scientific reports

OPEN

A novel RF-CEEMD-LSTM model for predicting water pollution

Jinlou Ruan, Yang Cui✉, Yuchen Song & Yawei Mao

Accurate water pollution prediction is an important basis for water environment prevention and control. The uncertainty of input variables and the nonstationary and nonlinear characteristics of water pollution series hinder the accuracy and reliability of water pollution prediction. This study proposed a novel water pollution prediction model (RF-CEEMD-LSTM) to improve the performance of water pollution prediction by combining advantages of the random forest (RF) and Long short-term memory (LSTM) models and Complementary ensemble empirical mode decomposition (CEEMD). The experimental results based on measured data show that the proposed RF-CEEMD-LSTM model can accurately predict water pollution trends, with a mean ab-solute percentage error (MAPE) of less than 8%. The RMSE of the RF-CEEMD-LSTM model is reduced by 62.6%, 39.9%, and 15.5% compared to those of the LSTM, RF-LSTM, and CEEMD-LSTM models, respectively, proving that the proposed method has good advantages in predicting non-linear and nonstationary water pollution sequences. The driving force analysis results showed that TN has the most significant impact on water pollution prediction. The research results could provide references for identifying and explaining water pollution variables and improving water pollution prediction method.

With the rapid advancement of industrialization and urbanization, water pollution problems have become increasingly serious in China[1]. Accurate and reliable water environment prediction models can provide real-time water pollution concentration change information, enabling health risks to be avoided in a timely manner and providing an intuitive reference for environmental protection departments. Water environments are nonlinear, nonstationary, and noisy systems, making water environment prediction difficult[2]. How to deeply explore and extract the laws contained in the concentration series of water pollutants and accurately predict the change trends of water pollutants in the future has become a difficult and urgent problem to solve.

Numerical models and data-driven models are mainly used to predict the concentration of pollutants in water environments[3]. Numerical models can simulate the development of pollutants and predict the quality of a water environment[4]. Fu et al.[5] analyzed the characteristics and challenges of existing water quality models and found that in terms of model parameter calibration, although a comprehensive calibration plan has been established[6], there are still difficulties in combining the model with on-site or laboratory observation results. Due to limited data, high-dimensional models, and overreliance, the numerical models may not be able to accurately capture all functional properties of the water quality variables of interest, resulting in significant difficulties in calibrating the model[5]. This uncertainty of model parameters makes it difficult for the constructed model to accurately simulate the potential relationship between input and output variables. Data-driven models do not require complex fluid dynamics theories and complex processes[7]. They can effectively explore the potential relationship between input variables and target variables by utilizing a large amount of historical monitoring data, and have superior applicability compared to numerical models[8]. Among them, artificial neural networks (ANNs) models are widely used in pollutant prediction due to their excellent ability to learn linear and nonlinear information from historical data[9]. Rustam et al.[10] used ANN in water quality and water quantity predictions, and the results showed that the accuracy of water quality prediction was 0.96, verifying the feasibility of artificial neural networks in water quality prediction. Najwa Mohd Rizal et al.[11] compared the performance of regression models, support vector machines, and ANNs in water quality prediction, and the results demonstrated that the ANN model was superior to the other models. Although these data-driven models can achieve good predictive performance, a single machine learning model is susceptible to overfitting and often fall into local optima[12].

In response to this problem, scholars have attempted to use combination forecasting models to predict water pollution. The composite model skillfully combines multiple models, aiming to solve the defects of a single model[13]. Common combination models include the residual processing model[14], weight combination model[15], and data decomposition model[16]. The residual processing model improves the prediction accuracy by processing the residuals of the prediction results, but it does not change the scope of application of a single model[17].

Henan Provincial Communications Planning and Design Institute Co., Ltd, Zhengzhou 450000, People's Republic of China. ✉email: cuiyangsg126@126.com

Therefore, it has great limitations in dealing with highly noisy, nonstationary and nonlinear systems. The weight combination model improves the accuracy and stability by assigning appropriate weights to each submodel to offset the residual prediction results. One weight combination model is often only applicable to specific data[18]. Water pollution data have significant nonlinear and nonstationary characteristics, leading to the phenomenon of high training accuracy but low verification accuracy when using weighted combination models to predict water pollution. Recent studies have shown that the empirical mode decomposition (EMD) is an effective data preprocessing method, which can decompose the original time series data into multiple subsequences with different frequencies, enabling the regular information contained in the data to be fully recognized and extracted, and is widely used in sequence prediction[16]. Hybrid models coupling EMD and machine learning tools have been commonly used in the water environment fields[19]. To improve the accuracy of prediction methods, Zhang et al.[20] used empirical mode decomposition (EMD) to preprocess the data and then used LSTM to predict water quality indicators, and found the performance of the hybrid models was superior to that of the single model. Due to the strong dependence of EMD on signal frequency, amplitude, and their differences, mode mixing often occurs during data decomposition[21]. The ensemble EMD (EEMD) method is an improved form of EMD that can overcome the modal mixing[22]. Eze et al.[21] developed a new combined prediction method using EEMD and LSTM neural networks to improve the accuracy of water quality parameter prediction, and found that the performance of the hybrid model is superior to similar water quality parameter prediction models.

One major concern of EEMD is that the introduction of noise assisted analysis increases computational complexity and time consumption. Moreover, the introduction of noise has a certain degree of damage on the original signal, leading to potential uncertainty in the decomposition results[23]. The complementary ensemble empirical mode decomposition (CEEMD) proposed by Yeh et al.[24] effectively overcomes these difficulties. CEEMD decomposes nonlinear and nonstationary sequence data into multiple components and residual terms by introducing complementary white noise, reducing the impact of residual noise while allowing outlier data to potentially play positive roles[25]. Nevertheless, for short-term water pollution time series with nonlinear and nonstationary traits, whether or to what degree, the hybrid model coupling CEEMD and deep learning models can improve the prediction accuracy remains unclear.

Selecting predictive variables is crucial in determining the performance and accuracy of the model, and recent research suggests using other techniques to select predictive factors before constructing a water pollution prediction model[19]. However, most data-driven models directly use machine learning methods to predict subsequences, ignoring the impact of feature selection on model performance, resulting in significant deficiencies in the interpretation of water pollution causes. There are also significant shortcomings in the analysis of water pollution characteristics. The effective interpretation of water pollution characteristics is often an important basis for water pollution prevention and control.

Considering the above factors, this study attempts to propose a RF-CEEMD-LSTM model for different water quality indicators by combining the advantages of the RF, LSTM model and CEEMD. First, RF was used to analyze the importance of water quality, meteorology, and air quality indicators to water pollution. Second, CEEMD was used to decompose the water pollution prediction indicators into intrinsic mode function (IMF) components and residual terms. Third, the LSTM algorithm was used to construct a combined water pollution prediction model for different water quality prediction indicator components. Finally, the proposed model was used to predict pondus hydrogenii (PH), ammonia nitrogen (NH3-N), and dissolved oxygen (DO) at the Kangdian and Banqiao stations in the Huaihe River basin. RF-CEEMD-LSTM was compared with other models to verify the effectiveness of the model.

Materials and methods

Study area and data

The Huaihe River Basin (111°55′E–121°25′E, 30°55′N–36°36′N) is located in central and eastern China and is the third largest water system in China. Its main stream flows through Hubei, Henan, Anhui, and Jiangsu provinces, with a watershed area of 270,000 km^2. The average annual temperature in the Huaihe River basin is between 11 and 16 ℃, with the highest and lowest temperatures occurring in July and January, respectively, and the average annual precipitation is 920 mm[26]. Due to the developed economy in the area where the Huaihe River flows, the Huaihe River is heavily affected by artificial intervention, resulting in a low capacity to absorb pollution and relatively serious pollution. Since the 1980s, water pollution accidents in the Huaihe River have occurred frequently, with Zhoukou in Henan being the most serious[27]. The water pollution problem has seriously restricted the development of the economy in the basin. Exploring accurate methods for analyzing and predicting water pollution characteristics is of great significance for utilizing water resources, preventing water pollution incidents, and comprehensively managing the environment.

The water pollution data, meteorological data, and air quality data of Kangdian station and Banqiao station in the Huaihe River Basin were selected to establish the water pollution prediction model (Fig. 1). The water pollution data are daily monitoring data from January 1, 2021, to December 31, 2022, at the Kangdian and Banqiao stations, these data include pH, water temperature (W-Temp), chemical oxygen demand (COD_{Mn}), DO, NH3-N, total nitrogen (TN), total phosphorus (TP), electrical conductivity (EC), and turbidity (Turb). The water pollution data were obtained from the China Environmental Monitoring Station and the hydrological monitoring division of the Huaihe River Hydrology Bureau. The meteorological data, including daily temperature (A-Temp), wind speed (WS), and precipitation (Prep), from January 1, 2021, to December 31, 2022, were obtained from the Resource and Environmental Science and Data Center of the Chinese Academy of Sciences. The air quality data, including daily $PM_{2.5}$, SO_2, NO_2, and CO, from January 1, 2021, to December 31, 2022, were obtained from the national air quality real-time release platform.

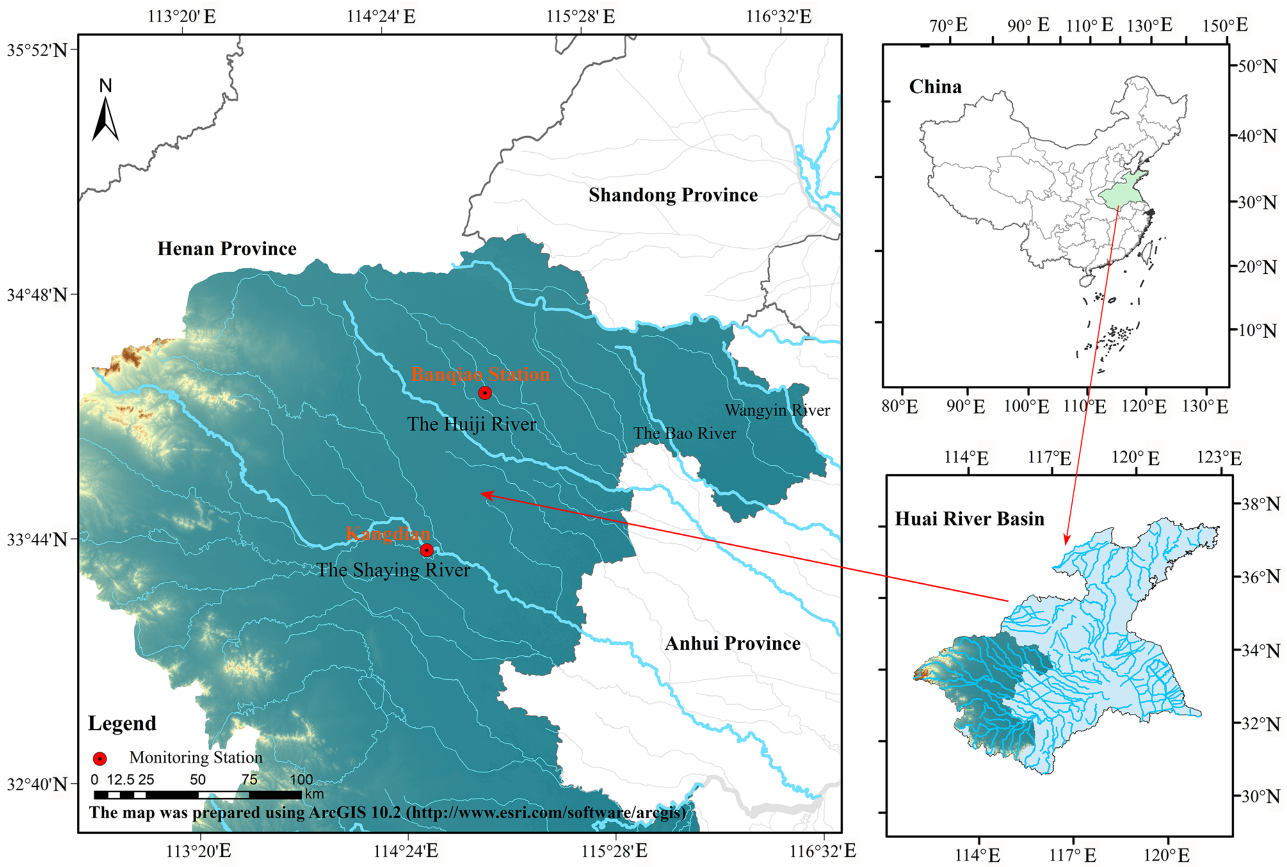


Figure 1. The location of the study area.

DO, pH, and NH3-N are important factors that affect the water environment quality and are also the focus of attention in water pollution prevention. The pH value is the negative logarithm of the hydrogen ion concentration in water, reflecting the degree of acid and alkali pollution of a water body. DO reflects the self-purification ability of a water body. A higher DO content indicates a strong self-purification ability of the water body. NH3-N is an important indicator that reflects the nutritional status of water bodies. NH3-N is present in water as free ammonia or ammonium salt and is the main oxygen-consuming pollutant in water bodies. Therefore, a water pollution prediction model was constructed with these three indicators.

Data filling

Due to the impact of monitoring station maintenance, there are a few missing data points in the sample. Therefore, the sample data were processed into continuous and complete sequence data. Cubic spline interpolation was used to process the missing data. Cubic spline interpolation can effectively overcome the limitations of polynomial interpolation and is widely used in the interpolation solution process. In cubic spline interpolation, the interval [a, b] is divided into n intervals [(a, x_1), (x_1, x_2), …, (x_{n-1}, b)], and the cubic spline interpolation results are calculated by solving matrix equations. The detailed mathematical description of cubic spline interpolation can be found in Chand et al.[28]. Twenty-three missing data points were processed using cubic spline interpolation (Eq. 1), accounting for 0.53% of the total sample data. The December 2022 data were selected as the test set, and the data from January 2021 to November 2022 were selected as the training set.

$$y = a_i + b_i x + c_i x^2 + d_i x^3. \tag{1}$$

a_i, b_i, c_i, d_i refers to the coefficients that need to be solved in each interval.

RF for feature selection

Input variables are the key factors that affect the performance of water pollution prediction. Quantifying the impact differences of water pollution indicators can explain the characteristics and causes of water pollution and provide a reference for water pollution prevention and control. But there is no unified paradigm for selecting and quantifying water pollution prediction indicators. In view of the above considerations, a method for selecting water pollution prediction indicators based on the RF algorithm was proposed in this study.

The RF algorithm is an integrated learning method based on decision trees[29]. In the RF algorithm, K training sets are randomly extracted using bootstrap resampling technology, and K decision trees are trained to form the random forest[30]. A RF has the advantages of simple modeling and strong generalizability[31] and exhibits better performance than that of individual decision trees in many classification analyses[31] and prediction tasks[30]. RF can also be used as a feature selection method.

In the quantitative analysis of water pollution prediction features using RF algorithms, the out-of-bag (OOB) error was used to measure the features. OOB data refers to data that have not been sampled during the RF resampling process, which can be used to evaluate the performance of decision trees. The RF algorithm analyzes the importance of a feature by perturbing it. If the OOB error of the RF model decreases significantly after perturbing the feature, it indicates that the feature is of high importance. The detailed steps for the quantitative analysis of water pollution prediction features based on the RF algorithm[29] can be divided into four steps.

(1) Randomly extract n data from the initial dataset and generate K new training sets. The data that were not extracted constitute K OOB datasets.
(2) For each decision tree in the RF, use the corresponding OOB data to calculate the OOB error (eo_k).
(3) Randomly perturb the characteristics of the OOB datasets $x_i, i = 1, 2, ..., M$ and calculate the OOB error (eo_{ki}).
(4) Calculate the feature importance of each feature.

k refers to the number of decision trees, eo_k refers to the OOB error of the kth decision tree, eo_{ki} refers to the OOB error of the decision tree after perturbing the ith feature, and IM_i refers to the importance score of the ith feature.

After obtaining the feature importance of each input variable to the predicted variable, the variable perturbation method is used to select the fewest number of inputs that offer the best predictive power and the most interpretation of the water pollution prediction model.

CEEMD prediction sequences

Due to the significant nonstationarity and uncertainty of a pollutant sequence, it is difficult for a model to accurately capture all the characteristics of the sequence, resulting in poor fitting and prediction performance. Therefore, the CEEMD method was used to decompose water pollution prediction sequences into stable components and residual terms before establishing the prediction model. CEEMD is an improvement of the EMD method[24]. EMD can process sequence data into IMF and residual terms that vary in frequency and are relatively stable. But EMD often exhibits the phenomenon of modal aliasing when decomposing sequence data that contain a large amount of noise[32]. To solve this problem, scholars in signal research have proposed various improved sequence decomposition methods, including EEMD[22] and CEEMD. CEEMD reduces the phenomenon of modal aliasing and the number of iterations required for decomposition[33] by adding a white noise pair with opposite signs during the data decomposition process. The decomposition process of the water pollution prediction sequence using CEEMD[24] was shown in Eqs. (2)–(4).

(1) Add a white noise pair $\varepsilon_i(n)$, denoting the sign, to a given sequence of pollutant concentrations $x(n), n = 1, 2, ..., N$ (Eq. 2).

$$\begin{cases} x_i^+(n) = x(n) + \varepsilon_i^+(n) \\ x_i^-(n) = x(n) + \varepsilon_i^-(n) \end{cases} \tag{2}$$

(2) Use CEEMD to decompose each original water pollution sequence with white noise and obtain m IMF components and one residual component *Res* (Eq. 3).

$$\begin{cases} x_i^+ = \sum_{j=1}^{m} c_{ij}^+(n) \\ x_i^- = \sum_{j=1}^{m} c_{ij}^-(n) \end{cases} \tag{3}$$

c_{ij} refers to the jth modal component of the i-th sequence after CEEMD.

(3) Calculate the average of all IMF components to obtain the final modal component group $c_i(t)$ by using Eq. (4).

$$c_i(t) = \frac{1}{2m} \sum_{j=1}^{2m} c_{ij} \tag{4}$$

Water pollution prediction model based on RF-CEEMD-LSTM

Based on the feature importance analysis of water pollution indicators and modal decomposition of prediction sequences, the water pollution prediction model was constructed using the LSTM method for each decomposed sequence. LSTM is a very important recurrent neural network (RNN) in deep learning methods[34]. An RNN creates a loop by adding additional weights to the network, making its input dependent not only on the current input but also on previous inputs. RNNs often experience gradient disappearance and explosion in long sequence modeling, resulting in a significant decrease in prediction performance. LSTM effectively solves the long-term dependency problem of RNNs by introducing input gates, forget gates, and output gates to store and update cell states, achieving selective retention of the sequence information[35].

The structure of LSTM is shown in Fig. 2. The core of LSTM lies in the cell state and the "gate" structure[34]. The cell state can continuously transmit relevant information during sequence processing. The earlier information can be carried to later cells, overcoming the impact of short-term memory. The three "gate" structures pass through σ functions to process data and learn which information to retain or forget during training, overcoming the long-term dependence of information. The main principles of using LSTM[34] for water pollution prediction are shown in Eqs. (5)–(10).

First, the input gate determines how much new water pollution data information and information output by the previous layer can be transmitted to the cell state.

$$i(t) = \sigma[w_i \times (h_{t-1}, x_t) + b_i] \tag{5}$$

$$\tilde{C}_t = \tanh[w_c \times (h_{t-1}, x_t) + b_c] \tag{6}$$

$i(t)$ in Eq. (5) refers to the output value of the input gate, w_i and b_i refer to the weights and biases of the input gate, respectively, w_c and b_c refer to the weights and biases of the update cell, respectively, σ refers to the activation function (sigmoid), h_{t-1} refers to the output of the memory cells at time t − 1, x_t refers to the input at time t, and $\tilde{C}_t$ in Eq. (6) refers to the state of the cell to be updated.

When information from the input gate is passed to the forget gate, the forget gate determines how much real-time data information and information output by the previous layer will be discarded.

$$f_t = \sigma[w_f \times (h_{t-1}, x_t) + b_f] \tag{7}$$

f_t in Eq. (7) refers to the output of the forget gate and w_f and b_f refer to the weights and biases of the forget gate, respectively.

The update of the cell state determines the proportion in which past information and instant information are combined and transmitted to the new cell state.

$$c_t = i_t \times \tilde{c}_t + f_t \times c_{t-1} \tag{8}$$

c_t and c_{t-1} in Eq. (8) refer to the cell states at time t and time t − 1, respectively.

Finally, the output gate determines how much data information will be output from the cell state. The output data information is used as input to the new round of the model cycle.

$$o_t = \sigma[w_o \times (h_{t-1}, x_t) + b_o] \tag{9}$$

$$h_t = o_t \times \tan(c_t) \tag{10}$$

o_t in Eq. (9) refers to the output value of the output gate, w_o and b_o refer to the weights and biases of the output gate, and h_t in Eq. (10) refers to the output of the memory cells at time t.

The LSTM prediction models were constructed for all prediction sequences decomposed by CEEMD, and all prediction sequences were integrated to generate water pollution prediction results. Parameter optimization is an important step in constructing the LSTM prediction model. The mean square error (MSE) was selected as the loss function, and adaptive moment estimation (Adam) was used as the optimizer in this study. The Adam optimizer combines the advantages of the AdaGrad and RMSProp optimization algorithms, i.e., high computational efficiency and parameter interpretation[36]. The number of neurons, learning rate, iteration times, and sliding window step length of LSTM were optimized by Adam. When the error between the actual value and the predicted value met the accuracy requirements, the model was saved. Based on the characteristics of the data in this study and relevant research results, the allowable error was set to 0.001.

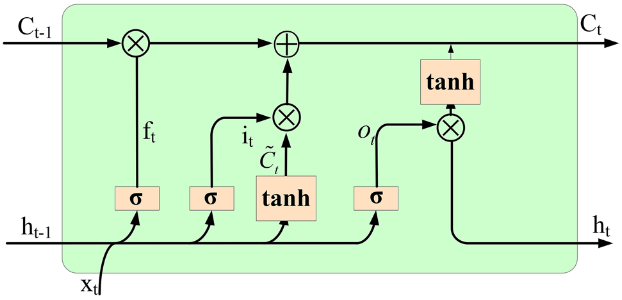


Figure 2. The principle of LSTM.

Schematic diagram of the proposed method

The RF algorithm was used for feature importance analysis to quantify the main indicators that affect water pollution prediction sequences in this study. Based on the results of feature importance analysis, an indicator set for different water pollution prediction sequences was proposed. CEEMD was used to reconstruct nonstationary and nonlinear prediction sequences into relatively stable components and residual terms for three types of water pollution prediction sequences from two stations. And LSTM was used to fit and predict trend components and integrate the results of all trend components to obtain the prediction model for different pollution sequences (Fig. 3). To verify the superiority of the proposed model, various other algorithms were used in this study for comparison, including the LSTM, RF-LSTM, and CEEMD-LSTM models.

Model performance evaluation

To evaluate the performance of the proposed RF-CEEMD-LSTM water pollution prediction model, four statistical indicators were selected to measure the prediction results: Nash–Sutcliffe efficiency (NSE), root mean square error (RMSE), mean absolute percentage error (MAPE), and mean absolute error (MAE). The MAPE (Eq. 13) and MAE (Eq. 14) can reflect the predicted errors in real-world scenarios. The RMSE (Eq. 12) is an evaluation index of the average error and volatility of the predicted results. The MAPE can measure the accuracy of time sequence prediction. The NSE (Eq. 10) can evaluate the fitting ability of the model. The closer the values of MAE, MAPE, and RMSE are to 0 and the closer the value of NSE is to 1, the better the prediction accuracy of the model. The indicator calculation methods are shown in Eqs. (11)–(14).

$$NSE = 1 - \frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{\sum_{i=1}^{n} (y_i - \overline{y}_i)^2} \tag{11}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{N}} \tag{12}$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n} \left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100\% \tag{13}$$

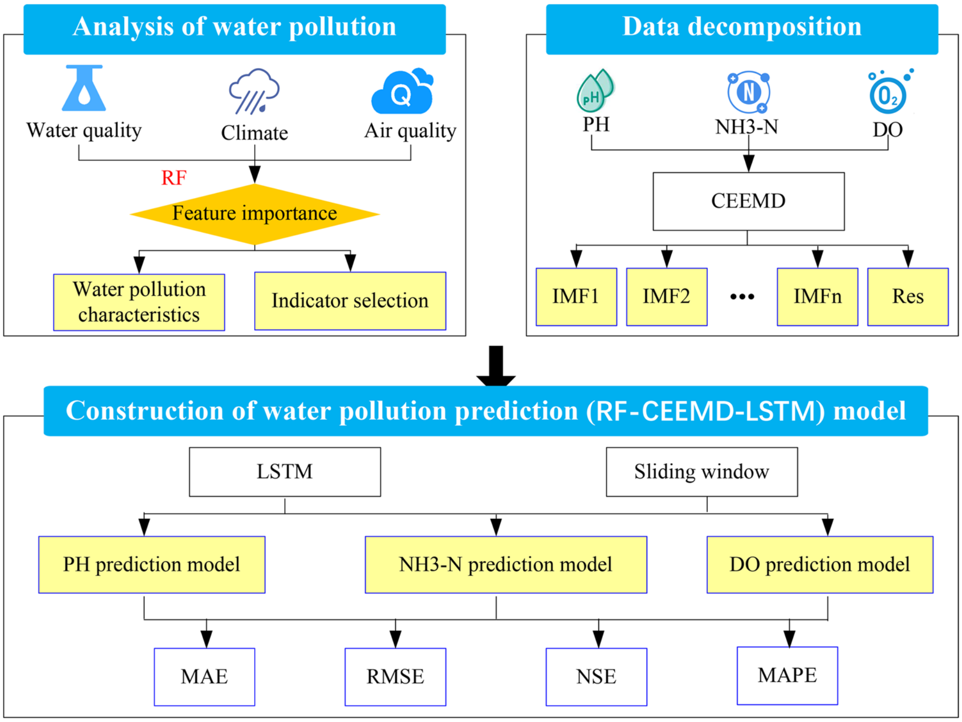


Figure 3. Modeling process of the RF-CEEMD-LSTM model.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{14}$$

y_i is the measured value, $\hat{y}_i$ is the predicted value of the water pollution, and n is the total number of validation samples.

Results

Characteristics of water pollution data

A descriptive statistical method was used to analyze the basic characteristics of the water pollution data used in this study to understand the quality and volatility of these data. As shown in Table 1, the missing rates of water pollution data at the two stations were relatively low, 0.37% and 0.69%. The main reason for missing data was that the stations were under maintenance. Cubic spline interpolation was used to process complete missing data in the sample to ensure sequence prediction continuity. Table 1 reflects the overall water pollution at the Kangdian and Banqiao stations. In terms of pH value, the average, maximum, and minimum values of the two stations were between 6 and 9, which conforms to the Class III water quality standard in China[37]. The minimum value of DO (0.83) and maximum value of NH3-H (9.67) at the Kangdian and Banqiao stations were far lower than the Class III water quality standard in China (DO > 5, NH3-H < 5)[37]. These results indicate that there is a certain degree of water pollution in the Huaihe River basin. In addition, Fig. 4 shows that the NH3-H and DO data of the two

Station	Feature	Mean value	Max value	Min value	Standard deviation	Missing data rate (%)
Kangdian	pH	7.91	8.64	7.38	0.24	0.35
	NH3-N (mg/L)	0.76	9.67	0.66	0.68	0.36
	DO (mg/L)	7.63	14.57	1.31	2.89	0.39
Banqiao	pH	8.17	9	7.11	0.31	0.61
	NH3-N (mg/L)	0.64	3.32	0.025	0.45	0.73
	DO (mg/L)	8.44	17.64	0.83	2.69	0.74

Table 1. Descriptive statistics of the water pollution data.

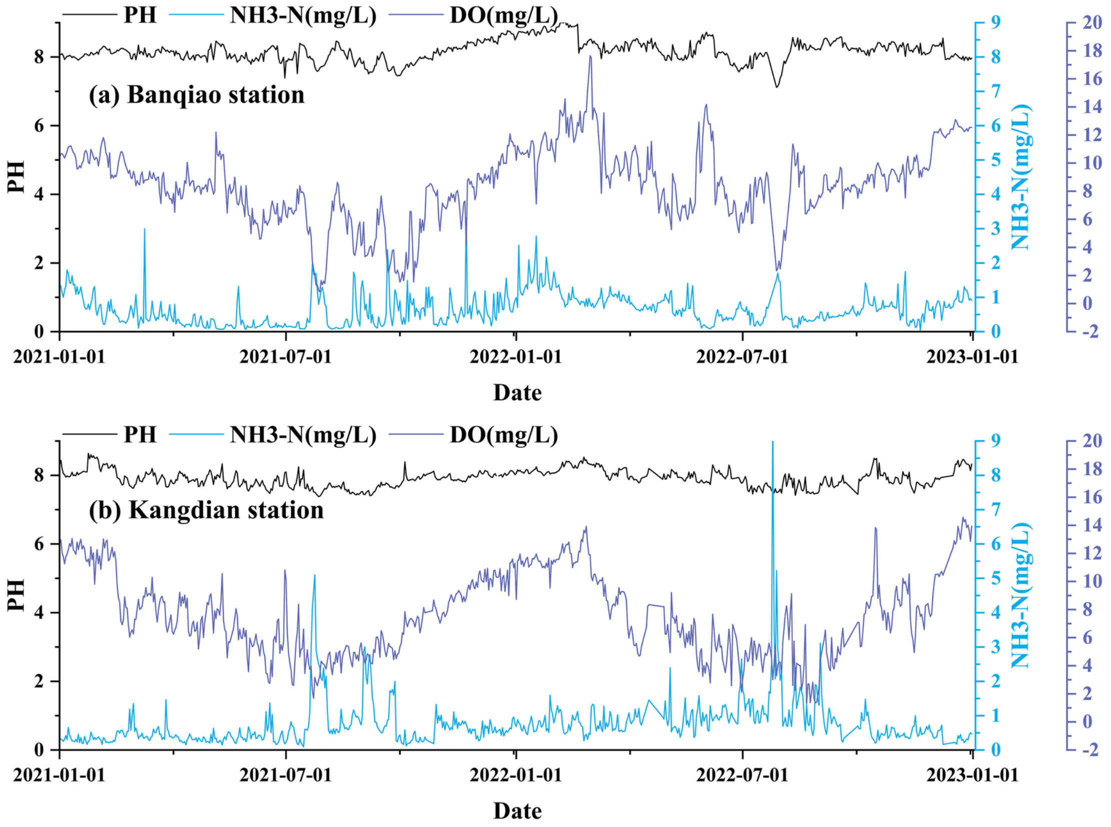


Figure 4. The data characteristics of water pollution prediction indicators (above is the Banqiao Station and below is the Kangdian Station).

stations during the flood season (June to September) in the Huaihe River Basin were significantly lower than those during the nonflood season. The main reason is that the flow in the nonflood season is small, resulting in worse water environment quality under the premise of the same pollutant emissions. Therefore, the phenomenon of water pollution in the Huaihe River Basin during the nonflood season is more prominent.

Table 1 and Fig. 4 demonstrate the presence of significant nonlinearity and uncertainty in the water pollution data collected at the Kangdian and Banqiao stations, and there were significant volatility and notable variations among the NH3-H and DO data. The pH data exhibited minimal fluctuation and had the lowest standard deviation among the data collected at the Kangdian and Banqiao stations. The three sets of sample data (pH, NH3-H, and DO) have different data characteristics. Thus, the prediction performance of the combined prediction model can be fully verified.

Feature selection results based on the RF algorithm

The RF algorithm was used to calculate the importance of water quality indicators, hydrometeorological indicators, and air quality indicators for COD_{Mn}, NH3-N, TN, and TP, and the input variable was selected based on the characteristic importance of each indicator. To avoid the contingency of feature importance analysis, the average of 10 feature importance calculations was used as the final importance of each feature. It was confirmed that indicators with a feature importance score exceeding 0.1 have a significant impact on predictive variables[38]. The indicator with importance score exceeding 0.1 was selected as the input variable for the water pollution prediction model in this study.

Figure 5 shows the results of the feature importance analysis. At the Kangdian and Banqiao stations, the main factors affecting pH were NH3-N, TN, W-Temp, COD_{Mn}, SO_2, DO, and TP, of which NH3-N had the highest characteristic contribution to pH, indicating that changes in pH are closely related to changes in NH3-N. The results of the importance analysis indicated that SO_2 in air was also an important factor affecting pH. The main reason is that SO_2 mainly comes from industrial emissions. The higher the SO_2 content is, the higher the amount of industrial pollutants discharged. The presence of industrial pollutants in the river significantly affect the pH of the water body. Therefore, pH changes are not only affected by water quality indicators but are also closely related to air quality indicators. Figure 5 also demonstrates that the main factors affecting NH3-N include TN, pH, TP, EC, DO, and W-Temp, of which TN and pH have the greatest impact on NH3-N. The main reason is that TN reflects the total amount of organic and inorganic nitrogen (including NH3-N) in the water body. Therefore, as a component of TN, NH3-N is directly affected by changes in TN. Furthermore, it has become an indisputable fact that water temperature is an important factor affecting DO. The results of the feature importance analysis also indicate that W-Temp and A-Temp are important factors affecting DO. From the results of the feature importance analysis, it was found that TN also had a significant impact on DO. Special attention should be given to changes in TN and temperature in the prediction of DO. Based on the results of feature importance analysis, NH3-H, TN, W-Temp, COD_{Mn}, SO_2, DO, and TP were selected as input variables for pH prediction, TN, pH, TP, EC,

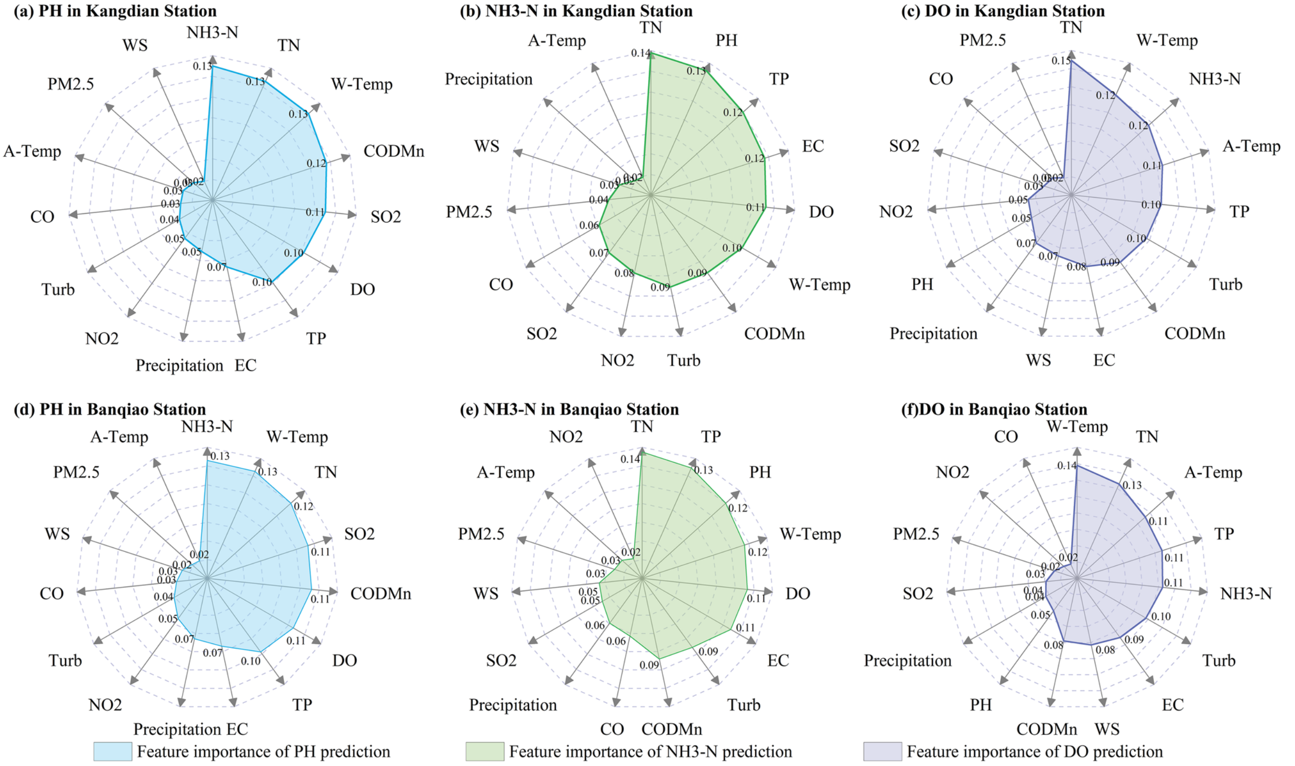


Figure 5. Results of feature importance analysis.

DO, and W-Temp were selected as input variables for NH3-N prediction, and TN, W-Temp, NH3-N, A-Temp, TP, and Turb were selected as input variables for DO prediction.

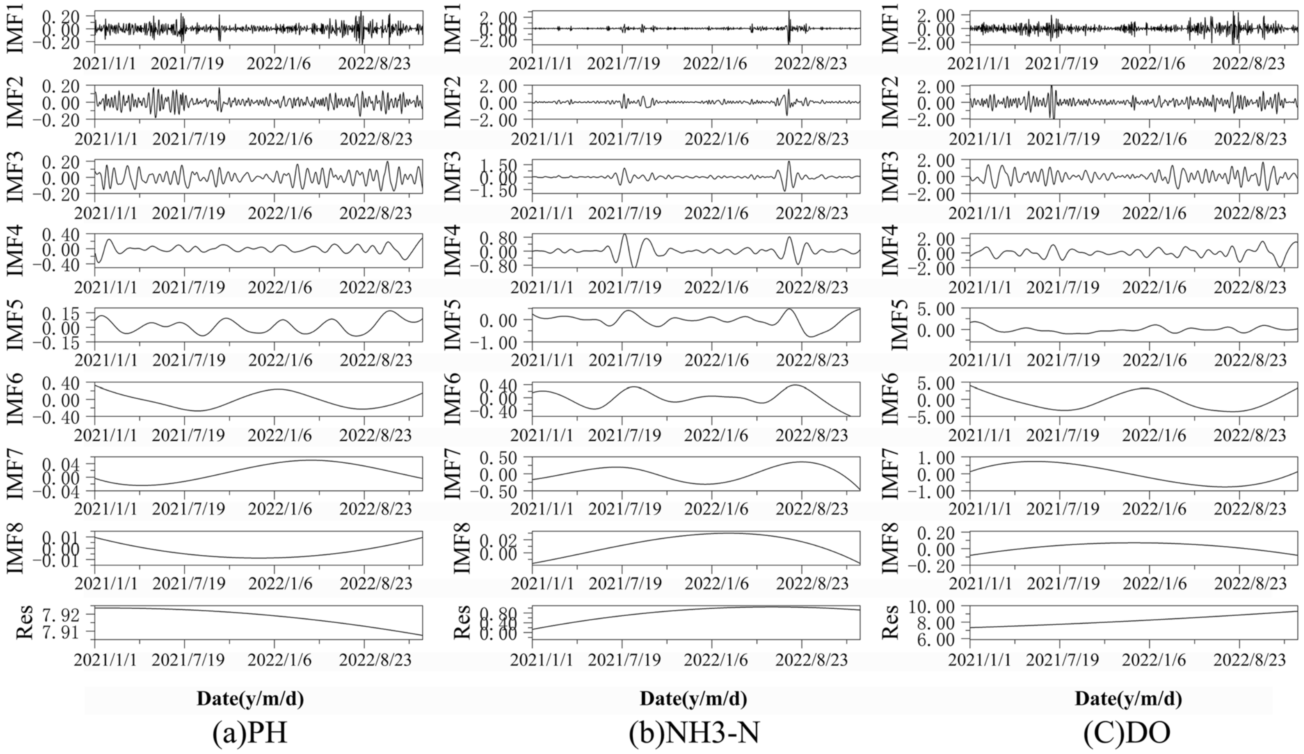
Predictive sequence decomposition based on CEEMD

It can be seen from Fig. 5 that there are high-frequency components and noise in the original water pollution prediction sequences, and it may be difficult to fit the model by directly using these sequences for prediction. Each sequence needs to be decomposed before modeling to reduce the impact of high-frequency components and noise on the model performance. Therefore, CEEMD was used to decompose the three prediction sequences of pH, NH3-N and DO at the Kangdian and Banqiao stations. As shown in Figs. 6 and 7, the three prediction sequences of the two stations were decomposed into eight high-frequency components and a residual term, and the signal curve of the CEEMD-decomposed predicted sequence component gradually tended to become stable as the frequency decreased. The fluctuation characteristics of the subsequence from IMF4 to IMF8 gradually became weaker, with an obvious periodic trend. Compared with the original sequence, the decomposed sequence obviously has better stationarity and periodicity.

To evaluate the effect of the CEEMD on the prediction sequences, the index of completion (IC) was selected to compare the CEEMD, EMD and EEMD data differences. The IC is the root mean square error of the decomposition sequence and the original sequence and is often used to compare the differences between the decomposed component sequence data and the original data. As shown in Table 2, EEMD significantly increased the error after reconstruction. The main reason is that EEMD eliminates the noise of the original sequence by adding auxiliary white noise, aiming to overcome the modal aliasing phenomenon existing in EMD. Added white noise is difficult to completely eliminate, resulting in a significant increase in the reconstruction error of the model. CEEMD is used to introduce white noise into each decomposition, ensuring that the error after reconstruction returns to the original order of magnitude. The guarantees accurate subsequent prediction results, effectively solves the modal oscillation phenomenon in EMD and verifies the validity of the method.

RF-CEEMD-LSTM prediction

The feature selection results of RF analysis were used as the input variables of the LSTM model, and LSTM was used to predict the 8 IMF components and 1 residual item obtained from the CEEMD and finally integrate the prediction sequence to obtain the pH, NH3-N and DO model predictions. The data from January 2021 to November 2022 were used as training samples to train the model and continuously adjust the parameters. After comprehensive consideration and multiple experiments, the network parameters of each water quality index prediction model were determined and are shown in Table 3. The water quality data from December 1st to December 31st, 2022, were input as test samples into the trained model. The results of each water quality index prediction model are shown in Fig. 8. The NSE of the prediction results was 0.99, demonstrating that the RF-CEEMD-LSTM model has a good prediction effect. From the MAPE values of the prediction results, the prediction accuracy of the model for pH (0.69% and 0.76% MAPE) was significantly higher than the prediction accuracies of NH3-N



(a)PH

(b)NH3-N

(C)DO

Figure 6. Forecast sequence decomposition of Kangdian station data.

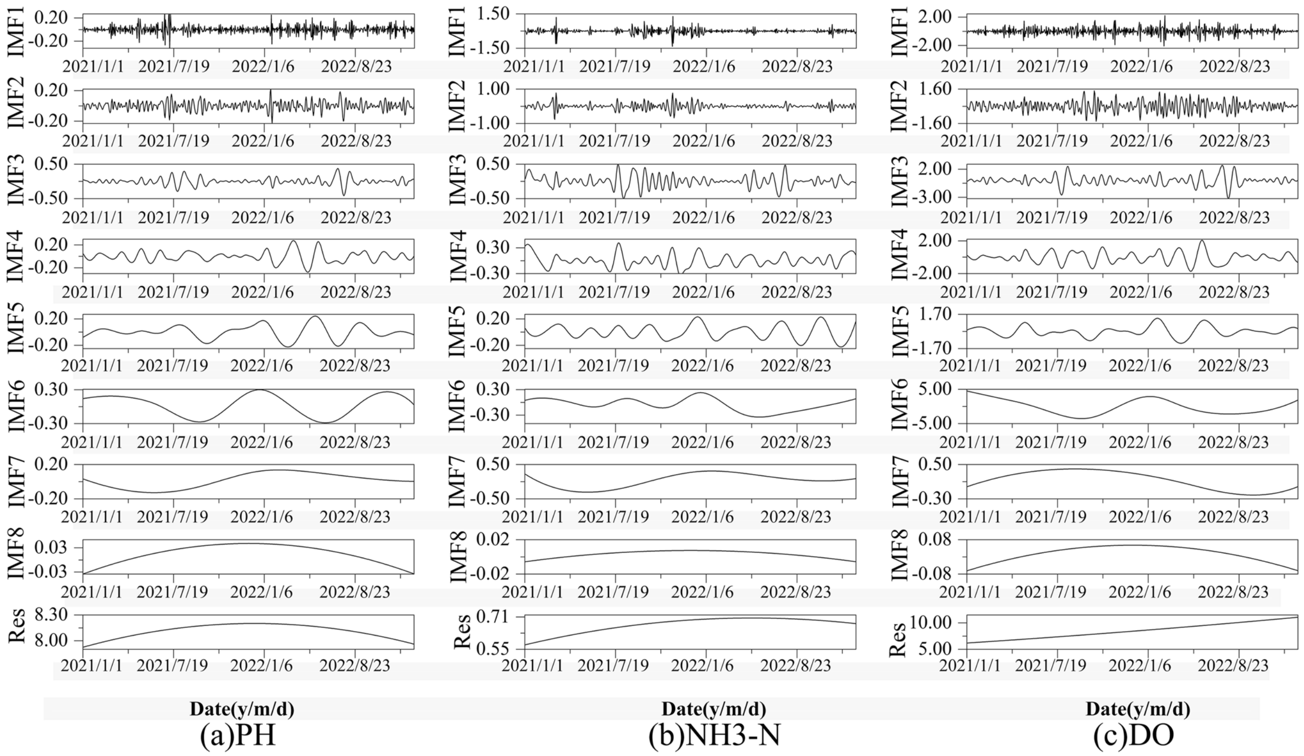


Figure 7. Forecast sequence decomposition of Banqiao station data.

Evaluation index	EMD	EEMD	CEEMD
IC	3.13×10^{-15}	0.017	3.19×10^{-15}
IMF number	6	8	8

Table 2. Evaluation index comparisons among EMD, EEMD and CEEMD.

Index	Number of hidden layer neurons	Learning rate	Iterations	Step
pH	20; 40	0.001	300	24
NH3-N	10; 20	0.001	250	24
DO	10; 10	0.003	250	24

Table 3. Network parameters corresponding to each water quality index.

(6.68% and 7.81% MAPE) and DO (1.58% and 1.19% MAPE). The main reason may be that the pH sequence is more stable (Fig. 4), and the RF-CEEMD-LSTM model can easily capture the potential rules of the sequence. Due to the high volatility of the NH3-N sequence, even if the original sequence was decomposed into 8 IMF components and 1 residual term by CEEMD, the decomposed IMF1 and IMF2 components still have certain volatility, which leads to the relatively poor prediction effect of the RF-CEEMD-LSTM model on NH3-N. Nevertheless, the proposed RF-CEEMD-LSTM model has a MAPE of water pollution prediction within 8%. It can predict the trend of water pollution more accurately.

Model comparative analysis

To further analyze the performance of the proposed RF-CEEMD-LSTM model, the differences in water pollution prediction based on the LSTM, RF-LSTM, CEEMD-LSTM and RF-CEEMD-LSTM models were compared. As shown in Table 4, the RF-CEEMD-LSTM prediction results have the smallest deviation from the actual value. The MAPE values of the RF-CEEMD-LSTM model were 0.73% (pH), 7.25% (NH3-N) and 1.39% (DO), and the NSE values of the prediction results were all above 0.99. In contrast, there was a large deviation between the prediction results of the single LSTM model and the actual results. In the prediction of NH3-N, the MAPE and NSE of the LSTM model were 14.58% and 0.88, respectively, which indicates that combined forecasting methods can effectively improve the overall forecasting performance by preprocessing the data through feature

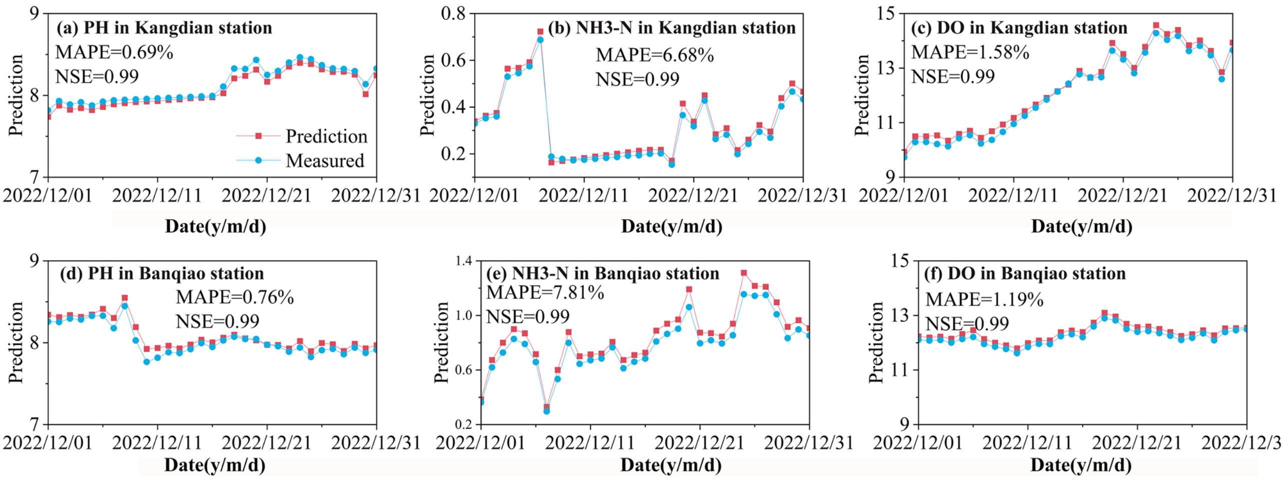


Figure 8. The prediction results of the RF- CEEMD-LSTM model.

Index	Model	NSE	RMSE (mg/L)	MAPE (%)	MAE (mg/L)
pH	LSTM	0.91	0.218	9.31	0.071
	RF-LSTM	0.94	0.101	7.52	0.068
	CEEMD-LSTM	0.975	0.079	3.46	0.061
	RF-CEEMD-LSTM	0.99	0.061	0.73	0.057
NH3-N	LSTM	0.88	0.256	14.58	0.073
	RF-LSTM	0.92	0.128	11.03	0.061
	CEEMD-LSTM	0.95	0.087	8.97	0.049
	RF-CEEMD-LSTM	0.99	0.071	7.25	0.044
DO	LSTM	0.91	0.32	10.21	0.25
	RF-LSTM	0.93	0.28	8.22	0.21
	CEEMD-LSTM	0.97	0.19	4.99	0.18
	RF-CEEMD-LSTM	0.99	0.18	1.39	0.16

Table 4. Comparison of the model prediction accuracy.

selection and decomposition. By comparing the prediction effects of the RF-LSTM and RF-CEEMD-LSTM models, it was found that the RMSE values of RF-LSTM for the three index predictions were 0.101, 0.128 and 0.28, and the RMSE values of the RF-CEEMD-LSTM model were 0.061, 0.071 and 0.18. The RMSE of the RF-CEEMD-LSTM model was 35.7–44.5% lower than that of the RF-LSTM model. Therefore, CEEMD significantly improved the prediction accuracy of the model. By comparing the prediction effects of the CEEMD-LSTM and RF-CEEMD-LSTM models, it was found that the RMSE of the prediction results of the RF-CEEMD-LSTM model was 5.3–22.8% lower than that of CEEMD-LSTM. This result proves that the use of RF for feature selection is effective in improving the model accuracy.

Discussion

In this work, RF was used to analyze the main factors affecting different water quality characteristics, and on this basis, input variables for different prediction indicators were identified. The results of the feature importance analysis indicated that the main factors affecting the prediction of pH are not only water pollution indicators (NH3-N, TN, W-Temp, COD_{Mn}, DO, and TP) but also the concentration of SO_2 in the air. Previous studies have shown that industrial activities are one of the most closely related factors[1] affecting the water quality of the Huaihe River Basin. The combustion and emission of sulfur containing fuels in industrial activities exacerbate the severity of SO_2 pollution in the air, and SO_2 entering the river with rainwater changes the pH balance of the river[39]. In the prevention and control process of river water pollution, it is possible to reduce or block SO_2 pollution and acidic water quality by increasing the emission standards of industrial waste gas. It was also found that the main factors affecting the change in NH3-N are TN and pH, and the main factors affecting the change in DO in water bodies are temperature and TN. TN has an important impact on pH, NH3-N, and DO in the Huaihe River Basin and is one of the most critical indicators affecting water quality. This discovery is consistent with Feng et al.[40]. The results in Table 4 show that by selecting input variables for different prediction indicators using RF, the RMSE of the model prediction results was reduced by 5.3–22.8%, significantly improving the model prediction accuracy. Therefore, in the construction of water pollution prediction models, the impact of indicator selection on the model performance should be considered[19]. Although this impact is not as significant as the impact of sequence decomposition on the model performance, it is an aspect that cannot be ignored.

Before constructing the water pollution prediction model, three prediction sequences were decomposed into eight IMF components and one residual using CEEMD in this study. By comparing the decomposition effects of EMD, EEMD, and CEEMD, it was found that the CEEMD model not only overcame the error caused by EMD mode aliasing but also significantly reduced the reconstruction error of EEMD by adding complementary white noise sequences, resulting in the best decomposition effect. This is also an important reason why some recent studies on hydrology[41] and water environment[23] have recommended the use of CEEMD method for data decomposition.

In order to compare the model performance more fairly, the performance of similar models in water pollution prediction in the latest research was statistically. As shown in Table 5, the proposed RF-CEEMD-LSTM model has the highest NSE, and the lowest MAPE, RMSE, and MAE, indicating that the proposed RF-CEEMD-LSTM has significantly better prediction accuracy than these similar water pollution prediction models. The main reason is that this study used RF algorithm to screen the most suitable input variables before constructing water pollution prediction model, and used CEEMD algorithm to decompose the filtered sequence dataset, the LSTM model can more comprehensively capture the fluctuation characteristics of these sequence data. In the prevention and control of water pollution, the proposed high accuracy water pollution prediction can reflect the current pollution situation of water bodies and the future trend of water body changes. Combining the analysis results of the feature importance, the main factors affecting water pollution can be quantified, which can provide a certain decision-making basis for relevant governance work. The trend perception of water pollution can transform the management of water resource pollution from post treatment to pre prevention and control, has long-term significance for improving the existing water pollution prevention and control situation and promoting the scientific development of the water environment.

There are still some limitations in this study due to the limited data and research subject. This study uses daily water pollution data to construct water pollution prediction model, which can be used for short-term water pollution prediction and prevention. The applicability of the proposed method for water pollution prediction at other scales still needs further research. Future research can collect longer sequence data to construct weekly and monthly scale water pollution prediction models, which can provide more comprehensive support for water pollution prevention and control. And this study only constructed a prediction model for pH, NH3-N and DO. Future research can explore the effectiveness of the proposed method in predicting other water pollution indicators, especially for heavy metal pollution prediction, which is very useful for water supply safety prevention and control.

Conclusions

In order to improve the performance of water pollution prediction models and effectively explain the main influencing factors of water pollution. This study proposed a hybrid model for water pollution prediction based on RF-CEEMD-LSTM to predict the changes in pH, NH3-N and DO. A hybrid water pollution prediction model was constructed for the Kangdian and Banqiao stations in the Huaihe River Basin, and the performance of the model was verified using statistical evaluation indicators. Various similar models were used to compare the performance of the constructed model. The main conclusions can be divided into three aspects.

(1) The RF algorithm was used to analyze the feature importance of various water pollution prediction variables in this study, and the results showed that TN is the most critical factor affecting water quality changes. Water pollution prevention and policy formulation need pay more attention to TN reduction strategies.

(2) The IC value of CEEMD in data decomposition is significantly lower than EMD, and the RMSE of the prediction model using CEEMD (RF-CEEMD-LSTM) is lower than that using EMD (EMD-LSTM) and EEMD (EEMD-LSTM), indicating that using CEEMD can effectively improve the performance of water pollution sequence prediction models. CEEMD can be one of the most effective methods for nonlinear non-stationary data decomposition.

(3) The comparisons between different models and the experimental results show that the RMSE value of the proposed RF-CEEMD-LSTM model is 62.6%, 39.9% and 15.5% lower than those of the LSTM, RF-LSTM and CEEMD-LSTM models, indicating that proposed model can provide superior predictive performance. Proposed RF-CEEMD-LSTM model could provide references for improving water pollution prediction method.

Model	NSE	RMSE (mg/L)	MAPE (%)	MAE (mg/L)	Source
RF-CEEMD-LSTM	0.99	0.061	0.73	0.057	This work
EEMD-LSTM	0.95	0.094	1.47	0.070	Luo et al.[42]
EMD-LSTM	0.94	0.27	2.66	–	Zhang et al.[20]
BPNN	–	0.189	1.70	0.165	Li et al.[43]
LSTM	–	0.085	1.11	0.075	Li et al.[43]

Table 5. Comparison of prediction accuracy against similar models.

Data availability

The datasets used during the current study available from the corresponding author on reasonable request.

Received: 5 July 2023; Accepted: 26 November 2023
Published online: 28 November 2023

References

1. Li, H., Chen, S., Ma, T. & Ruan, X. The quantification of the influencing factors for spatial and temporal variations in surface water quality in recent ten years of the Huaihe River Basin, China. *Environ. Sci. Pollut. Res.* **29**, 44490–44503. https://doi.org/10.1007/s11356-021-18282-9 (2022).
2. Kim, J. *et al.* A novel hybrid water quality forecast model based on real-time data decomposition and error correction. *Process Saf. Environ. Prot.* **162**, 553–565. https://doi.org/10.1016/j.psep.2022.04.020 (2022).
3. Qiu, R. *et al.* Water temperature forecasting based on modified artificial neural network methods: Two cases of the Yangtze River. *Sci. Total Environ.* **737**, 139729. https://doi.org/10.1016/j.scitotenv.2020.139729 (2020).
4. KarimaeiTabarestani, M. & Fouladfar, H. Effect of reservoir size on water quality in coastal reservoirs during desalinization period using numerical model. *Water Supply* **22**, 5080–5094. https://doi.org/10.2166/ws.2022.182 (2022).
5. Fu, B. *et al.* Modeling water quality in watersheds: from here to the next generation. *Water Resour. Res.* **56**, e2020WR027721. https://doi.org/10.1029/2020WR027721 (2020).
6. Sao, D. *et al.* Evaluation of different objective functions used in the SUFI-2 calibration process of SWAT-CUP on water balance analysis: A case study of the Pursat river basin, Cambodia. *Water* **12**, 2901. https://doi.org/10.3390/w12102901 (2020).
7. Liu, Y. *et al.* Predicting urban water quality with ubiquitous data—A data-driven approach. *IEEE Trans. Big Data.* **8**, 564–578. https://doi.org/10.1109/TBDATA.2020.2972564 (2022).
8. Chen, Y., Song, L., Liu, Y., Yang, L. & Li, D. A review of the artificial neural network models for water quality prediction. *Appl. Sci.* **10**, 5776. https://doi.org/10.3390/app10175776 (2020).
9. Nguyen, K., François, B., Balasubramanian, H., Dufour, A. & Brown, C. Prediction of water quality extremes with composite quantile regression neural network. *Environ. Monit. Assess.* **195**, 284. https://doi.org/10.1007/s10661-022-10870-7 (2023).
10. Rustam, F. *et al.* An artificial neural network model for water quality and water consumption prediction. *Water.* **14**, 3359. https://doi.org/10.3390/w14213359 (2022).
11. Najwa Mohd Rizal, N. *et al.* Comparison between regression models, support vector machine (SVM), and artificial neural network (ANN) in river water quality prediction. *Processes.* **10**, 1652. https://doi.org/10.3390/pr10081652 (2022).
12. Wang, J., An, Y., Li, Z. & Lu, H. A novel combined forecasting model based on neural networks, deep learning approaches, and multi-objective optimization for short-term wind speed forecasting. *Energy.* **251**, 123960. https://doi.org/10.1016/j.energy.2022.123960 (2022).
13. Ma, J., Ding, Y., Cheng, J. C. P., Jiang, F. & Xu, Z. Soft detection of 5-day BOD with sparse matrix in city harbor water using deep learning techniques. *Water Res.* **170**, 115350. https://doi.org/10.1016/j.watres.2019.115350 (2020).
14. Cho, K. H., Pachepsky, Y., Ligaray, M., Kwon, Y. & Kim, K. H. Data assimilation in surface water quality modeling: A review. *Water Res.* **186**, 116307. https://doi.org/10.1016/j.watres.2020.116307 (2020).
15. Ahmadianfar, I., Jamei, M. & Chu, X. A novel hybrid wavelet-locally weighted linear regression (W-LWLR) model for electrical conductivity (EC) prediction in surface water. *J. Contam. Hydrol.* **232**, 103641. https://doi.org/10.1016/j.jconhyd.2020.103641 (2020).
16. Xiao, X. *et al.* A novel single-parameter approach for forecasting algal blooms. *Water Res.* **108**, 222–231. https://doi.org/10.1016/j.watres.2016.10.076 (2017).
17. Zhou, Y. Real-time probabilistic forecasting of river water quality under data missing situation: Deep learning plus post-processing techniques. *J. Hydrol.* **589**, 125164. https://doi.org/10.1016/j.jhydrol.2020.125164 (2020).
18. Hooftman, D. A. P. *et al.* Reducing uncertainty in ecosystem service modelling through weighted ensembles. *Ecosyst. Serv.* **53**, 101398. https://doi.org/10.1016/j.ecoser.2021.101398 (2022).
19. Khudhair, Z. S. *et al.* A review of hybrid soft computing and data pre-processing techniques to forecast freshwater quality's parameters: Current trends and future directions. *Environments.* **9**, 85. https://doi.org/10.3390/environments9070085 (2022).
20. Zhang, Y. *et al.* Accurate prediction of water quality in urban drainage network with integrated EMD-LSTM model. *J. Clean Prod.* **354**, 131724. https://doi.org/10.1016/j.jclepro.2022.131724 (2022).
21. Eze, E., Halse, S. & Ajmal, T. Developing a novel water quality prediction model for a South African aquaculture farm. *Water.* **13**, 1782. https://doi.org/10.3390/w13131782 (2021).
22. Wu, Z. & Huang, N. E. Ensemble empirical mode decomposition: A noise-assisted data analysis method. *Adv. Adapt. Data Anal.* **1**, 1–41. https://doi.org/10.1142/S1793536909000047 (2009).
23. Li, J., Wu, Z., He, H. & Lu, W. Application of the complementary ensemble empirical mode decomposition for the identification of simulation model parameters and groundwater contaminant sources. *J. Hydrol.* **612**, 128244. https://doi.org/10.1016/j.jhydrol.2022.128244 (2022).
24. Yeh, J. R., Shieh, J. S. & Huang, N. E. Complementary ensemble empirical mode decomposition: A novel noise enhanced data analysis method. *Adv. Adapt. Data Anal.* **2**, 135–156. https://doi.org/10.1142/S1793536910000422 (2010).
25. Ding, Y., Chen, Z., Zhang, H., Wang, X. & Guo, Y. A short-term wind power prediction model based on CEEMD and WOA-KELM. *Renew. Energy.* **189**, 188–198. https://doi.org/10.1016/j.renene.2022.02.108 (2022).
26. Xu, Y., Sun, H. & Ji, X. Spatial-temporal evolution and driving forces of rainfall erosivity in a climatic transitional zone: A case in Huaihe River Basin, Eastern China. *Catena.* **198**, 104993. https://doi.org/10.1016/j.catena.2020.104993 (2021).
27. Li, P. *et al.* Helicobacter pylori infection and immune factors on residents in high-incidence areas of cancer along S river. *Life Sci. J.* **8**, 500–504 (2011).
28. Chand, A. K. B. & Kapoor, G. P. Generalized cubic spline fractal interpolation functions. *SIAM J. Numer. Anal.* **44**, 655–676. https://doi.org/10.1137/040611070 (2006).
29. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32. https://doi.org/10.1023/A:1010933404324 (2001).
30. He, S., Wu, J., Wang, D. & He, X. Predictive modeling of groundwater nitrate pollution and evaluating its main impact factors using random forest. *Chemosphere.* **290**, 133388. https://doi.org/10.1016/j.chemosphere.2021.133388 (2022).
31. Guo, M. *et al.* Quantitative analysis of polycyclic aromatic hydrocarbons (PAHs) in water by surface-enhanced Raman spectroscopy (SERS) combined with random forest. *Spectroc. Acta Pt. A-Mol. Biomol. Spectr.* **287**, 122057. https://doi.org/10.1016/j.saa.2022.122057 (2023).
32. Zhu, S. *et al.* Two-step-hybrid model based on data preprocessing and intelligent optimization algorithms (CS and GWO) for NO_2 and SO_2 forecasting. *Atmos. Pollut. Res.* **10**, 1326–1335. https://doi.org/10.1016/j.apr.2019.03.004 (2019).
33. Liu, W., Cao, S. & Chen, Y. Applications of variational mode decomposition in seismic time-frequency analysis. *Geophysics.* **81**, 365–378. https://doi.org/10.1190/geo2015-0489.1 (2016).
34. Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural Comput.* **9**, 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735 (1997).

35. Zhang, D., Lindholm, G. & Ratnaweera, H. Use long short-term memory to enhance Internet of Things for combined sewer overflow monitoring. *J. Hydrol.* **556**, 409–418. https://doi.org/10.1016/j.jhydrol.2017.11.018 (2018).
36. Mohtasham Moein, M. *et al.* Predictive models for concrete properties using machine learning and deep learning approaches: A review. *J. Build. Eng.* **63**, 105444. https://doi.org/10.1016/j.jobe.2022.105444 (2023).
37. Su, J. *et al.* Developing surface water quality standards in China. *Resour. Conserv. Recycl.* **117**, 294–303. https://doi.org/10.1016/j.resconrec.2016.08.003 (2017).
38. Huang, H., Liu, L., Cao, R. & Cao, Y. Prediction and feature importance of earth pressure in shields using machine learning algorithms. *KSCE J. Civ. Eng.* **27**, 862–877. https://doi.org/10.1007/s12205-022-1241-8 (2023).
39. Zhao, S. *et al.* Temporal dynamics of SO_2 and NO_X pollution and contributions of driving forces in urban areas in China. *Environ. Pollut.* **242**, 239–248. https://doi.org/10.1016/j.envpol.2018.06.085 (2018).
40. Feng, H. *et al.* Mapping multiple water pollutants across China using the grey water footprint. *Sci. Total Environ.* **785**, 147255. https://doi.org/10.1016/j.scitotenv.2021.147255 (2021).
41. Chen, Y., Yeh, H., Kao, S., Wei, C. & Su, P. Water level forecasting in tidal rivers during typhoon periods through ensemble empirical mode decomposition. *Hydrology.* **10**, 47. https://doi.org/10.3390/hydrology10020047 (2023).
42. Luo, L., Zhang, Y., Dong, W., Zhang, J. & Zhang, L. Ensemble empirical mode decomposition and a long short-term memory neural network for surface water quality prediction of the Xiaofu river, China. *Water.* **15**, 1625. https://doi.org/10.3390/w15081625 (2023).
43. Li, Z. *et al.* Water quality prediction model combining sparse auto-encoder and LSTM network. *IFAC-PapersOnLine* **51**, 831–836. https://doi.org/10.1016/j.ifacol.2018.08.091 (2018).

Acknowledgements

This Research was supported by Transportation Science and Technology project of Henan Province (Grant No. 2021J1). We thank the hydrological and transportation departments of Henan Province for their assistance in obtaining research data. We also thank the anonymous reviewers for their valuable comments.

Author contributions

Conceptualization, J.R. and Y.C.; methodology, J.R.; software, Y.S.; validation, J.R. and Y.C.; in-vestigation, J.R.; resources, Y.C.; data curation, Y.C.; writing—original draft preparation, J.R.; writing—review and editing, Y.M.; visualization, Y.S.; supervision, Y.S.; project administration, Y.M.; funding acquisition, J.R. All authors have read and agreed to the published version of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Y.C.

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.