Desalination and Water Treatment
www.deswater.com
doi: 10.5004/dwt.2019.24034

154 (2019) 147–159
June

Regression models (SVR, EMD and FastICA) in forecasting water quality of the Haihe River of China

Nan Liang[a], Zhihong Zou[a], Yigang Wei[a,b,*]

[a]*School of Economics and Management, Beihang University, No. 37 Xueyuan Road, Haidian District, Beijing 100083, China, Tel. +86 15901019164; Fax: +86 82315854; email: weiyg@buaa.edu.cn (Y. Wei), Tel. +86 18911688596; Fax: +86 68218354; email: liangnan@buaa.edu.cn (N. Liang), Tel. +86 13671038207; Fax: +86 82315854; email: zouzhihong@buaa.edu.cn (Z. Zou)*
[b]*Beijing Key Laboratory of Emergency Support Simulation Technologies for City Operation*

Received 24 July 2018; Accepted 25 February 2019

ABSTRACT

Dissolved oxygen (DO) concentration is an important indicator for monitoring water quality. A reliable prediction of DO concentration has significant implications for safeguarding human health and promoting environmental sustainability. A novel model that employs support vector regression (SVR) algorithm combined with empirical mode decomposition (EMD) and fast independent component analysis (FastICA) noise reduction is proposed to ensure the accuracy of DO forecasting. On the basis of weekly data from 2008 to 2013, the model is applied to the Sanchakou section of the Haihe River in China to forecast DO levels. A host of traditional forecasting models was also studied and compared to derive the best-performing model. The mean absolute percentage error (MAPE) and maximum relative error (MRE) were used as the criteria to assess the models. The estimation results indicate that the MAPE and MRE obtained from the model developed in this study were 27% and 1.35%, respectively. Compared with other traditional methods, including SVR, SVR based on FastICA and SVR based on EMD, the estimation results reveal that the proposed model can improve forecasting accuracy and forecast time in the case of an extreme DO situation. The comparison clearly points to the enormous potential of SVR based on EMD and FastICA to provide early warning of an extreme DO situation and forecast a specific value for a specific week. Thus, the model can be considered a viable alternative for promoting an effective environmental protection of the Haihe River.

Keywords: Water quality forecasting; Dissolved oxygen (DO); Empirical mode decomposition (EMD) algorithm; Fast independent component analysis (FastICA) algorithm; Noise reduction; Support vector regression (SVR) algorithm

1. Introduction

Water is a precious natural resource with strategic economic importance. It serves as the basic element of ecological environment and is the most essential resource for socio-economic development [1]. China is one of the countries in the world that is facing a serious shortage of water resources. Although China possesses 7% of global water resources, the country's water resources per capita only account for 25% of the world's average [2]. In a survey of more than 600 Chinese cities, two-thirds of cities suffered from inadequate water supplies, and every one in six cities experienced severe water shortages [3]. Moreover, the surface water environment is widespread and heavily polluted. To overcome these challenges, China has established a national surface water environmental monitoring network for regular water quality monitoring. Among the 749 sections of the 406 rivers monitored by the Data Center of Ministry of Environmental Protection of the People's Republic of China in August 2013, 69% of the rivers were Class I to III water resources, 23% were Class IV water resources and 8% were

* Corresponding author.

Class V water resources. The statistics indicate that 69% of Class I to III water resources (69%) met the water quality standard for drinking water. The trend of water quality deterioration due to pollution has not yet been fundamentally reversed [4]. The data indicate that water quality in the seven biggest rivers in China varies greatly. Fig. 1 shows that the water quality of Zhujiang River is the best, followed by Chang Jiang River, Huang He River, Huaihe River, Liaohe River and Songhuajiang River ranking second to sixth in terms of water quality, but these rivers are slightly polluted. Meanwhile, Haihe River has the worst water quality and is also the most seriously polluted. Chemical oxygen demand (COD), biological oxygen demand (BOD), nitrate (NH_3), permanganate and DO are the main indicators for monitoring water quality. These indicators in the seven biggest rivers are all below the national water sanitary criteria. As the worst polluted river, Haihe River has attracted intense attention for promoting effective pollution treatment.

Dissolved oxygen (DO) refers to the amount of oxygen that is dissolved in water, and it is widely recognised as the representative parameter for measuring the quality of water in rivers and streams [5]. DO concentration is usually employed as an indicator to measure water quality for the following reasons: (1) DO concentration is highly related to a host of water quality indicators, such as COD, BOD, NH_3 and permanganate, and thus is capable of reflecting the general quality of a water body to some extent [5]. (2) DO is also an indication of the self-purification capacity of a water body. The self-purification processes of a water body entail the consumption of oxygen, thereby diminishing the DO concentration in water. If a water body is heavily polluted, then the decomposition of organic matter will consume oxygen, which results in a dramatic reduction of DO concentration. In addition, DO concentration is an important parameter in understanding how well the water can support aquatic plant and animal life. Fish will not survive if DO concentration in a water body is below 2 mg L^{-1} [6]. (3) If DO level is above the saturation value (7.5 mg L^{-1}), then the water body is possibly under eutrophication [6].

As mentioned above, high DO levels generate a significant and an adverse impact on the ecological environment. Therefore, monitoring and forecasting the fluctuations in DO levels are critically important. Accurate forecasting of DO levels has been widely studied in the disciplines of water source planning, protection and treatment. However, DO concentration is influenced by a wide variety of factors, such as environmental uncertainty, human disturbance and climate change, among others [7]. Thus, changes in DO levels show the characteristics of irregularity, high nonlinearity and non-smoothness, which add to the difficulty of DO forecasting [8].

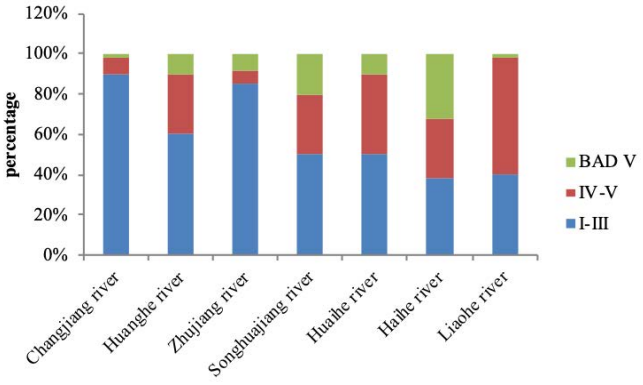

Fig. 1. Water quality condition of the seven rivers in China.

The purpose of the present study is to develop a reliable forecasting model for measuring DO concentration and predicting the advent of extreme DO concentration events. Support vector regression (SVR) forecasting algorithm based on empirical mode decomposition (EMD)–fast independent component analysis (FastICA) noise reduction is proposed for the forecasting of DO concentration. The model is capable of predicting the specific date on which an extreme situation is going to happen (i.e., DO concentration is above or below the normal standard). By forecasting the specific DO level and specific date of its occurrence, the model can provide early warning for authorities to adopt measures and avoid extreme consequences of precarious DO levels. The EMD–FastICA noise reduction is relatively new in water quality forecasting and has been applied with SVR for the first time in this study to forecast DO levels.

The rest of this paper is organised as follows: Section 2 provides an extensive literature review on water quality forecast. Section 3 presents data and research methods. Section 4 outlines the structure, flow and principle of the model. Section 5 discusses the prediction results. The last section summarises and concludes the paper.

2. Literature review

A large number of time-series-based forecasting models have been used in water quality prediction in recent years [9–12]. Specifically, these models include spatial autocorrelation [13], time sequence [14], artificial neural network (ANN) method [15], regression analysis [16] and support vector machine [17–22].

An extensive literature review was conducted to summarise the pros and cons of these models (summarised in Table 1). (1) Palani et al. [23] developed a neural network model to forecast the amount of DO in seawater. The ANN method has self-adaptive estimation of input–output responses without a predefined mathematical model, and it is effective in dealing with dynamic or nonlinear water data [24]. However, this method has several key shortcomings, such as being local optimal, over-fitting and having generalisation ability defects [25]. (2) Meryem et al. [16] performed an analysis based on a time sequence method to choose a suitable predictive method. A time sequence method is virtually based on rigid assumptions, such as linearity, normality and independence among predictor variables. However, the strong prerequisites limit its application in the real world, and the method cannot effectively depict a complex nonlinear relationship in water resource management [26–28]. (3) Li [29] indicated that the water simulation model is suitable for small-sized data application and can properly reflect uncertain changes of water data. However, its forecasting accuracy is low and the model is difficult to be established. (4) Partalas et al. [30] studied a greedy ensemble selection family of algorithms for ensembles of regression models to

Table 1
Comparison of different predictive methods

Predictive methods	Reference	Applicable condition	Accuracy	Advantages	Defects
Regression analysis	Partalas et al. [30]	Linear or nonlinear regression on a large sample	High	Good generalisation ability and convergence for complex or high-dimensional models	Weak fault Tolerant ability
Artificial neural network	Kariniotakis et al. [25]	Dynamic and uncertain water environment	High	Self-adaptively estimates linear or nonlinear relationships	Local optimal and overfitting problems
Time series method	Meryem et al. [16]	Exponential change or small sample	Low	Based on sophisticated theoretical basis	Hardly depicts complex nonlinear relationship
Water simulation model	Li [29]	Simulation prediction of small data	Low	Reflects uncertain change of data	Difficult to establish a model

solve the forecasting of water quality. Xiang and Jiang [24] employed the least squares support vector machine and particle swarm optimisation model to predict the quality of a drinking water source. Regression analysis, such as SVR, performs excellently in dealing with the nonlinear mapping problem. The model is theoretically straightforward with the ease of using short training time and high-precision characteristics [31,32]. SVR can effectively minimise the empirical uncertainty of training samples and overcome the defects of overfitting and local optimal solution [33–37]. Li [29] compared various statistical models for forecasting and concluded that SVR performs the best among the available approaches.

Given that input–output response for DO data is nonlinear complex and of high dimensionality, the key problem for prediction is to properly and self-adaptively depict the input–output nonlinear relationship. The extensive literature review concludes that ANN and regression analysis are more applicable for nonlinear mapping problems and big data samples than time sequence method and water simulation model. Moreover, the multiple nonlinear regression analysis is more suitable than the ANN method in effectively forecasting the changing trends in DO levels. Therefore, SVR is the best-performing multiple nonlinear regression method for DO data prediction.

However, the method also has limitations. Water quality data are irregularly distributed and influenced by multiple factors. If all data used to train the model in SVR have the same weight, then local fluctuation and randomness among data will still influence the training model and further weaken forecasting accuracy [38,39]. Many researchers have endeavoured to overcome the shortcomings of SVR and enhance its forecasting accuracy.

The composite forecasting model is an effective approach. EMD was employed to self-adaptively decompose water quality data into a series of functions with different frequencies. Such data transformation can help explore the frequency distribution of functions. SVR forecasting on each function can reduce the forecasting error of inflection points [40–43]. However, high-frequency functions from the EMD decomposition results are considered noise signals [44]. In addition, it will influence the forecasting accuracy of SVR. To effectively solve the problem, FastICA was employed to transform groups of experiment data into independent components and combined with SVR to enhance its forecasting accuracy [45–49]. The FastICA transformation results still contain many noise signals and the forecasting error rate remains high, especially on inflection points. The forecasting results are still unsatisfactory. Hence, the key problem is to further study the benefits of EMD and FastICA for data processing, which will help overcome limitations of SVR.

3. Data and research methods

3.1. Study area

Haihe River is approximately 1,090 km long and is one of the 10 longest rivers in China. It flows through Tianjin City, which is a megacity with a population of more than 10 million. The main stream of the river is the one that feeds into the sea, and it performs comprehensive functions of drainage, water storage, water supply, shipping and environmental protection for Tianjin City [50]. It plays an important role in North China as it helps in environmental protection and provides economic significance. However, water in the Haihe River has been seriously polluted. Among all the 64 monitoring sections, only 42% of the water in the Haihe River falls under Class I to Class III, 28% is categorised as Class IV to Class V and 30% is worse than Class V (National Surface Water Quality Report of August 2013). Fig. 2 shows the section numbers in the Haihe River, which is worse than Class V in terms of COD, BOD, NH_3, permanganate and DO. Water pollution has seriously threatened the lives of the residents along its course.

DO concentration is very low around the Sanchakou section of the Haihe River in summer and high in winter due to a reduction in rainfall, and its minimum value can even reach 1 mg L^{-1} due to the presence of various pollutants. In rainy and flood seasons, the concentration of DO usually drops significantly and the self-purification capability of the Haihe River drops dramatically because rainwater flushes into the river's course. However, in winter, much of the organic matter pollutants are discharged into the river and cannot be diluted in a timely manner in the same season. Algae and aquatic plants grow fast, and a strong photosynthesis produces more oxygen, resulting in a sharp rise in

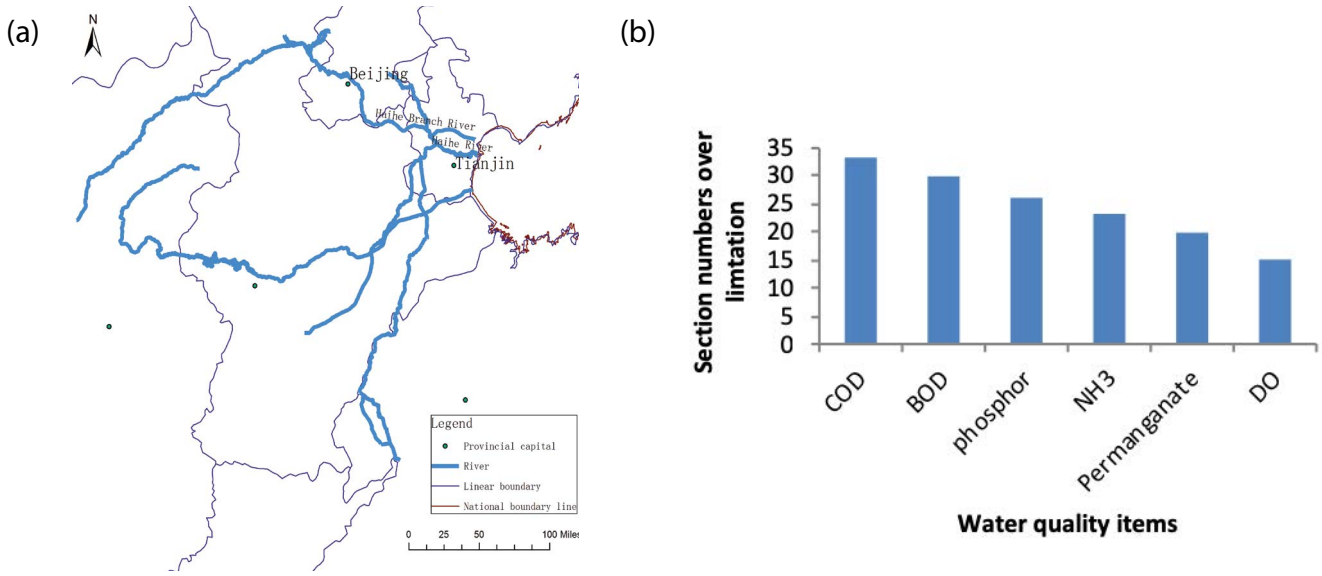

Fig. 2. (a) Geographic location and (b) water quality of Haihe River.

DO levels. Although the saturation of DO in water is usually 7.5 mg L^{-1}, the actual measured value in the river often surpasses the saturation value and reaches 22 mg L^{-1} in winter. Ultrahigh DO levels also mean that the water quality is bad [51–56]. Effectively dealing with the water pollution problem is vitally important to the economy, society and environment around the Haihe River.

3.2. Research data

The DO concentration data around the Sanchakou section of the Haihe River was obtained from the Data Center of Ministry of Environmental Protection of the People's Republic of China. The weekly time series data cover the period from the 33rd week in 2008 to the 29th week in 2013, consisting of a total of 256 data sets. The raw data are shown in Fig. 3. The figure shows several patterns: (1) the average value of DO concentration was generally constant at 7.56 mg L^{-1} and it was higher than the saturation standard (7.5 mg L^{-1}) because of organic matter pollution. (2) DO concentration fluctuated dramatically and irregularly. The maximum and minimum values were different per year and swept across a wide range and DO level in each season showed a wide variation. (3) An extreme value of DO in water frequently occurred. In general, the DO level reached its maximum value in winter (6th week of 2009, 8th week of 2010, 3rd week of 2011, 5th week of 2012 and 1st week of 2013). The highest DO concentration can reach 18 mg L^{-1}, which is far higher than the normal water quality standard, clearly showing that the Haihe River is heavily polluted. The DO concentration was the lowest in summer (41st week of 2009, 34th week of 2010 and 31st week of 2011 and 2012) and decreased to 1 mg L^{-1}, which is far beyond the highest level of poor-quality water (Class V). A dramatic fluctuation reflected significant changes in water quality. In addition, predicting the extreme value and specific time when this occurs is essential to deploy suitable measures for environmental protection and economic development.

3.3. Research methods

This section introduces the underlining rationale and calculation process of EMD, FastICA and SVR algorithms. Fig. 4 shows the relations and steps of model development: (1) First, EMD and FastICA achieve the functions of data noise reduction. (2) Processed data are used to train the model for forecasting future DO levels. (3) Finally, the forecasted results are compared and assessed.

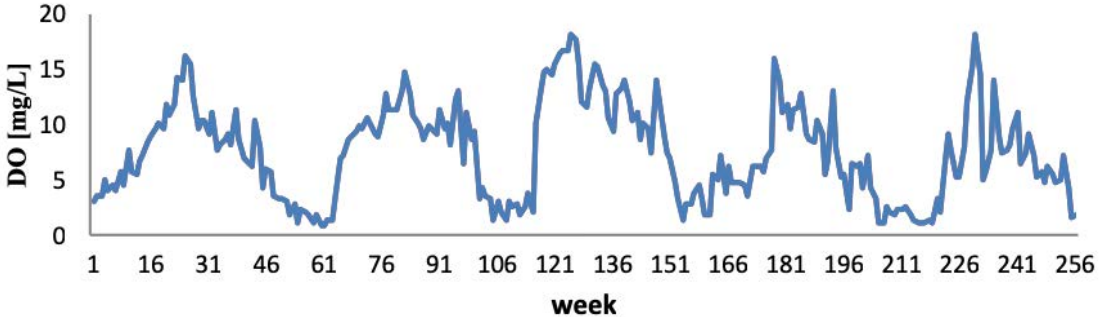

Fig. 3. DO concentration of 256 weeks around Sanchakou section of the Haihe River.

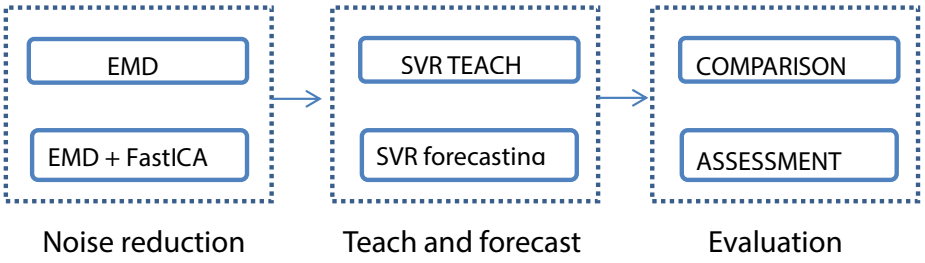

Fig. 4. Algorithms in the model.

3.3.1. EMD algorithm

EMD decomposes the data into a series of intrinsic mode functions (IMFs) with different frequencies.

The specific decomposing process is explained as follows:

(1) Use cubic spline functions to fit the upper and lower enveloping curve of source signal $s(t)$ according to all maximum and minimum points in the IMF.
Set $m_i(t)$ as the average value of the upper and lower enveloping curves.
Set $h_1(t)$ as a new data series that $m_i(t)$ is subtracted from $s(t)$.
Repeat the above process and get $\mathrm{IMF}_1(t) = \sum m_i(t)$ if $h_1(t)$ satisfies the IMF specified conditions.

(2) Repeat step 1. A series of $\mathrm{IMF}_i(t)$ and the rest of the indecomposable function $r(t)$ are obtained until $r(t)$ is smaller than the set value or turns into a monotonic function:

$$s(t) = \sum_{i=1}^{n} \mathrm{IMF}_i(t) + r(t) \tag{1}$$

where $\mathrm{IMF}_i(t)$ is the value of IMF at time t.

3.3.2. FastICA algorithm

ICA aims at separating input signals and making components of output Y independent. Non-Gaussianity criterion is utilised to predict the mutual independence of output components, which means that the separation process is finished when its non-Gaussianity reaches the maximum value. The process is shown in Fig. 5. The input component of S is assumed to be random and mixed independent signal X is separated by the B system. Thus, output signal Y is close to S. Entropy is applied to measure non-Gaussianity in FastICA, and one direction is obtained to ensure that Y has the biggest non-Gaussianity through transformation $Y = W^T X$.

3.3.3. SVR algorithm

The linear regression function $f(x) = wx + b$ can be utilised to fit (x_i, y_i) with $x_i \in R^n$ as input and $y_i \in R^n$ as output,

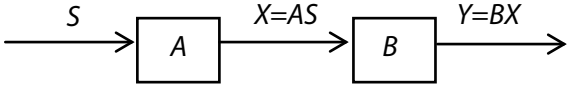

Fig. 5. ICA process.

where w and b need to be calculated. Assume that all training data are fitted by the linear function with an error of ε. Import slack variables ξ_i and ξ_i^*, and then w and b can then be determined by solving the minimum of optimisation function as follows:

$$\begin{aligned} &L = \frac{1}{2}\omega \cdot \omega + C\sum_{i=1}^{n}\left(\xi_i + \xi_i^*\right) \\ &y_i - f\left(x_i\right) \le \varepsilon + \xi \\ &f\left(x_i\right) - y_i \le \varepsilon + \xi^* \\ &\xi_i, \xi_i^* \ge 0 \end{aligned} \tag{2}$$

The above optimisation function is in a quadratic form with a linear constraint condition. It can be solved by the Lagrange multiplier. If Lagrange multiplier α_i, α_i^*, γ_i, γ_i^* is imported, then function L is changed into Eq. (3):

$$\begin{aligned} L = {} & \frac{1}{2}\omega \cdot \omega + C\sum_{i=1}^{n}\left(\xi_i + \xi_i^*\right) - \sum_{i=1}^{n}\alpha_i\left[\xi_i + \varepsilon - y_i + f\left(x_i\right)\right] - \\ & \sum_{i=1}^{n}\alpha_i^*\left[\xi_i^* + \varepsilon - y_i + f\left(x_i\right)\right] - \sum_{i=1}^{n}\left(\xi_i\gamma_i + \xi_i^*\gamma_i^*\right) \end{aligned} \tag{3}$$

The minimisation of function L to ω, ξ_i and ξ_i^* and the maximisation of function L into α_i, α_i^*, γ_i, γ_i^* can be calculated and imported into Eq. (3), achieving the maximisation function in dual form as shown below:

$$\begin{aligned} W\left(\alpha_i, \alpha_i^*\right) = {} & \frac{1}{2}\sum_{i=1,j=1}^{n}\left(\alpha_i - \alpha_i^*\right)\left(\alpha_j - \alpha_j^*\right)\left(x_i \cdot x_j\right) + \\ & \sum_{i=1}^{n}\left(\alpha_i - \alpha_i^*\right)y_i - \sum_{i=1}^{n}\left(\alpha_i - \alpha_i^*\right)\varepsilon \end{aligned} \tag{4}$$

The corresponding constraint condition is as follows:

$$\sum_{i=1}^{I}\left(\alpha_i - \alpha_i^*\right) = 0, \quad 0 \le \alpha_i,\ \alpha_i^* \le C \tag{5}$$

Variable b can be calculated through Eqs. (3) and (4):

$$b = y_i - \frac{1}{2}\sum_{i=1,j=1}^{n}\left(\alpha_i - \alpha_i^*\right)x_i x_j - \varepsilon \tag{6}$$

Linear fitting function can be obtained as follows:

$$f\left(x\right) = wx + b = \sum_{i=1}^{n}\left(\alpha_i - \alpha_i^*\right)x_i x + b \tag{7}$$

The main idea of SVR is to map an input vector into a high-dimensional feature space (Hilbert space) through the nonlinear map and perform linear regression in a high-dimension space. SVR maps input vector x into a high-dimensional feature space, employing nonlinear function $f(x) = K(x) + b$ to fit data (x_i, y_i). Thus, Eq. (4) turns into

$$W\left(\alpha_i,\alpha_i^*\right)=\frac{1}{2}\sum_{i=1,j=1}^{n}\left(\alpha_i-\alpha_i^*\right)\left(\alpha_j-\alpha_j^*\right)K\left(x_i\cdot x_j\right)+\sum_{i=1}^{n}\left(\alpha_i-\alpha_i^*\right)y_i-\sum_{i=1}^{n}\left(\alpha_i-\alpha_i^*\right)\varepsilon \tag{8}$$

Correspondingly, the nonlinear fitting function can be obtained as follows:

$$f\left(x\right)=\sum_{i=1}^{n}\left(\alpha_i-\alpha_i^*\right)K\left(x_i,x\right)+b \tag{9}$$

4. Development of the model

4.1. Step one: EMD noise reduction

The EMD algorithm can decompose DO data into functions of different frequencies. Power spectrums of IMFs can show frequency distribution of IMFs. The highest frequency is the noise signal, and its range is low. If it is eliminated, then the data will be more regular and forecasting accuracy will be enhanced. Therefore, the IMFs of the highest frequency should be estimated as noise signals and eliminated in the first stage.

4.2. Step two: EMD–FastICA noise reduction

Real DO data are statistically independent of noise signals. Given that FastICA can transform IMFs into independent components by calculating non-Gaussianity, it may be implemented on IMFs to effectively separate noise signals and real DO data. To indicate the component containing noise, Kurtosis coefficient can be employed. EMD can be utilised again to decompose the indicated component. After decomposition, the IMF of the highest frequency can be extracted and eliminated, completing noise reduction at the second stage.

4.3. Step three: SVR training and forecasting

After the EMD–FastICA noise reduction procedure is completed, the SVR algorithm is employed to train models among the output IMFs and the changing trend is forecasted on the final stage.

The combined EMD–FastICA algorithm can extract more noise signals and will not significantly change the range of the original signal. EMD is self-adaptive to extract the high-frequency noise signal, which is efficient in performing preliminary noise reduction. To further extract the noise signal, FastICA is used to separate noise and real DO data and EMD extract the highest-frequency noise signal again. Consecutive noise reduction efficiently eliminates noise signals of comparatively high frequencies, and the processed data become smoother and regular. Correspondingly, the SVR model will be trained convincingly and forecasting errors can be effectively decreased.

On the basis of the above analysis, the water quality forecasting model presented in Fig. 6 is developed in seven basic steps:

(1) Decomposition of the water quality data series

EMD algorithm is employed to decompose water data $x(t)$ into a series of IMFs $\{\mathrm{IMF}_i(t)|t = 1,2,3,\ldots n;\ i = 1,2,3,\ldots d\}$:

$$x\left(t\right)=\sum_{i=1}^{d}\mathrm{IMF}_i\left(t\right) \tag{10}$$

(2) EMD noise reduction

Power spectrums of all functions are calculated, and the function of the highest frequency is selected and eliminated. After the calculation, $x(t)$ is converted into $x\left(t\right)=\sum_{i=2}^{d}\mathrm{IMF}_i\left(t\right)$.

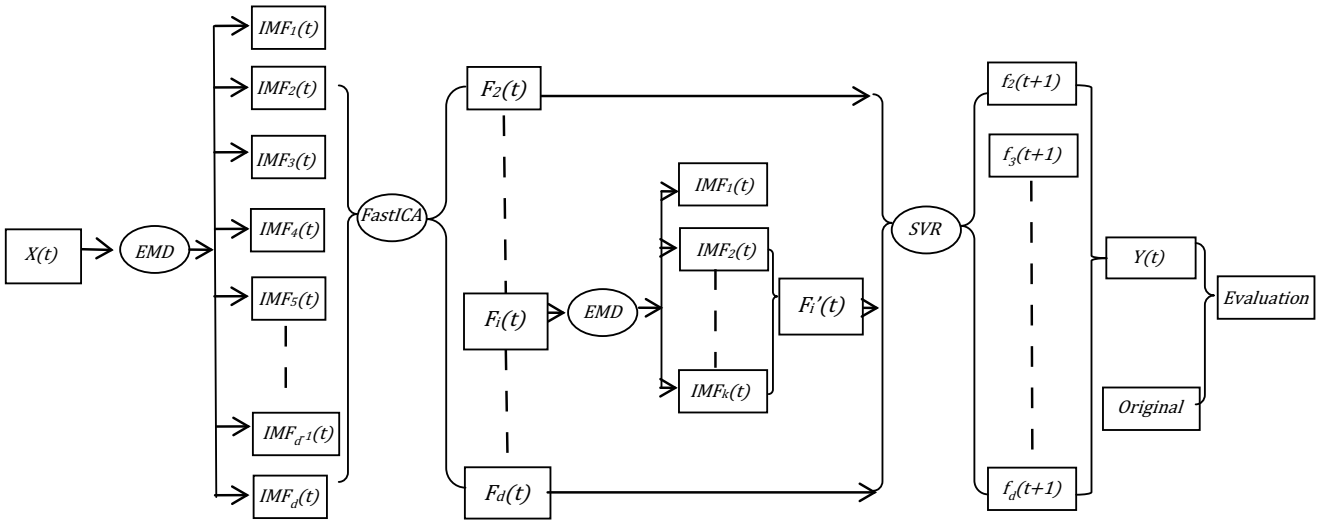

Fig. 6. Flow chart of the model.

(3) FastICA transformations

FastICA is employed to transform $d - 1$ principal components. The component with Kurtosis close to 0 is selected. The process is expressed by the following equation:

$$\left\{\mathrm{IMF}_i(t)\middle|i=2,3,\ldots d;\mathrm{t}=1,2,3,\ldots\mathrm{n}\right\}=W\left\{F_i(t)\middle|i=2,3,\ldots d;t=1,2,3,\ldots n\right\} \tag{11}$$

(4) EMD–FastICA noise reduction

The EMD algorithm is employed to decompose the component into a series of IMFs. The IMF of the highest frequency is selected and eliminated, completing the noise reduction procedure.

Hereafter, $\{\mathrm{IMF}_2\ldots\mathrm{IMF}_d\}$ are reconstructed by the processed components.

(5) SVR training

Based on the training of items of $\{\mathrm{IMF}_2\ldots\mathrm{IMF}_d\}$ and adaptive parameter adjustment, $d - 1$ SVR models are obtained.

(6) SVR forecasting

$d - 1$ trained SVR models are employed to forecast water quality at the $n + 1$ moment and derive $f_2(n + 1)$, $f_3(n + 1)\ldots$ $f_d(n + 1)$. The final result of $n + 1$ moment is obtained by summing up these data.

$$y(n+1)=\sum_{i=2}^{d}f_i(n+1) \tag{12}$$

(7) Evaluation of results

The mean absolute percentage error (MAPE), root mean square error (RMSE), maximum relative error (MRE) and maximum absolute error (MAE) are chosen to evaluate the forecasting results.

MAPE is the most common index of forecasting error, directly reflecting the MAE between the original and forecasting value (Eq. (13)):

$$E_{\mathrm{MAPE}}=\frac{1}{N}\sum\frac{\left|W_R-W_F\right|}{W_R} \tag{13}$$

$$E_{\mathrm{RMSE}}=\sqrt{\frac{1}{N}\sum\left|W_R-W_F\right|^2} \tag{14}$$

$$E_{\mathrm{MRE}}=\max\frac{W_R-W_F}{W_R} \tag{15}$$

$$E_{\mathrm{MAE}}=\max\left|W_R-W_F\right| \tag{16}$$

5. Results and discussion

5.1. Forecasting DO in Sanchakou section of the Haihe River

5.1.1. Step one: EMD noise reduction

The DO data in the Haihe River were decomposed into functions by the EMD algorithm, and eight-dimension functions from high to low frequencies are shown in Fig. 7. Then, the power spectrums of eight functions are analysed in Fig. 8. Among the functions, the first function covered a wide band of high frequencies (Fig. 8), but the range was low. This finding shows that the first function contained many noise signals and can be deleted.

5.1.2. Step two: EMD–FastICA noise reduction

The rest of the seven-dimension functions were then transformed to seven dimension-independent components

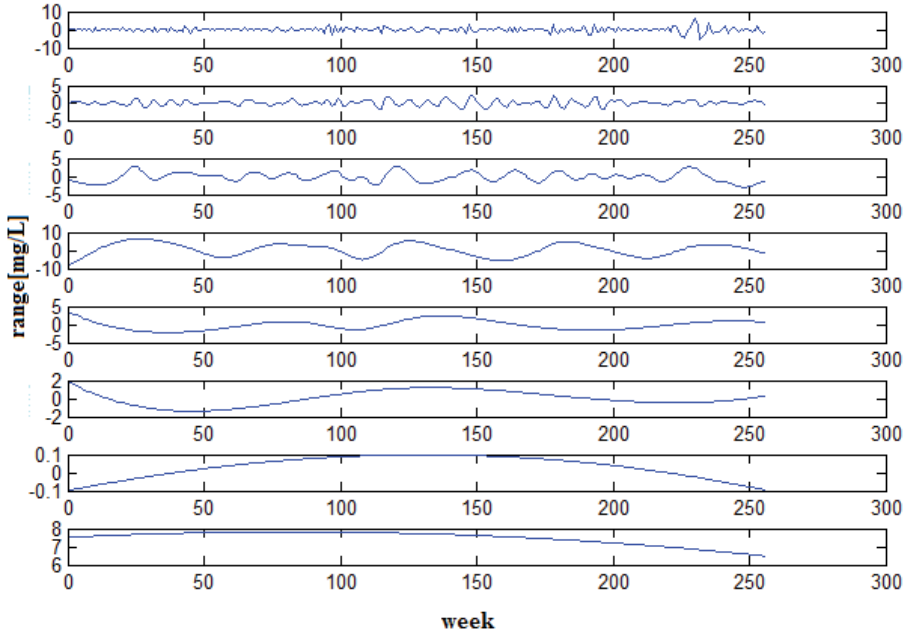

Fig. 7. EMD decomposition results.

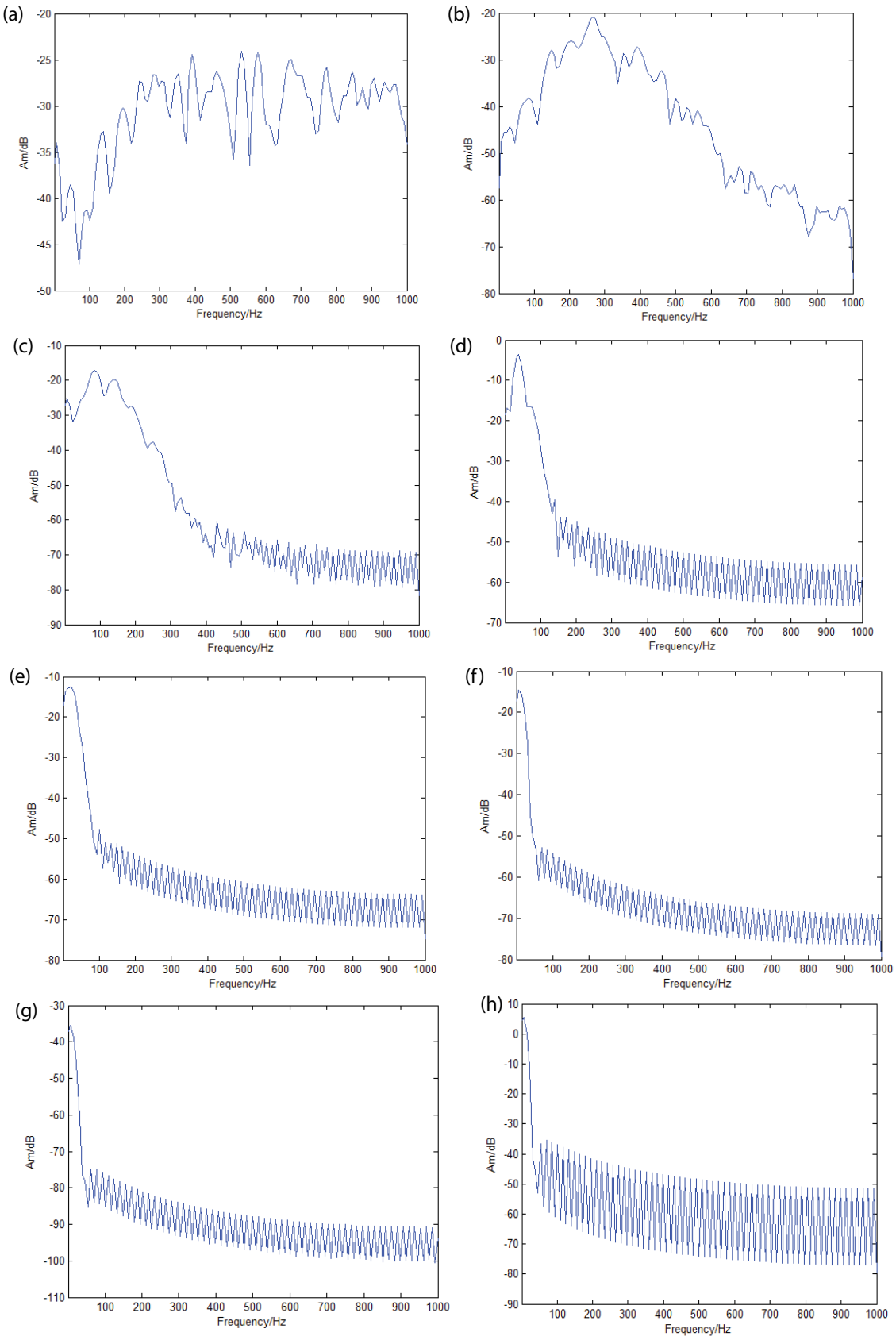

Fig. 8. Power spectrums of EMD decomposition results. Power spectrum of (a) IMF_1, (b) IMF_2, (c) IMF_3, (d) IMF_4, (e) IMF_5, (f) IMF_6, (g) IMF_7 and (h) IMF_8.

by FastICA (Fig. 9), further separating the noise signals. The power spectrums and Kurtosis coefficients of the seven components were analysed in Fig. 10 and Table 2, respectively. As shown by the power spectrum results in Fig. 10, components 5, 7 and 8 contain high noise signals, whereas the Kurtosis coefficient result in Table 2 meant that component 8 was mainly composed of noise signals. Thus, noise in component 8 can be reduced through the EMD decomposition again. The power spectrums proved that the first of the eight-dimension functions contained many noise signals and can be deleted. After deletion, component 8 was reconstructed by the rest of the functions, indicating that the noise reduction calculation was completed. Input DO data were regrouped by components 2–8.

5.1.3. Step three: SVR training

Given China's environment administrative authority conducting weekly water quality monitoring of big rivers, rich water data were provided to train the SVR model. The study was designed to employ one season's data (12 weeks of data) for training the model and forecasting the 13th week's data. SVR was applied to train and forecast components 2–8. By training for 200 times, the first 212 data sets of each component established seven teaching sets for model training and achieved seven trained models.

5.1.4. Step four: SVR forecasting

Of the 256 data sets obtained from the Data Center of Ministry of Environmental Protection of the People's Republic of China, 44 data sets were employed to establish the forecasting set. In accordance with the seven models, forecasting can be implemented and 7 × 44 forecasting data were obtained for all the models. By adding up the seven-dimension data, 44 final forecasting data sets were finally achieved.

5.2. Comparison of the forecasting results

To compare the forecasting results and verify the performance of the model, a series of prediction experiments were conducted. Single SVR algorithm was first implemented to train the model and forecast the DO data. In the second experiment, the DO data were decomposed into functions by the EMD algorithm. Then, SVR was applied to train and forecast the data of each function. Forecasting data of each function were added up to one-dimensional data, and the final forecasting data were achieved. In the third experiment, the EMD decomposition functions of the second experiment were transformed into multi-dimension components by the FastICA algorithm. Then, SVR was applied to train and forecast the data of each component.

The results of the forecasting curves are shown in Fig. 11. The results of the experiment were better than those of SVR, SVR based on FastICA or SVR based on EMD. MAPE and RMSE help determine the forecasting performance of all the methods, and its comparison results of the four algorithms are shown in Table 3. The MAPE and RMSE of SVR based on EMD–FastICA were 27.5% and 2.19%, respectively, and were far below the other three algorithms by 27%, 28.9%, 44%, 26.3%, 24% and 42.2%, respectively. Thus, the model can be recognised as the best-performing forecast model with the least prediction errors. The forecasting results are also more consistent with the actual observed values when compared with the forecast results from other models.

Furthermore, MRE and MAE indicate the local deviation degree between the original and forecasting values. The

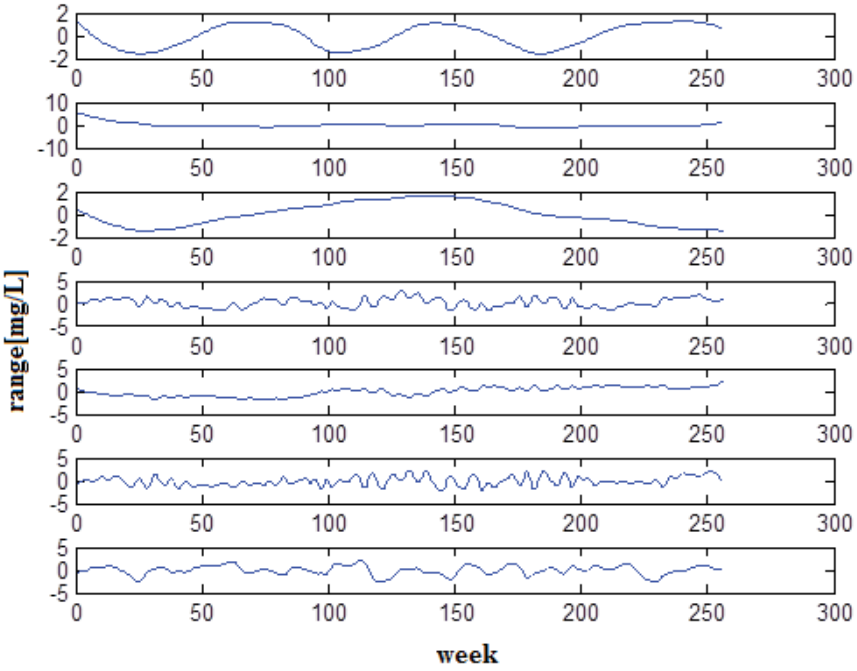

Fig. 9. FastICA transformation results after the first EMD noise reduction.

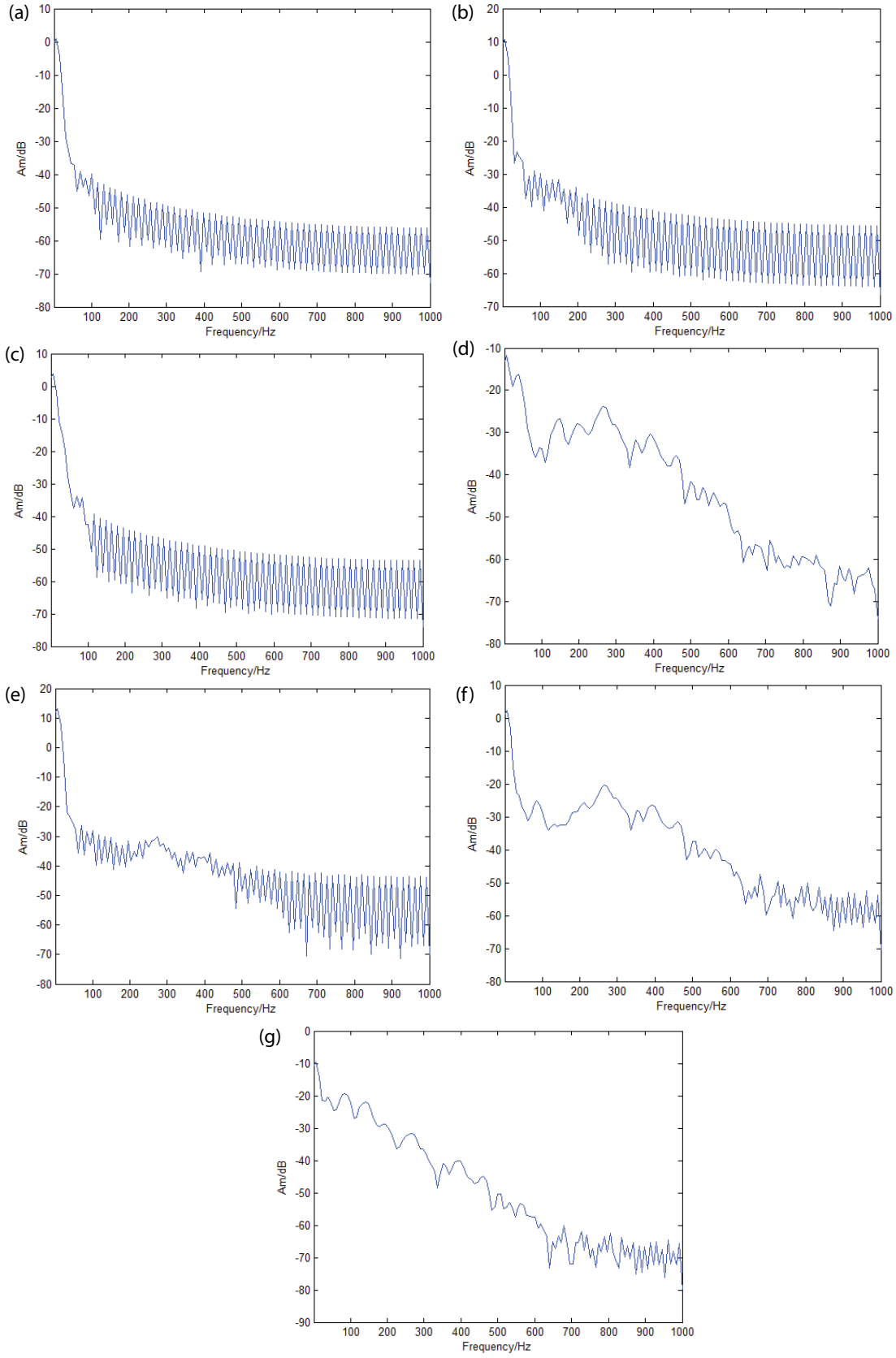

Fig. 10. Power spectrums of FastICA transformation results. Power spectrum of (a) IMF_2, (b) IMF_3, (c) IMF_4, (d) IMF_5, (e) IMF_6, (f) IMF_7 and (g) IMF_8.

MRE and MAE of SVR based on EMD–FastICA were 1.35 and 6.59, which were far below the other three algorithms by 11.2%, 29.9%, 37.4%, 8.7%, 6.252% and 46.6%. In other words, the model can best predict the trend in local changes of DO data. Fig. 11 also shows that forecasting results of SVR based on EMD–FastICA were nearly the same as the original value when the DO level is below 2 mg L^{-1}. When the DO level is higher than 6 mg L^{-1}, the forecasting results of SVR based on EMD–FastICA were close to the original value, but the difference is not large. However, SVR, SVR–EMD and SVR–FastICA cannot forecast all the wave summits and valleys and lag behind the trend of DO changes of the original signal.

In addition, the forecasting results reflect every up and down change, and the position of the maximum or minimum value coincided with the original position. These results demonstrate that SVR based on EMD–FastICA can provide early warning and forecast for the specific week when the DO value changes drastically and specific value when DO level surges or drops.

Table 2
Kurtosis coefficient analysis on FastICA components

Index	Kurtosis
F2	−1.52786
F3	9.401198
F4	−1.30199
F5	−0.84692
F6	−1.24776
F7	−0.51681
F8	0.14695

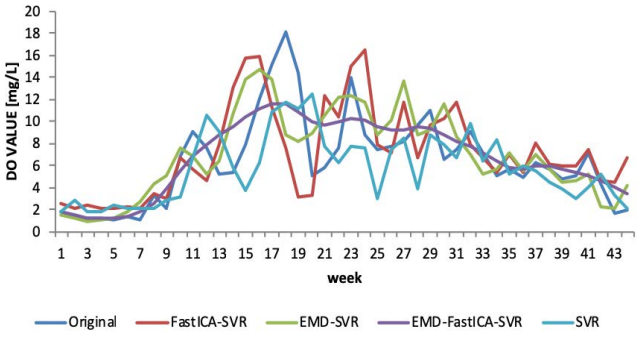

Fig. 11. Comparison of the forecasting results.

Table 3
Comparison of the forecasting results

Index	SVR based on EMD–FastICA	SVR based on EMD	SVR`	SVR–FastICA
MAPE	0.2754	0.3782	0.3938	0.4914
RMSE	2.19	2.97	2.88	3.79
MRE	1.35	1.44	1.48	2.53
MAE	6.59	9.4	7.42	10.52

6. Conclusions

This study has explored the potentials of the SVR method combined with EMD–FastICA noise reduction in water quality forecasting for the Haihe River in China. For long-term forecasting, the results clearly indicate that the proposed model's results are very much close to real values. A comparison of the forecasting results with SVR, SVR based on FastICA and SVR based on EMD shows that the model performs better and can be a viable alternative in projecting future water quality. The empirical results demonstrate that the model is a powerful tool for automatic monitoring systems of water quality and the forecasting results can give an early warning about extreme events of DO concentrations. The data generated from this model can be very helpful and supplementary for framing suitable environmental protection policy. In accordance with the forecast results, some policy recommendations are provided: (1) DO level is strongly related with COD, BOD, NH_3, permanganate and DO indices. Therefore, the establishment of a DO level monitoring and forecasting system is strongly suggested as it can be helpful and supplementary in framing environmental policies and undertaking early warning and emergency response measures. (2) A combination of a monitoring and forecasting systems is significant for the environmental protection of water.

Acknowledgments

The study is funded by the National Natural Science Foundation of China (No. 51478025), Beijing Science and Technology Plan (No. Z161100005016037) and MOE (Ministry of Education in China) Project of Humanities and Social Sciences (No. 18YJC840041). The authors gratefully acknowledge the support of the Beijing Key Laboratory of Emergency Support Simulation Technologies for City Operations. In addition, the authors wish to thank the anonymous reviewers for the insightful comments that helped us improve the quality of this paper.

References

[1] Y. Wei, C. Huang, J. Li, L. Xie, An evaluation model for urban carrying capacity: a case study of China's mega-cities, Habitat Int., 53 (2016) 87–96.
[2] Y. Jiang, China's water scarcity, J. Environ. Manage., 90 (2009) 3185–3196.
[3] S.G. Li, Research on carrying capacity of urban water resource and its adjusting method (PhD dissertation), Peking University, Beijing, China, 2003 (in Chinese).
[4] L.N. Zheng, R.B. Lin, Analysis of water Pollution in China in Recent Years, Guide of Sci-tech Magazine, Vol. 5, 2012, p. 246 (in Chinese).
[5] X.X. Song, H.W. Yan, B.H. Tian, Analysis of pollution in Haihe River and its conventional indicators, South-to-North Water Divers, Water Sci. Technol., 10 (2012) 98–101 (in Chinese).
[6] Y.H. Jia, S.X. Liu, Analysis and discussion on the dissolved oxygen pollution index of environmental quality assessment, China Chem. Trade, 6 (2012) 213–213 (in Chinese).
[7] Y.C. Chen, F.U. Jian, Z.W. Liu, Analysis of the variety and impact factors of dissolved oxygen downstream of Three Gorges Dam after the impoundment, Adv. Water Sci., 20 (2009) 526–530.
[8] J. Huan, X. Liu, Dissolved oxygen prediction in water based on *K*-means clustering and ELM neural network for aquaculture, Trans. Chin. Soc. Agric. Eng., 32 (2016) 174–181.

[9] B. Bai, B. Yoo, X.Q. Deng, I. Kim, D.H. Gao, Linking routines to the evolution of IT capability on agent-based modeling and simulation: a dynamic perspective, J. Comput. Math. Organ. Theory, 22 (2016) 184–211.
[10] S.Y. Liu, L.Q. Xu, Y. Jiang, D.L. Li, Y.Y. Chen, Z.B. Li, A hybrid WA–CPSO-LSSVR model for dissolved oxygen content prediction in crab culture, Eng. Appl. Artif. Intell., 29 (2014) 114–124.
[11] J.S. Matos, E.D.R. Sousa, Prediction of dissolved oxygen concentration along sanitary sewers, Water Sci. Technol., 34 (1996) 525–532.
[12] K.P. Singh, S. Gupta, P. Rai, Predicting dissolved oxygen concentration using kernel regression modeling approaches with nonlinear hydro-chemical data, Environ. Monit. Assess., 186 (2014) 2749–2765.
[13] W.L. Li, Z.N. Wei, G.Q. Sun, Multi-interval wind speed forecast model based on improved spatial correlation and RBF neural network, Electr. Power Autom., 29 (2009) 89–92.
[14] S. Pritpal, B. Bhogeswar, An efficient time series forecasting model based on fuzzy time series, Eng. Appl. Artif. Intell., 26 (2013) 2443–2457.
[15] S. Heddam, Modeling hourly dissolved oxygen concentration (DO) using two different adaptive neuro-fuzzy inference systems (ANFIS): a comparative study, Environ. Monit. Assess., 186 (2014) 597–619.
[16] O. Meryem, J. Ismail, E.M. Mohammed, A Comparative Study of Predictive Algorithms for Time Series Forecasting, IEEE 5th International Conference on Information Science and Technology (ICIST), Changsha, China, 2015, pp. 68–73.
[17] L. Chen, J.M. Liu, X.X. Liu, Application of support vector machine in the ground water quality evaluation, J. Northwest A&F Univ. (Nat. Sci. Ed.), 8 (2010) 221–226.
[18] P.J. García Nieto, E. García-Gonzalo, J.R. Alonso Fernández, C. Díaz Muñiz, Hybrid PSO–SVM-based method for long-term forecasting of turbidity in the Nalón river basin: a case study in Northern Spain, Ecol. Eng., 73 (2014) 192–200.
[19] H. Huang, W.X. Lu, Assessment of water quality based on support vector machine model, Water Saving Irrig., 2 (2012) 57–63.
[20] X.C. Liang, Y.B. Gong, D. Xiao, Novel method for water quality prediction based on multi-kernel weighted support vector machine, J. Southeast Univ., 41 (2011) 14–17.
[21] S. Roghayeh, M.Z. Rahm, S. Karim, V.D. Patrick, Use of support vector machines (SVMs) to predict distribution of an invasive water fern *Azolla filiculoides* (Lam.) in Anzali wetland, southern Caspian Sea, Iran, Eco. Modell., 244 (2012) 117–126.
[22] H. Yang, S.P. Gu, M.D. Cui, Forecast of short-term wind speed in wind farms based on GA optimized LS-SVM, Power Syst. Prot. Control., 39 (2011) 44–48.
[23] S. Palani, S.Y. Liong, P. Tkalich, Development of a neural network model for dissolved oxygen in seawater, Indian J. Mar. Sci., 38 (2009) 151–159.
[24] Y.R. Xiang, L.Z. Jiang, Water Quality Prediction Using LS-SVM With Particle Swarm Optimization, IEEE 2nd International Workshop on Knowledge Discovery and Data Mining, Moscow, Russia, 2009, pp. 900–904.
[25] G.N. Kariniotakis, G.S. Stavrakakis, E.F. Nogaret, Wind power forecasting using advanced neural networks models, IEEE Trans. Energy Convers., 11 (1996) 762–767.
[26] P. Fang, R.H. Shao, Q.Y. Si, J. Ren, The application of least squares support vector machine regression in water quality forecast of Xi'an Ba River, Syst. Eng., 6 (2011) 113–116.
[27] T. He, P. Chen, Prediction of Water-quality Based on Wavelet Transform Using Vector Machine, IEEE: 2010 Ninth International Symposium on Distributed Computing and Applications to Business, Engineering and Science (DCABES), HongKong, 2010, pp. 76–81.
[28] X. Wang, J. Lv, D. Xie, A Hybrid Approach of Support Vector Machine With Particle Swarm Optimization for Water Quality Prediction, IEEE 2010 5th International Conference on Computer Science & Education (ICCSE), Hefei, China, 2010, pp. 1158–1163.
[29] R.Z. Li, Advance and trend analysis of theoretical methodology for water quality forecast, J. Hefei Univ. Technol., 29 (2006) 26–30.
[30] I. Partalas, G. Tsoumakas, E.V. Hatzikos, I. Vlahavas, Greedy regression ensemble selection: theory and an application to water quality prediction, Inform. Sci., 178 (2008) 3867–3879.
[31] K. Hojat, K. Sohrab, R. Mohammad, F. Saeed, Predicting discharge coefficient of triangular labyrinth weir using support vector regression, support vector regression-firefly, response surface methodology and principal component analysis, Flow. Meas. Instrum., 55 (2017) 75–81.
[32] L. Tang, A.Y. Wang, Z.J. Xu, J. Li, Online-purchasing behavior forecasting with a firefly algorithm-based SVM model considering shopping cart use, Eurasia J. Math. Sci. Technol. Ed., 13 (2017) 7967–7983.
[33] E. Ceperic, V. Ceperic, A. Baric, A strategy for short-term load forecasting by support vector regression machines, IEEE Trans. Power Syst., 28 (2013) 4356–4364.
[34] F. Chen, B. Tang, R. Chen, A novel fault diagnosis model for gearbox based on wavelet support vector machine with immune genetic algorithm, Measurement, 46 (2013) 220–232.
[35] W.C. Hong, Traffic flow forecasting by seasonal SVR with chaotic simulated annealing algorithm, Neurocomputing, 74 (2011) 2096–2107.
[36] C.N. Ko, C.M. Lee, Short-term load forecasting using SVR (support vector regression)-based radial basis function neural network with dual extended Kalman filter, Energy, 49 (2013) 413–422.
[37] R.J. Liao, H.B. Zheng, G. Stanislaw, L.J. Yang, Particle swarm optimization-least squares support vector regression based forecasting model on dissolved gases in oil-filled power transformers, Electr. Power Syst. Res., 84 (2011) 2074–2080.
[38] H.M. Sheng, J. Xiao, Electric vehicle state of charge estimation: nonlinear correlation and fuzzy support vector machine, J. Power Sources, 281 (2015) 131–137.
[39] Z. Wun, H. Zhang, J.H. Liu, A fuzzy support vector machine algorithm for classification based on a novel PIM fuzzy clustering method, Neurocomputing, 125 (2014) 119–124.
[40] X.J. Liu, Z.Q. Mi, Q.X. Yang, Wind speed forecasting based on EMD and time series analysis, Acta Energiae Solaris Sinica, 31 (2010) 1037–1041.
[41] Q. Ouyang, W.X. Lu, X. Xin, Y. Zhang, W.G. Cheng, T. Yu, Monthly rainfall forecasting using EEMD-SVR based on phase-space reconstruction, Water Resour. Manage., 30 (2016) 2311–2325.
[42] N. Ramesh Babu, B. Jagan Mohan, Fault classification in power systems using EMD and SVM, Ain Shams Eng. J., 8 (2017) 103–111.
[43] L. Ye, P. Liu, Combined Model Based on EMD-SVM for Short-Term Wind Power Prediction, Proc. CSEE., 31 (2011) 102–108.
[44] H.Q. Li, X.F. Wang, L. Chen, E.B. Li, Denoising and R-peak detection of electrocardiogram signal based on EMD and improved approximate envelope, Circuits Syst. Signal Process., 33 (2014) 1261–1276.
[45] D. Boutte, B. Santhanam, A Feature Weighted Hybrid ICA-SVM Approach to Automatic Modulation Recognition, IEEE 13th Digital Signal Processing Workshop and 5th IEEE Signal Processing Education Workshop, Marco Island, United States, 16 (2009) 399–403.
[46] J.G. Chen, Z.X. Zhang, Z.G. Guo, Application of independent component analysis in empirical mode decomposition, J. Vib. Shock, 28 (2009) 109–111.
[47] P. Nirmal Kumar, H. Kareemullah, EEG Signal with Feature Extraction Using SVM and ICA Classifiers, IEEE: International Conference on Information Communication and Embedded Systems (ICICES2014), Tokyo, Japan, 2014, pp. 1–7.
[48] B. Sun, Short-Term Wind Speed Forecasting Based on FastICA Algorithm and Improved LSSVM Model, Proc. CSU-EPSA., 26 (2014) 22–27.
[49] H. Shao, X.H. Shi, L. Li, Power signal separation in milling process based on wavelet transform and independent component analysis, Int. Int. J. Mach. Tools Manuf., 51 (2011) 701–710.

[50] J. Li, Prevention and cure on Haihe valley's pollution, J. Changchun Normal Univ., 10 (2008) 20.
[51] H.Y. Wang, J. Teng, J. Zhang, Reason analysis of the high DO in winter of Haihe River in Tianjin, Urban Environ. Urban Ecol., 18 (2005) 27–28 (in Chinese).
[52] Y.G. Wei, C. Huang, P.T.I. Lam, Y. Sha, Y. Feng, Using urban-carrying capacity as a benchmark for sustainable urban development: an empirical study of Beijing, Sustainability, 7 (2015) 3244–3268.
[53] Y. Li, Y.G. Wei, S.Q. Shan, Y. Tao, Pathways to a low-carbon economy: estimations on macroeconomic costs and potential of carbon emission abatement in Beijing, J. Cleaner Prod., 199 (2018) 603–615.
[54] Y.G. Wei, C. Huang, P.T.I. Lam, Z.Y. Yuan, Sustainable urban development: a review on urban carrying capacity assessment, Habitat Int., 46 (2015) 64–71.
[55] Y.G. Wei, Y. Li, M.Y. Wu, Y. Li, The decomposition of total-factor CO_2 emission efficiency of 97 contracting countries in Paris agreement, Energy Econ., 2019 (78) 365–378.
[56] Y.G. Wei, Z.C. Wang, H.W. Wang, T. Yao, Y. Li, Promoting inclusive water governance and forecasting the structure of water consumption based on compositional data: a case study of Beijing, Sci. Total Environ., 634 (2018) 407–416.