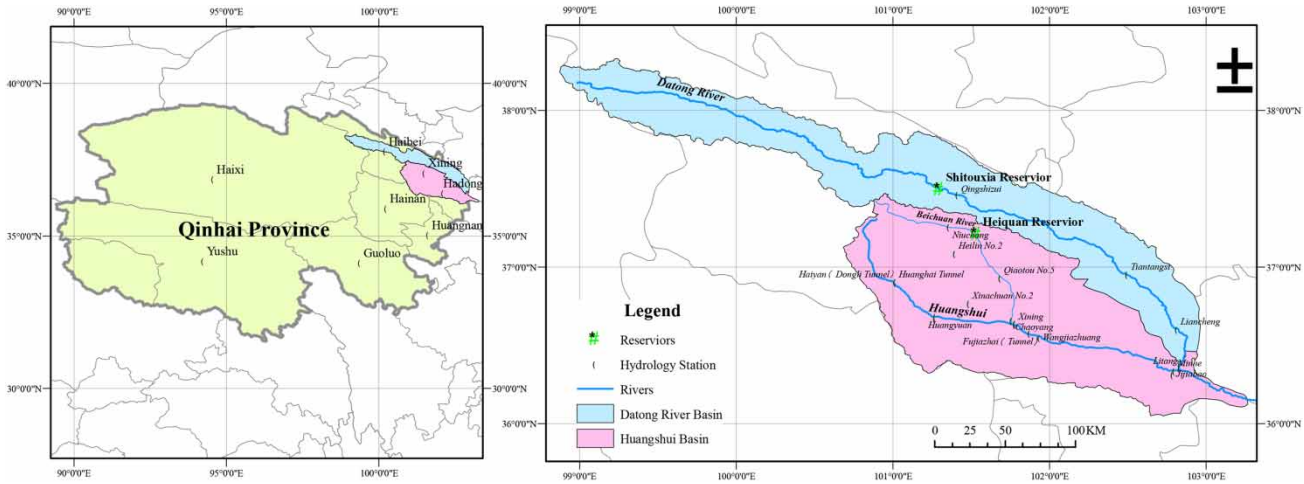

Figure 3 | Map of the Datong Basin.

absolute relative error (MARE) and mean absolute error (MAE) were used to evaluate the performance of forecasting models:

$$NS = 1 - \frac{\sum_{i=1}^{n} (y'_i - y_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \tag{9}$$

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

$$MARE = \frac{1}{n}\sum_{i=1}^{n} \left|\frac{(y_i - y'_i)}{y_i}\right| \tag{11}$$

$$R = \frac{\sum_{i=1}^{n} (y_i - \bar{y})(y'_i - \bar{y'})}{\sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2 \sum_{i=1}^{n} (y'_i - \bar{y'})^2}} \tag{12}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - y'_i| \tag{13}$$

where n is the length of runoff time series; y_i and y_i' are the observed and forecasted runoff at time i, $\bar{y}$ and $\bar{y'}$ are the mean of the observed and forecasted runoff, respectively. The best fit between observed and forecasted values would be NS = 1, R = 1, RMSE = 0, MARE = 0 and MAE = 0. The closer the NS and R values are to 1, and the RMSE, MARE and MAE values are to 0, the better the performance of the forecasting model.

Results and discussion

Characteristic analysis of runoff decomposition signals

In previous forecasting experiments, the runoff series in the validation period are processed using the decomposition model saved in the training period. However, the decomposition results may differ significantly depending on the length of the series. The longer the runoff sequence, the more sub-series are decomposed (Huang *et al.* 1998). The original runoff sequence was decomposed into 10 sub-families, including nine IMFs and one residue. These sub-series have different frequencies and amplitudes.

Application analysis and discussion

In the forecasting model, the entire time series was decomposed into several IMFs and one residue, and then each decomposition component was divided into a training period and a validation period. The training period data was used to build the model, and the validation period was used for forecast testing. The results showed that IMF8, IMF9 and the residue of the runoff sequence were highly accurate by the multivariate linear fitting method, and other components were forecast by LSSVM. Figure 5 shows the forecasting results of the runoff sequence decomposition. Table 1 shows the forecast results of the components. It can be seen that in addition to the higher frequency IMF1 and IMF2, the other IMFs and the residue are

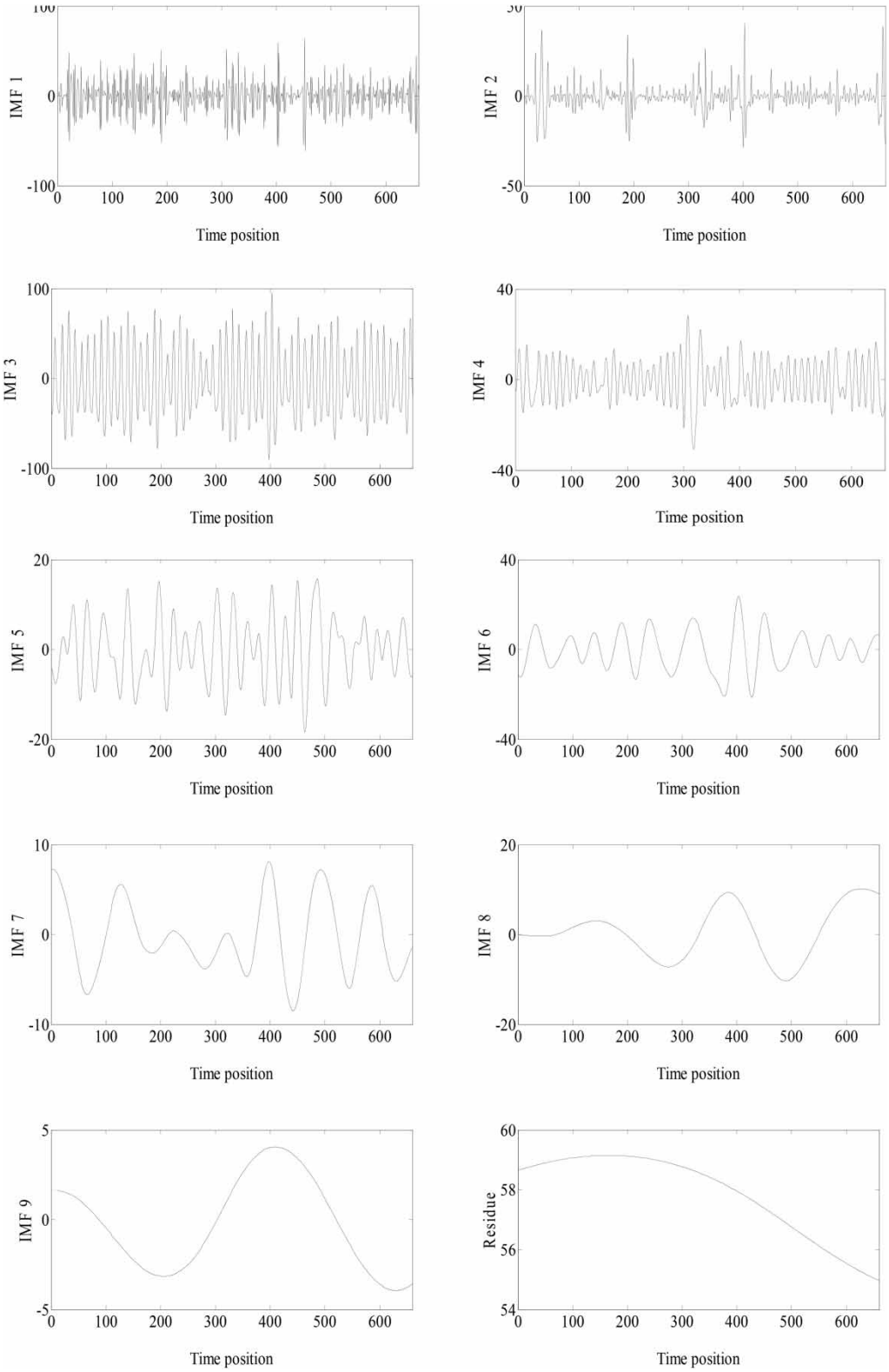

Figure 4 | The decomposition results of the runoff series for the STX Reservoir based on CEEMD.

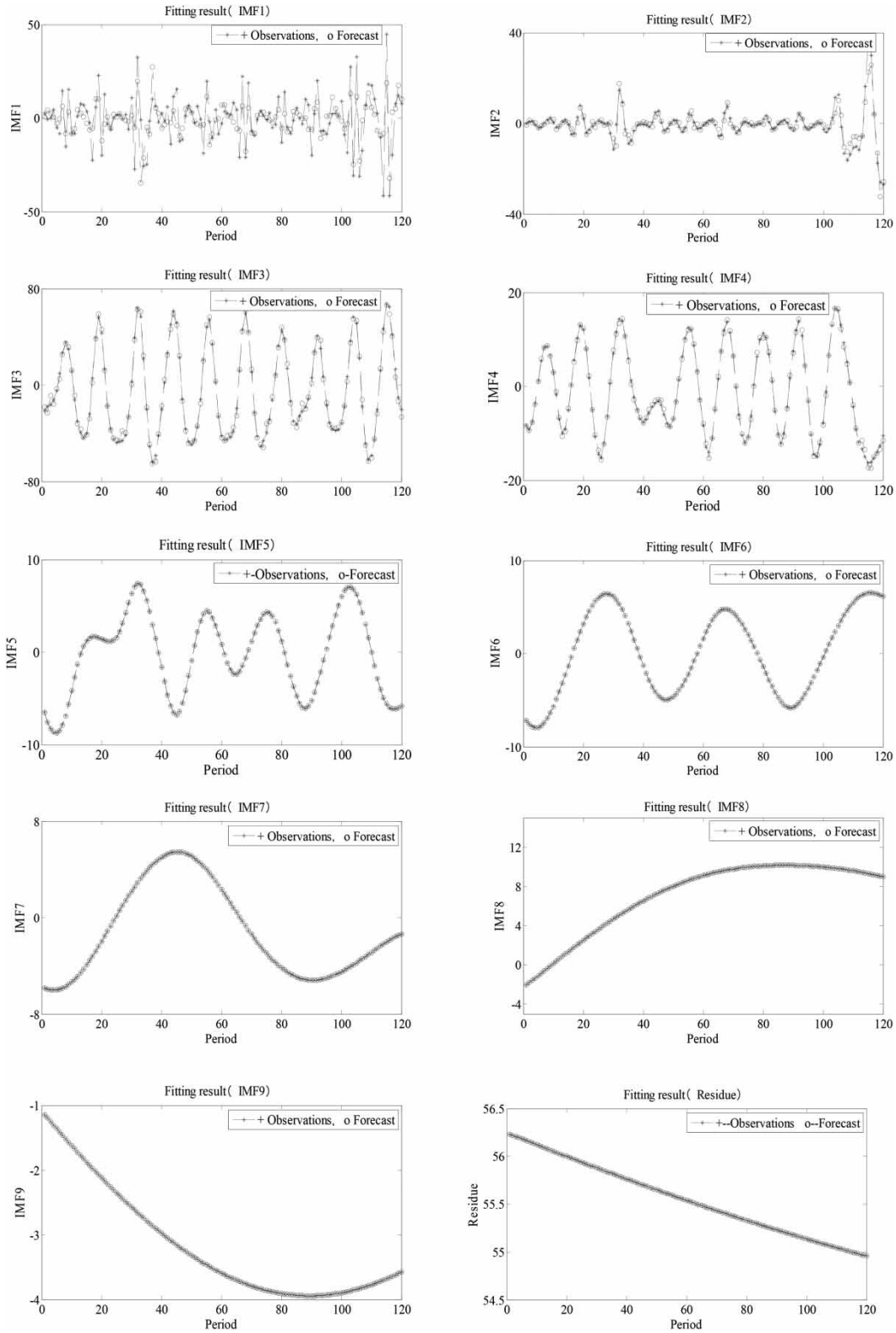

Figure 5 | The forecasting results of runoff sequence decomposition.

Table 1 | Forecasting performance of the components

	Criteria	IMF_1	IMF_2	IMF_3	IMF_4	IMF_5	IMF_6	IMF_7	IMF_8	IMF_9	Residue
Training dataset	R	0.857	0.9643	0.999	1	1	1	1	1	1	1
	NS	0.5022	0.8663	0.9964	0.999	1	1	1	1	1	1
	MAE	5.6519	1.2482	1.4208	0.2048	0.0014	0.0001	0	0	0	0
	MARE (%)	175.57	126.35	9.7520	5.4598	0.9614	0.2853	0.0537	0	0	0
	RMSE	6.3545	1.2346	1.4652	0.1631	0.0083	0.0014	0.0001	0	0	0
Validating dataset	R	0.7119	0.9312	0.9969	0.9983	1	1	1	1	1	1
	NS	0.5022	0.8663	0.9937	0.9966	1	1	1	1	1	1
	MAE	7.3310	1.6842	2.2209	0.4052	0.0225	0.0023	0.0014	0.0001	0	0
	MARE (%)	229.44	198.92	12.2750	8.9845	1.6914	0.3594	0.0892	0.0039	0.0001	0
	RMSE	9.8449	2.7634	2.8642	0.5371	0.0280	0.0027	0.0017	0.0001	0	0

predicted to perform well. When the original sequence was decomposed layer by layer, the frequency of the decomposition component was lowered and there was a gradual apparent periodic process. Therefore, the predicted performance of the IMFs gradually increased, resulting in NS and R close to 1, while RMSE, MARE and MAE were approaching 0. The component forecasting results were used to reconstruct the information to obtain the runoff forecasting value of the STX Reservoir. R and NS were close to 1, RMSE, MARE and MAE were lower, indicating that the forecasting result was better.

Table 2 shows the forecasting performance of the LSSVM, NNBR and CEEMD-LSSVM models for the STX Reservoir. The performance of the CEEMD-LSSVM model was significantly improved, and the NS and R values were higher than for the other models. This indicates that the CEEMD-LSSVM model produced better agreement between the forecasted and observed runoff and the RMSE and MAE values were lower than for the other models, so the average forecasting error was low. However, the MARE value of the CEEMD-LSSVM model was higher than that of the NNBR model, which can be attributed to the fact that the CEEMD-LSSVM model is not suitable in the dry season. According to Equation (11), the runoff in the dry season is relatively low, and thus even a small deviation will lead to a large relative error. The performance of the CEEMD-LSSVM and LSSVM models is shown in Figure 6. It can be seen that CEEMD-LSSVM is significantly better than the LSSVM model for the forecasting of runoff, especially in the peak forecasting, which proves that the overall accuracy of the LSSVM model can be improved by the CEEMD method. Therefore, the decomposition can be helpful to transform non-linear and non-stationary time series to stationary time series and can be useful to improve the forecasting capacity. It can also be seen that in terms of minimum values, the CEEMD-LSSVM model does not perform well and even predicts negative values.

Table 2 | The performance of different models

Model	R	NS	MAE	MARE (%)	RMSE
CEEMD-LSSVM	0.978	0.954	7.562	26.864	10.235
LSSVM	0.903	0.815	13.167	32.167	20.654

MAE, RMSE and MARE were used to evaluate the forecasting accuracy in each month (Table 3). It is obvious that the CEEMD-LSSVM model is very suitable for runoff forecasting in the wet season, but the prediction effect in the dry season is not as good as the NNBR model. As shown in Table 3, for the STX Reservoir, the CEEMD-LSSVM model predicts better results than the other models from May to October, but in other months it is not as accurate as the NNBR model. The scatter plots of the runoff predictions and observations show that in the runoff prediction process of the reservoir, the prediction accuracy of NNBR when the runoff is small is higher than that of the CEEMD method, as shown in Figure 7. Figure 7(a) and 7(b) show the comparison of the NNBR model in the wet season and the dry season, respectively. Figure 7(c) shows the comparison between the observed and predicted values of the NNBR model in the dry season, respectively. It can be seen that the scatter points in Figure 7(b) is more

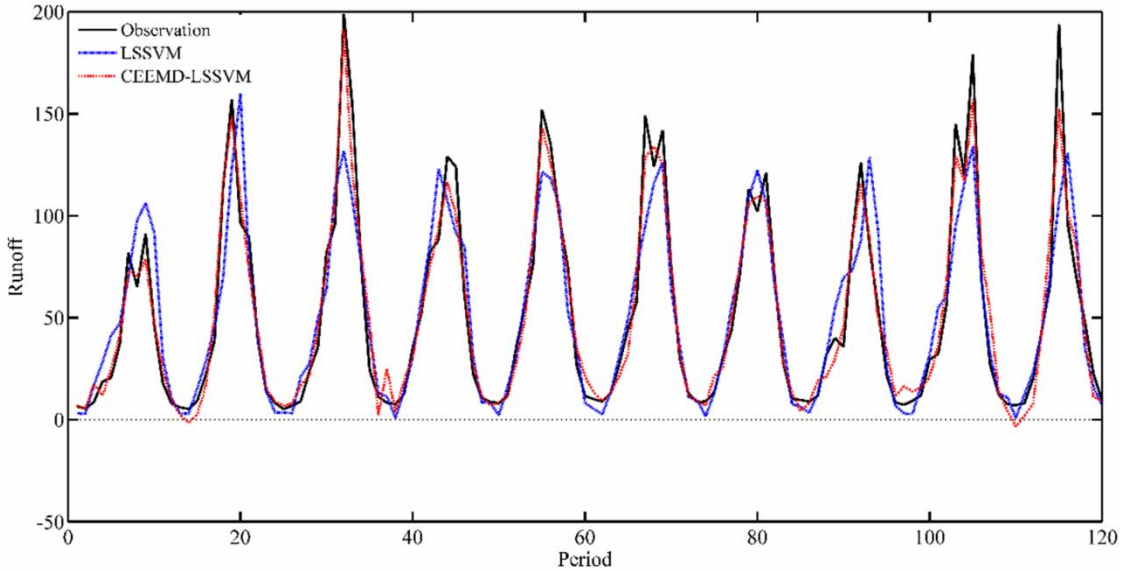

Figure 6 | The forecasting results of the validation period using the CEEMD-LSSVM and LSSVM models for the STX Reservoir.

Table 3 | The validation forecasting performance for each month using different models at the STX Reservoir

Model	Criteria	1	2	3	4	5	6	7	8	9	10	11	12
CEEMD-LSSVM	MAE	5	3	6	7	**7**	**10**	**13**	**8**	**16**	**5**	10	3
	MARE (%)	58	43	54	27	**18**	**16**	**9**	**7**	**14**	**8**	40	28
	RMSE	7	4	6	8	**8**	**12**	**17**	**9**	**18**	**5**	13	4
LSSVM	MAE	3	5	4	4	11	15	34	31	25	16	8	2
	MARE (%)	38	68	46	17	33	24	25	27	22	29	34	24
	RMSE	3	6	6	5	13	20	41	37	29	22	9	3
NNBR	MAE	**1.5**	**1.6**	**1.4**	**1.4**							**1.9**	**0.9**
	MARE (%)	**18.5**	**19.3**	**13.1**	**5.7**							**8.8**	**8.3**
	RMSE	**1.9**	**1.8**	**1.6**	**1.7**							**2.0**	**1.2**

Note: the bold values are the best values.

scattered and do not cluster near the diagonal. The scatter points in Figure 7(a) and 7(c) are distributed near the diagonal.

Therefore, in the runoff forecasting for the season, the accuracy of the CEEMD-LSSVM model is considerable; in the runoff forecasting for the dry season, the simulation effect of the NNBR model is better than the CEEMD-LSSVM model.

It can be seen from the Figure 7(b) that the runoff in the dry season forecast by the CEEMD model is even negative, because the runoff sequence decomposed by CEEMD is longer, there are more components obtained by decomposition the amplitude of the high-frequency term is larger, the fluctuation range is larger, and there is more noise. Each component is separately forecasted, and there may be accumulated errors when linearly superimposed, resulting in low forecasting performance or even negative values at minimum values.

A Taylor diagram can quantify the correspondence ratio between the simulation results and the measured data, including the correlation coefficient, RMSE and the standard deviation (Qiang *et al.* 2018). As shown in Figure 8, the standard deviations of the LSSVM model, the CEEMD-LSSVM model and the combined model are 45.03, 46.29, 46.51, respectively; RMSE values are 20.65, 10.23, 9.09, respectively; R values are 0.903, 0.978, 0.985

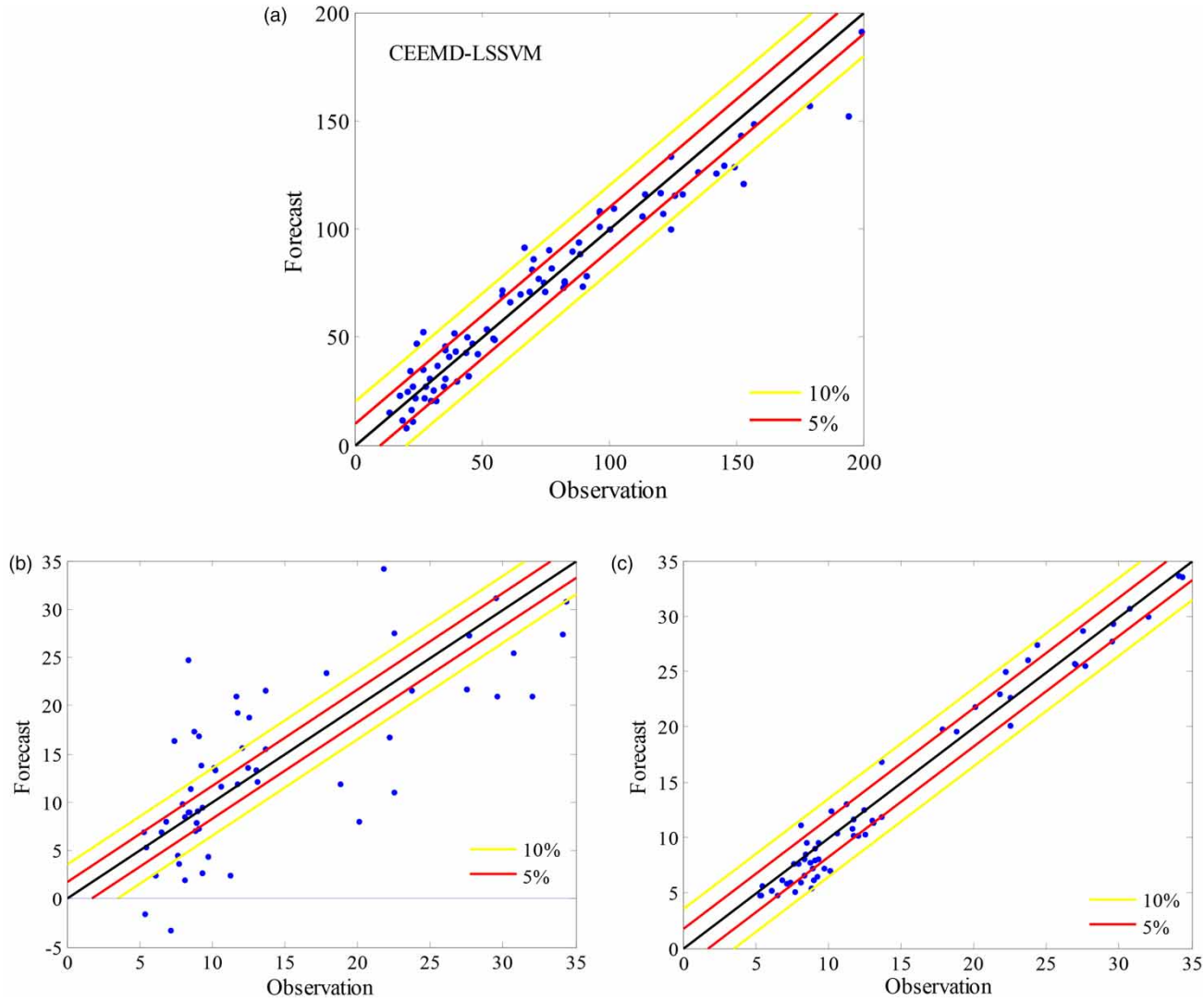

Figure 7 | Comparison of the forecasting results for the STX Reservoir runoff.

respectively. Although the standard deviation of the combined model is slightly larger than the other models, the RMSE and R values are smaller and higher than for the other models, respectively. In general, the combined model can effectively improve the accuracy and can be applied to the forecasting of the inflow runoff of the STX Reservoir.

Figure 9 shows that the forecasting accuracy in both the wet and dry seasons can be significantly improved. Table 4 shows that compared with the CEEMD-LSSVM model, the R and NS values are improved from 0.978 and 0.954 to 0.985 and 0.966, while the RMSE, MARE and MAE values are reduced from 10.235, 26.864% and 7.562 to 8.835, 12.112% and 5.563 for values of STX Reservoir.

In summary, values of results of values of analysis demonstrate that the proposed CEEMD-LSSVM model is able to attain better results than values of LSSVM model, a drastic improvement in terms of different evaluation measures for monthly runoff time series forecasting. This also indicates that the idea of decomposition-prediction-reconstruction is feasible and the proposed CEEMD-LSSVM model can overcome values of drawbacks of individual models by generating a synergetic effect in forecasting.

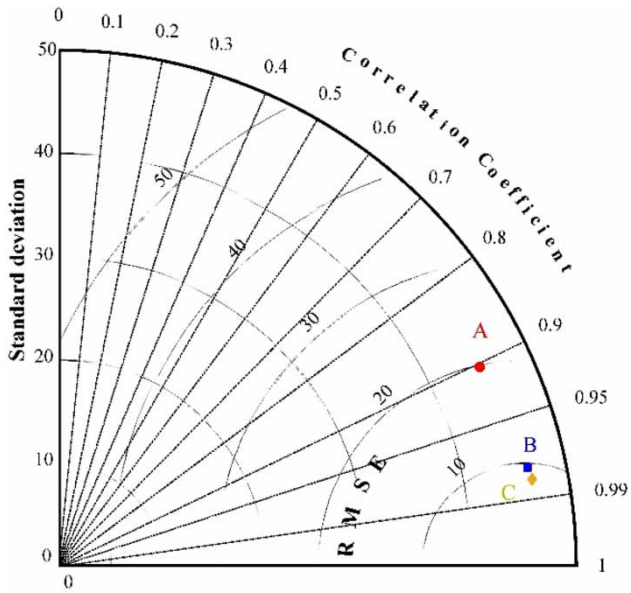

Figure 8 | Taylor diagram for comparison of model forecasting results.

CONCLUSIONS

In this study, runoff forecasting methods based on a decomposition-prediction-reconstruction model with the CEEMD, LSSVM and NNBR models were used to forecast runoff. The accuracy of the three forecasting methods was compared and analyzed. The forecasting methods suitable for different months were selected to forecast the runoff of the STX Reservoir by stages. The main conclusions are as follows:

(1) It is proved that the runoff decomposition model based on CEEMD can effectively identify the characteristic information of the original runoff series, and decompose the runoff series into several IMF components and one residual quantity whose frequencies are from high to low.

(2) The CEEMD-LSSVM model can significantly improve the accuracy of the dry season prediction, but it is relatively low for the dry season prediction.

(3) The NNBR model can better characterize the auto-correlation of runoff and has little variation in numerical fluctuations, so the NNBR model shows better performance than the other models in dry season prediction. It is therefore suggested that the monthly runoff forecasting for the STX Reservoir should combine the NNBR model with the CEEMD-LSSVM model, to use CEEMD-LSSVM model in the wet season, while in the dry season, the NNBR method with relatively high accuracy is used to forecast runoff, with the two methods being combined in stages to improve the prediction accuracy of inflow runoff.

The method in this study provided reliable results to simulate the runoff data to can provide support for

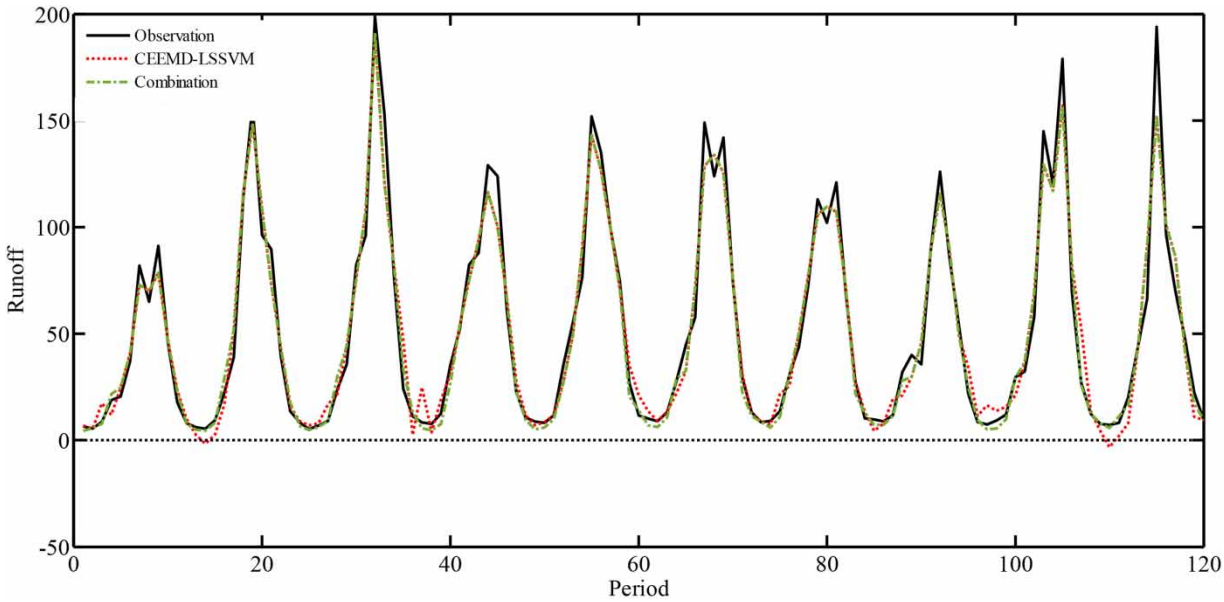

Figure 9 | The combined forecasting results for the STX Reservoir runoff.

Table 4 | The comprehensive validation performance combining NNBR and CEEMD-LSSVM models

Model	R	NS	MAE	MARE(%)	RMSE
CEEMD-LSSVM	0.978	0.954	7.562	26.864	10.235
Combination	0.985	0.966	5.563	12.112	8.835

decision-making and risk analysis in reservoir operation. In the future, the techniques can be applied to different reservoirs. In theory, because of the complexity of the hydrological system, in order to achieve more accurate prediction, it is necessary to analyze the physical meaning of each sub-series. With the increase of global temperature and atmospheric humidity, there will be more and more extreme hydrometeorological events such as extreme precipitation and drought. Reservoirs' real-time operation will be affected by extreme runoff more frequently. Consequently, when the trajectory of future hydrological elements no longer follows historical data, the prediction accuracy of the model needs to be improved.

The uncertainty of runoff prediction error should be considered in future reservoir real-time operation, so as to improve the reservoir real-time operation potentiality. This will be the focus of future research.

ACKNOWLEDGEMENTS

This research was supported by the National Key R&D Program of China (2017YFC0406004; 2018YFC0407303), The National Natural Science Foundation of China (51979038, 51569003, 51909033), the Heilongjiang Postdoctoral Fund (LBH-Z18020), and the Natural Science Foundation of Heilongjiang Province of China (E2015024; LH2019E010).

DATA AVAILABILITY STATEMENT

Data cannot be made publicly available readers should contact the corresponding author for details.

REFERENCES

Asefa, T., Kemblowski, M., McKee, M. & Khalil, A. 2006 Multi-time scale stream flow predictions: the support vector machines approach. *Journal of Hydrology (Amsterdam)* **318** (1–4), 0–16.

Ashrafi, M., Chua, L. H. C., Quek, C. & Qin, X. 2017 A fully-online neuro-fuzzy model for flow forecasting in basins with limited data. *Journal of Hydrology* **545**, 424–435.

Danandeh Mehr, A., Kahya, E. & Olyaie, E. 2013 Streamflow prediction using linear genetic programming in comparison with a neuro-wavelet technique. *Journal of Hydrology* **505** (Complete), 240–249.

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

Huang, N. E., Shen, Z., Long, S. R., Wu, M. C., Shih, H. H., Zheng, Q., Yen, N. C., Tung, C. C. & Liu, H. 1998 The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. *Proceedings A* **454** (1971), 903–995.

Huang, Y., Schmitt F, G., Lu, Z. & Liu, Y. 2009 Analysis of daily river flow fluctuations using empirical mode decomposition and arbitrary order Hilbert spectral analysis. *Journal of Hydrology* **373** (1–2), 103–111.

Huang, S., Chang, J., Huang, Q. & Chen, Y. 2014 Monthly streamflow prediction using modified EMD-based support vector machine. *Journal of Hydrology* **511**, 764–775.

Humphrey, G. B., Gibbs, M. S., Dandy, G. C. & Maier, H. R. 2016 A hybrid approach to monthly streamflow forecasting: integrating hydrological model outputs into a Bayesian artificial neural network. *Journal of Hydrology* **540**, 623–640.

Karthikeyan, L. & Nagesh Kumar, D. 2013 Predictability of nonstationary time series using wavelet and EMD based ARMA models. *Journal of Hydrology* **502**, 103–119.

Kisi, O. & Cimen, M. 2011 A wavelet-support vector machine conjunction model for monthly streamflow forecasting. *Journal of Hydrology (Amsterdam)* **399** (1–2), 132–140.

Lall, U. & Sharma, A. 1996 A nearest neighbor bootstrap for resampling hydrologic time series. *Water Resources Research* **32** (3), 679–693.

Li, J., Chen, L., Xia, S., Xu, P. & Liu, F. 2014 A complete ensemble empirical mode decomposition for GPR signal time-frequency analysis. In *Radar Sensor Technology XVIII. International Society for Optics and Photonics*.

Maity, R., Bhagwat, P. P. & Bhatnagar, A. 2010 Potential of support vector regression for prediction of monthly streamflow using endogenous property. *Hydrological Processes* **24** (7), 917–923.

Nayak, P. C., Sudheer, K. P. & Rangan, D. M. 2004 A neuro-fuzzy computing technique for modeling hydrological time series. *Journal of Hydrology* **291** (1–2), 52–66.

Niu, M., Wang, Y., Sun, S. & Yongwu, L. 2016 A novel hybrid decomposition-and-ensemble model based on CEEMD and

GWO for short-term PM2.5 concentration forecasting. *Atmospheric Environment* **134**, 168–180.

Partal, T. K. 2007 Wavelet and neuro-fuzzy conjunction model for precipitation forecasting. *Journal of Hydrology* **342** (1–2), 199–212.

Qiang, F., Linqi, L. & Mo, L. 2018 An interval parameter conditional value-at-risk two-stage stochastic programming model for sustainable regional water allocation under different representative concentration pathways scenarios. *Journal of Hydrology* **564**, 115–124.

Ravansalar, M., Rajaee, T. & Kisi, O. 2017 Wavelet-linear genetic programming: a new approach for modeling monthly streamflow. *Journal of Hydrology* **549**, 461–475.

Sang, Y. F., Wang, Z. & Liu, C. 2012 Period identification in hydrologic time series using empirical mode decomposition and maximum entropy spectral analysis. *Journal of Hydrology* **424–425**, 154–164.

Shiri, J. & Kisi, O. 2010 Short-term and long-term streamflow forecasting using a wavelet and neuro-fuzzy conjunction model. *Journal of Hydrology (Amsterdam)* **394** (3–4), 486–493.

Sivapragasam, C. & Vasudevan, G. 2010 Genetic programming model for forecast of short and noisy data. *Hydrological Processes* **21** (2), 266–272.

Sudheer, K. P., Gosain, A. K. & Ramasastri, K. S. 2010 A data-driven algorithm for constructing artificial neural network rainfall-runoff models. *Hydrological Processes* **16** (6), 1325–1330.

Suykens, J. A. K. & Vandewalle, J. 1999 Least squares support vector machine classifiers. *Neural Processing Letters* **9** (3), 293–300.

Tan, Q. F., Lei, X. H., Wang, X., Wang, H., Wen, X., Ji, Y. & Kang, A.-Q. 2018 An adaptive middle and long-term runoff forecast model using EEMD-ANN hybrid approach. *Journal of Hydrology* **567**, 767–780.

Tang, C. Y., Guan, X. W. & Zhang, S. M. 2008 *The Advanced Methods for mid-Long Term Hydrological Forecasting and its Application*. (In Chinese). China Water Power Press.

Torres, M. E., Colominas, M. A., Schlotthauer, G. & Flandrin, P. 2011 A complete ensemble empirical mode decomposition with adaptive noise. In *2011 IEEE International Conference on Acoustics, Speech, and Signal Processing*. pp. 4144–4147, doi:10.1109/ICASSP.2011.5947265.

Vapnik, V. N. 2000 *The Nature of Statistical Learning Theory*. Springer.

Wu, Z. A. & Huang, N. E. 2009 Ensemble empirical mode decomposition: a noise assisted data analysis method. *Advances in Adaptive Data Analysis* **1**, 1–41.

Ye, L. & Cheng-You, T. 2011 Application of nearest neighbor bootstrapping regressive model in the dry season monthly runoff forecast. *Water Sciences and Engineering Technology* (6), 14–17.

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.

First received 13 July 2020; accepted in revised form 15 November 2020. Available online 27 November 2020