scientific reports

Check for updates

OPEN

Research on water environmental indicators prediction method based on EEMD decomposition with CNN-BiLSTM

Zhaohua Wang[1], Longzhen Duan[1], Dongsheng Shuai[2] & Taorong Qiu[1]✉

Water resources protection is related to the development of the social economy, and the monitoring and prediction of water environmental indicators have important practical significance. In view of the seasonality, periodicity, uncertainty, and nonlinear characteristics of water quality indicators data, traditional prediction models have poor performance. To address this issue, this paper introduces a hybrid water quality index prediction model based on Ensemble Empirical Mode Decomposition (EEMD), combined with Convolutional Neural Network (CNN) and Bidirectional Long Short-Term Memory Network (BiLSTM). We have conducted out experiments to predict dissolved oxygen based on the water quality monitoring indicators of the Liaohe National Control Sanhongcun Village station in Yichun City. The results show that the model proposed in this paper improves the R^2 index by 5%, 7% and 5% compared to the suboptimal model in the 4-h, 1-day and 2-day index predictions, respectively.

In recent years, with the development of socio-economy, water pollution has garnered escalating public attention, leading to water resource protection being widely recognized as a societal consensus. The dynamic monitoring of changes in water quality, coupled with the implementation of water environment indicator predictions, holds profound practical significance for the preservation of water resources.

The prediction of water environment indicators involves the identification of temporal changes in water quality indicators and their correlation with hydrological, meteorological, and other factors within a specified spatiotemporal context[1]. Water environment indicator prediction can be categorized into mechanistic prediction methods and non-mechanistic prediction methods, depending on their underlying theoretical foundations.

Mechanistic prediction methods are holistic approaches grounded in the governing principles and evolving dynamics of the water environment, encompassing diverse disciplines such as hydrodynamics, ecology, and chemistry[2]. These methods typically employ models to encapsulate the intricate interplay among various elements. Commonly utilized models in this category include the Water Quality Analysis Simulation Program (WASP)[3], QUAL model[4], MIKE system[5], Generalized Watershed Loading Function (GWLF)[6], and others.

In contrast, non-mechanistic prediction methods adopt a 'black box' approach. These models rely on probabilistic statistical theories and are tailored to specific water environments, demonstrating effective predictive capabilities. Three prevalent non-mechanistic models can be identified: traditional probabilistic statistical models, such as the grey model[7] and Markov chain model[8]; time series models, such as Exponential Smoothing (ETS) and Auto Regressive Integrated Moving Average (ARIMA); and artificial intelligence models, including Support Vector Machine (SVM), Long Short-Term Memory (LSTM), eXtreme Gradient Boosting (XGBoost), Gate Recurrent Unit (GRU), and Informer, among others.

Surface water is an important type of water environment. Its water quality indicators exhibit characteristics such as seasonality, periodicity, uncertainty, and nonlinearity. There are also complex dependent relationships between the indicators[9]. Traditional probabilistic statistical methods are difficult to model such complex dependent relationships. At present, artificial intelligence methods represented by deep learning have made great progress in the application of surface water environment indicator prediction. Recurrent neural networks (RNNs) are suitable for processing time series data, but they suffer from the problem of gradient disappearance. To solve the problems in RNNs, Hochreiter et al.[10] proposed LSTM networks,which can perform long time series prediction tasks. Hu et al.[11] used LSTM to predict pH and water temperature in water quality indicators, and Zhang

[1]School of Mathematics and Computer Sciences, Nanchang University, Nanchang, China. [2]Jiangxi Zhonggan Investment Survey and Design Limited Company, Nanchang, China. ✉email: qiutaorong@ncu.edu.cn

Yiting et al.[12] applied LSTM to the prediction of ammonia nitrogen indicators in river water quality. However, a single LSTM model cannot avoid the interference of noise, resulting in unsatisfactory prediction accuracy. To solve the noise interference problem, convolutional neural networks (CNNs) are introduced to extract features from multidimensional time series, such as: Zhang Mingwei et al.[13] employed the CNN-LSTM model to predict the dissolved oxygen index of river water quality, and Wang Zhibo et al.[14] employed CNN-LSTM to predict the dissolved oxygen index of lake water quality. But LSTM can only make predictions based on historical data, while water quality indicators are not only related to historical data, but also related to future data. On the other hand, modal decomposition methods are introduced to eliminate the impact of noise, such as: Yuan Meixue et al.[15] employed wavelet decomposition to denoise water quality data, and then used a hybrid LSTM and Seq2Seq model for prediction. Benjamin et al.[16] applied the Empirical Mode Decomposition (EMD) method to decompose the dissolved oxygen indicator in the water quality time series, effectively isolating the trend and fluctuation components of the data. José et al.[17] employed EMD and LSTM to improve the performance of time series classification. Bai Wenrui et al.[18] first employed Variational Mode Decomposition(VMD) to decompose water quality indicators, and then used LSTM to predict water quality indicators.Wavelet decomposition has defects such as edge effects and difficulty in determining the basis function; while VMD requires higher data stability and linearity.

This paper proposes a CNN-BiLSTM water quality indicator prediction model based on Ensemble Empirical Mode Decomposition (EEMD) decomposition, aiming to overcome the prevalent challenges in deep learning applications for water quality indicator prediction, as well as to address the periodicity, uncertainty, and nonlinearity inherent in water quality monitoring data. EEMD effectively mitigates the issue of mode mixing encountered in EMD and imposes less stringent requirements on data stationarity and linearity compared to VMD. CNN is employed to extract local features from water quality indicator data, while BiLSTM handles sequential dependence modeling within this data, considering the impacts of both forward and backward data. To validate the efficacy of our proposed model, we conducted multivariate and multi-step prediction experiments using water quality data obtained from the national monitoring station in Sanhong Village, Liaohe.

Model and methods

Water environment indicator decomposition

EEMD was proposed by Wu et al.[19]based on Empirical Mode Decomposition (EMD) to overcome the problem of mode mixing in EMD decomposition.

EEMD is a method that involves adding Gaussian white noise to the original sequence, applying EMD to the sequence multiple times according to a predefined number of experiments, and then taking the average of the decomposition results to eliminate the influence of noise. This methodology imparts properties of uniform distribution and smoothness to the original sequence.The steps for sequence decomposition in EEMD are as follows:

(i) Add white noise of limited amplitude to the original indicator sequence to obtain a new sequence:

$$X^s = X + \varepsilon^s \tag{1}$$

where $X(X \in R^{(m\times n)})$ is the original sequence, ε^s is white noise,and X^s is the new sequence.

(ii) Decompose X^s into Intrinsic Mode Function(IMF) components using EMD:

$$X^s = \sum_{l}^{L} C_l^{EMD,s} + r(t) \tag{2}$$

where $C_l^{EMD,s}$ is the intrinsic mode function after EEMD decomposition, $r(t)$ is residual.

(iii) Repeat the above steps according to the set number of times and calculate the final result:

$$X = (C_1^{EMD,s}, C_2^{EMD,s}, ..., C_L^{EMD,s}, r(t)) \tag{3}$$

The process flow of EEMD decomposition for water quality indicators is illustrated in the Fig. 1.

Local correlation feature extraction of water environment indicators

Convolutional neural networks (CNN) are feedforward neural networks that use convolution and pooling operations for feature extraction. It is an important algorithm in deep learning. For time series data, 1D convolutions are often used.

In this paper, a sliding window is employed on the water environment indicator sequence to extract local features. Additional noise filtering is carried out through convolution and pooling operations to achieve enhanced outcomes. The specific formula is as follows:

$$Y = w * X \tag{4}$$

where w is the convolution kernel, $*$ denotes convolution, X represents the water quality indicator sequence that has been decomposed by EEMD, and Y is the extracted feature.

Temporal dependence modeling of water environment indicators

This paper chooses BiLSTM to model temporal dependencies. BiLSTM constitutes an advancement over the LSTM neural network. Relevant research[20] indicates that BiLSTM offers noteworthy enhancements in performance compared to LSTM for time series prediction tasks.

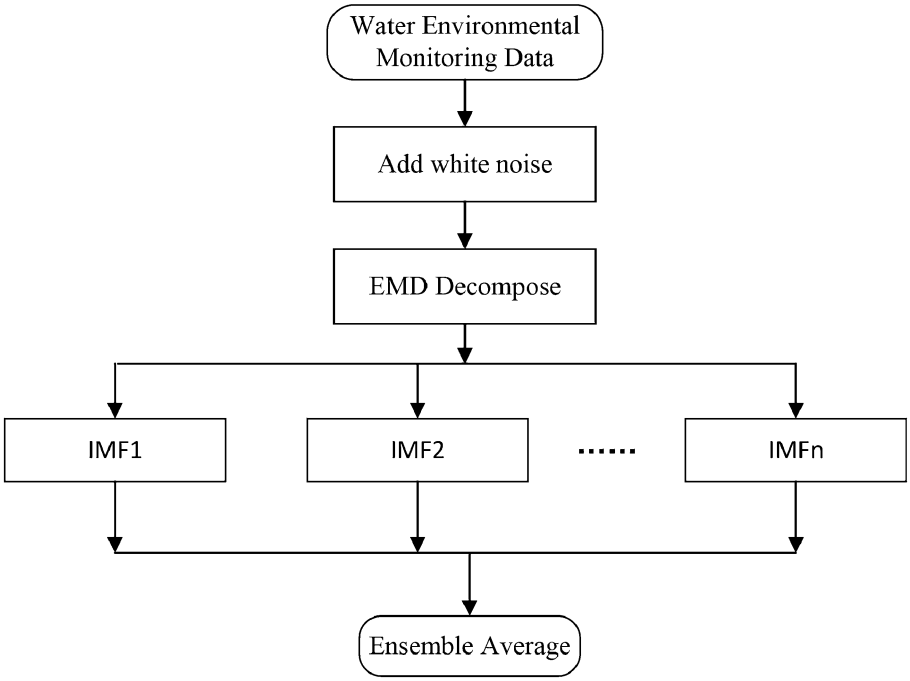


Figure 1. EEMD decomposition flowchart.

$$H = BiLSTM(Y) \tag{5}$$

where Y represents the vector of target variables to be predicted, H represents the prediction results. BiLSTM consists of two layers of LSTM neural networks that operate in opposing directions.Rather than merely stacking the two LSTM layers, it integrates data features from both forward and backward directions at the present time step for predictive purposes.

Model building

Given the strong coupling and nonlinear characteristics of water environment monitoring data, traditional prediction methods often yield subpar results.Accordingly, this paper introduces a CNN-BiLSTM hybrid model for water environment data prediction based on EEMD decomposition.

Initially, the preprocessed water environment data is decomposed by EEMD, yielding four modes. Each of these modes is subsequently fed into both CNN and BiLSTM for feature extraction. Ultimately, the extracted features are accumulated and reconstructed to derive the predictive outcome.

This hybrid model synergistically integrates EEMD, CNN, and BiLSTM to capitalize on the strengths of each component: EEMD for noise reduction, CNN for capturing local features, and BiLSTM for modeling sequential dependencies. The ensemble methodology has the potential to enhance prediction accuracy. In this experiment, dissolved oxygen is decomposed by EEMD, and then combined with other indicators to form new training data. The model structure is illustrated in the Fig. 2.

Experiments

Dataset

The research focuses on water quality monitoring data obtained from the national monitoring station in Sanhong Village, Liaohe.Liaohe is the largest tributary of Xiuhe River,which traverses Jing'an County in Yichun City. It holds significance as the primary river in the county and eventually merges into Poyang Lake via the Xiuhe River.

The monitoring dataset spans from November 2020 to December 2022,with measurements taken every four hours, amounting to a total of 4,700 data points. It encompasses nine indicators: water temperature (TEMP),pH,dissolved oxygen (DO),potassium permanganate (PP),ammonia nitrogen (TAN),total phosphorus (TP),total nitrogen (TN),electrical conductivity (EC),and turbidity (TUB).This dataset is obtained from the Environmental Quality Information Release Platform of Jiangxi Province.

In addition, meteorological data from Yichun City covering the same time period was also gathered, encompassing six indicators:temperature,atmospheric pressure,humidity,wind speed,dew point temperature,and precipitation.This data is obtained from the website "Reliable Prognosis".

Among the various water quality indicators, the concentration of dissolved oxygen serves as a crucial benchmark for assessing water quality[21]. Consequently, this paper focuses on utilizing dissolved oxygen as the target indicator for model prediction.

Through a series of experiments and evaluations, it was determined that '4' was the optimal number of modalities, as it demonstrated the best performance and accuracy during model training. In this paper, the EEMD

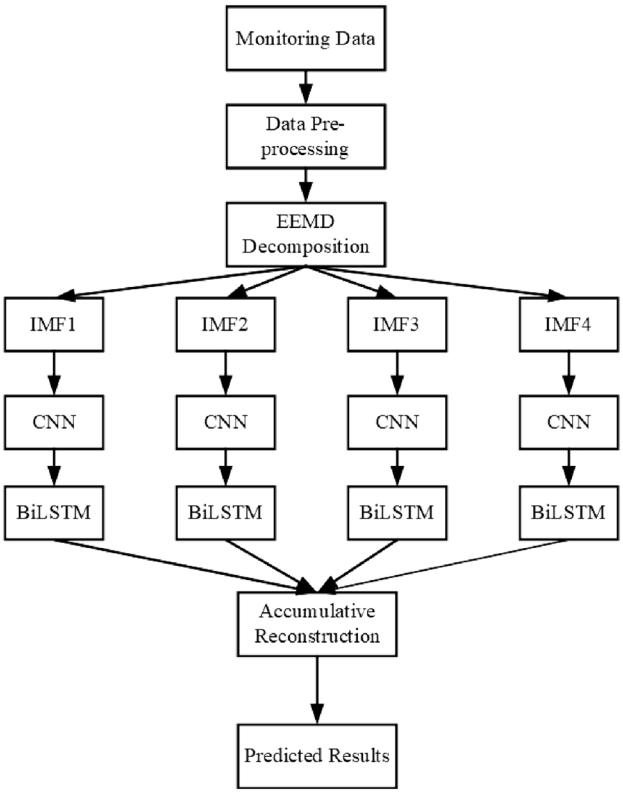


Figure 2. EEMD-CNN-BiLSTM Mixture model Diagram.

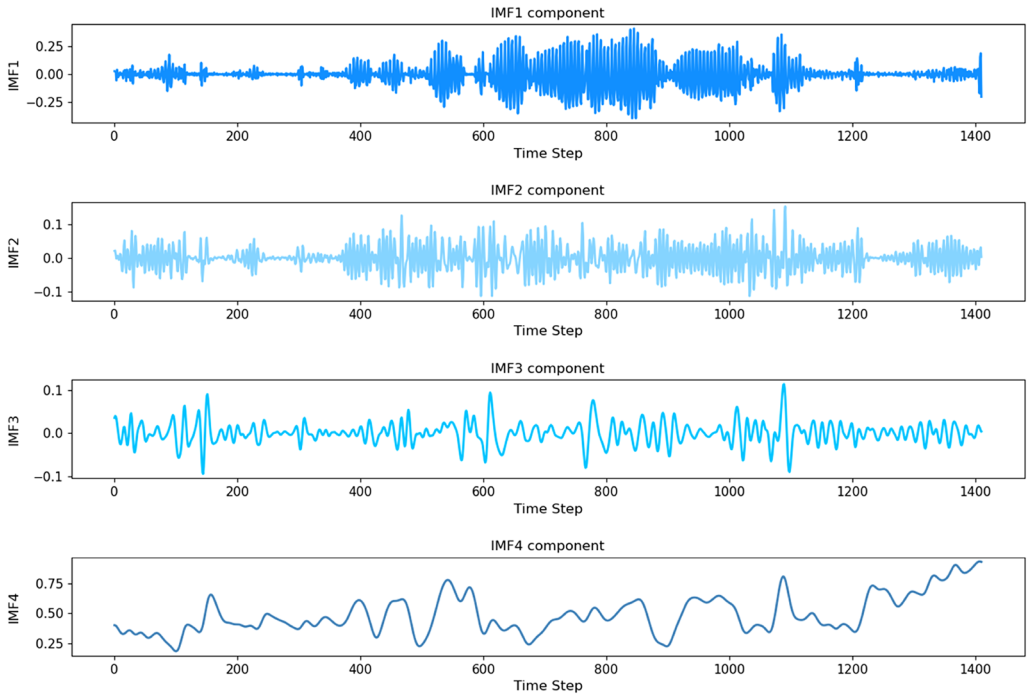


Figure 3. Dissolved oxygen index after decomposition of EEMD.

method (4 modes) is employed to decompose the dissolved oxygen indicator through experimental comparison. The waveform diagrams of each mode after decomposition in the validation and test sets are illustrated in Fig. 3:

Through autocorrelation experiments, we observed that the three modes: IMF1, IMF2, and IMF3 exhibit evident cyclical characteristics, while IMF4 retains the trend characteristic inherent in the data.

(i) Missing and outlier value handling

During the analysis of the data, it was discovered that certain issues such as missing values and outliers existed due to factors like equipment maintenance or malfunctions that occurred during the data collection process.

For indicators with a significant number of consecutive missing values, linear interpolation is employed to fill in the gaps according to the formula:

$$\varphi(x) = \frac{x - x_1}{x_0 - x_1} y_0 + \frac{x - x_0}{x_1 - x_0} y_1 \tag{6}$$

where x represents time, $\varphi(x)$ represents the estimated value at that specific time x. The coordinates x_0 and y_0 represent the first known data point, x_1 and y_1 represent the second known data point.

(ii)Normalization

As water quality indicators possess distinct scales, for optimal model training, each indicator is normalized using the formula:

$$x^{'} = \frac{x - min(x)}{max(x) - min(x)} \tag{7}$$

where x is the original data that needs to be normalized, $x^{'}$ is the normalized data, and its value range is [0,1], $max(x)$ and $min(x)$ are the maximum and minimum values in the dataset, respectively.

(iii) Correlation analysis

To investigate the significance of each indicator in the prediction process, correlation analysis is conducted on the data, and a correlation heat map is presented in the figure 4.

It is evident that following EEMD decomposition, the correlations between dissolved oxygen and various indicators such as temperature, electrical conductivity, ammonia nitrogen, and total nitrogen have demonstrated an increase.

Determination of model parameters

In this paper, grid search is employed to optimize the model parameters. Only one parameter is adjusted at a time, and grid search is utilized for fine-tuning. Through iterative execution of the aforementioned steps, the optimized model parameters are presented in Table 1:

Metrics for experimental evaluation

Mean absolute error (MAE),mean square error (MSE),Mean Absolute Percentage Error (MAPE) and correlation coefficient (R^2) are employed as quantitative metrics to assess the predictive performance of the model.

$$MAE = \frac{\sum |y - \hat{y}|}{n} \tag{8}$$

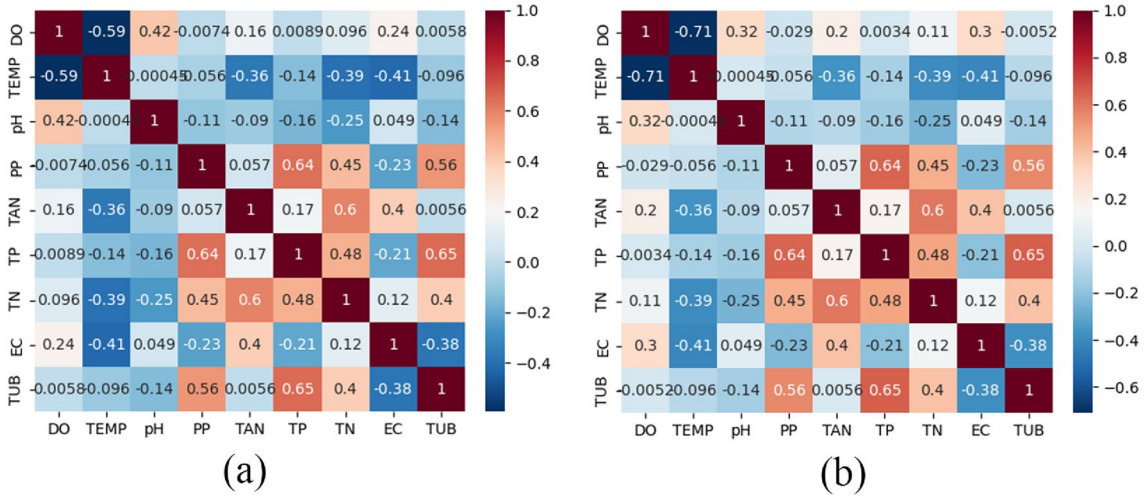


Figure 4. Heat map: (**a**) is correlation between water quality indicators, (**b**) is IMF4 correlation heat map after EEMD decomposition.

Indicators	XGboost	LSTM	GRU	BiLSTM	CNN-BiLSTM	Informer	Ours
Training set	0.7	0.7	0.7	0.7	0.7	0.7	0.7
Validation set	0.15	0.15	0.15	0.15	0.15	0.15	0.15
Test set	0.15	0.15	0.15	0.15	0.15	0.15	0.15
Batch size	–	256	256	256	512	256	256
Input window	8	8	8	8	8	8	8
Loss function	–	MSE	MSE	MSE	MSE	MSE	MSE
Learning rate	0.1	0.001	0.001	0.001	0.001	0.001	0.001
Weight decay	–	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
Stacking depth	–	1	1	1	1	–	1
Hidden layer dimensions	–	256	256	32	256	–	256
Training epochs	100	300	300	300	300	300	300
CNN Output channels	–	–	–	–	128	–	128
CNN Convolution kernel size	–	–	–	–	–	–	1
CNN Convolution stride	–	–	–	–	–	–	1

Table 1. Model parameters for each model. Since each model has different characteristics,the parameters that need to be set are not exactly the same.In the table, "–" indicates that the model does not need to set this parameter.In order to facilitate the comparison of model performance, the same parameters should be set as much as possible

$$MSE = \frac{\sum (y - \hat{y})^2}{n} \tag{9}$$

$$MAPE = \frac{100\%}{n} \sum \left| \frac{\hat{y} - y}{y} \right| \tag{10}$$

$$R^2 = \frac{\sum (\hat{y} - \bar{y})^2}{\sum (y - \bar{y})^2} \tag{11}$$

where y is the true value, $\hat{y}$ is the predicted value, and $\bar{y}$ is the mean of the indicator. When comparing models, a lower value of MAE, MSE, and MAPE indicates better model performance, while an R^2 value closer to 1 signifies a superior model.

Experimental design

Dissolved oxygen is chosen as the target variable for prediction, and both single-step and multi-step predictions are carried out. Based on data correlation analysis, the following four combinations of data have been designed as described in Table 2:

Based on the above 4 data combinations,the experiments are designed as follows:

(i) Window size experiment:Verify the impact of window size on results.
(ii) Model comparison:Compare with mainstream time series prediction models XGBoost, LSTM, GRU, Informer.
(iii) Correlation experiment:Conduct multi-step comparative prediction experiments on four data combinations.
(iv) Ablation experiment:Verify the role of each module through ablation experiment.

No.	Variable	Prediction target
Combination 1	TEMP, pH, PP, TAN, TP, TN, EC, TUB	DO
Combination 2	TEMP, pH, PP, TAN, TP, TN, EC, TUB + Meteorological information	DO
Combination 3	TEMP, pH, EC, TAN	DO
Combination 4	TEMP, pH, EC, TAN + Meteorological information	DO

Table 2. Combination of experimental data. Combination 1 employs the remaining 8 water quality indicators, excluding dissolved oxygen, as input variables. Combination 2 incorporates meteorological data into Combination 1 to assess its influence on the prediction. Combination 3 utilizes the top 4 most strongly correlated indicators as input variables. Combination 4 introduces meteorological data into Combination 3

Experimental results and analysis

In this paper, relevant experiments are conducted in accordance with the aforementioned plan.

(i) Sliding Window Size Experiment: To determine the optimal window size, comparative experiments are performed using window sizes of 8 and 48 for XGBoost, LSTM, GRU, and our proposed model.

Based on the experimental results, it appears that each model demonstrates a low sensitivity to the window size.Taking the R^2 metric as an example,in the XGBoost model, there is only a 2% improvement in prediction results when the window size was increased to 48. However, better prediction results were observed in the other models when the window size was set to 8. Consequently, this paper opts for a window size of 8 in subsequent experiments.

(ii) Popular prediction models commonly used in the field of time series forecasting, namely XGBoost, LSTM, and GRU, are selected for comparison. In the realm of time series forecasting, several popular prediction models are commonly employed for comparative analysis. These models include XGBoost, LSTM, and GRU. In light of the widespread adoption of transformer-based models for time series prediction, Temporal Fusion Transformer (TFT) was introduced by Bryan et al.[22] TFT is capable of learning intricate relationships between different temporal scales within time series data. Building upon this, Jitha et al.[23] leveraged the temporal fusion transformer architecture to model and predict river water quality indicators.

Additionally, Zhou et al.[24] proposed the Informer model for long-term time series prediction. Therefore, we conducted experiments incorporating the Informer model into our comparative analysis.

The comparison experiment is conducted at step sizes of 1 (4 hours), 6 (1 day), 12 (2 days), and 18 (3 days). The results are presented in Table 3, with the optimal results are in bold.

According to the results, the proposed model in this paper consistently achieves the best prediction performance at step 1, 6 and 12 in Combination 1, with improvements in R^2 of 5%, 7%, 5% compared to the second-best model. And in step 18, the model achieved a second-best result, with a difference of only 0.01 from the optimal value. When meteorological data is introduced (Combination 2), there is a little enhancement in prediction performance observed for any of the models, and the R^2 values remain relatively consistent across different step sizes. Notably, the proposed model continues to deliver optimal results at step sizes of 1, 6, and 12. At the step 18,Informer performed slightly better than our proposed model, proving the advantage of the informer in long-term prediction.

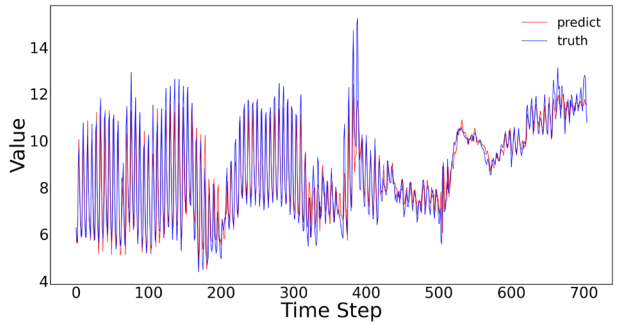
As the prediction step size increases, the forecasting performance of various models tends to decline. However, the proposed model consistently achieves the best results across nearly all step sizes, demonstrating its efficacy in dissolved oxygen prediction.

Examining the 1-step prediction curve, it is evident that the proposed model in this paper provides a better fit to the actual values, with the curves nearly overlapping the true values. The curves are depicted in Fig. 5.

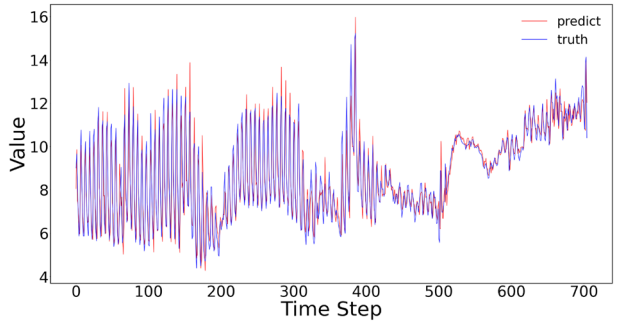
(iii) Following correlation analysis, the top 4 most strongly correlated indicators are selected and utilized in conjunction with the proposed model for multi-step prediction. The results are presented in Table 4, with the optimal value are in bold for reference.

Method	Metric	Combination 1				Combination 2			
		@1	@6	@12	@18	@1	@6	@12	@18
XGBoost	MSE	0.5690	0.8248	1.1159	1.2988	0.5863	0.7754	1.0562	1.2617
	MAE	0.5380	0.6502	0.7675	0.8400	0.5291	0.6298	0.7500	0.8317
	MAPE	0.0641	0.0778	0.0917	0.1012	0.0626	0.0756	0.0904	0.1002
	R^2	0.8597	0.7954	0.7217	0.6746	0.8554	0.8077	0.7366	0.6839
LSTM	MSE	0.4278	0.6828	1.0316	1.2628	0.4398	0.6940	1.0076	**1.2596**
	MAE	0.4528	0.6043	0.7354	**0.8071**	0.4682	0.5841	0.7122	**0.8080**
	MAPE	0.0539	0.0726	0.0860	**0.0935**	0.0556	0.0684	0.0826	**0.0928**
	R^2	0.8958	0.8325	0.7451	0.6861	0.8928	0.8298	0.7512	0.6870
GRU	MSE	0.4325	0.7187	1.0491	1.2644	0.3789	0.7087	1.0579	1.3190
	MAE	0.4498	0.6066	0.7311	0.8303	0.4221	0.5995	0.7324	0.8275
	MAPE	0.0536	0.0722	0.0849	0.0985	0.0505	0.0706	0.0842	0.0953
	R^2	0.8946	0.8237	0.7408	0.6857	0.9077	0.8261	0.7387	0.6721
Informer	MSE	0.4593	0.8761	1.2524	1.3496	0.3337	0.6643	1.3052	1.3650
	MAE	0.5354	0.7435	0.8839	0.9179	0.4260	0.6286	0.9033	0.9097
	MAPE	0.0745	0.0826	0.0921	0.1051	0.0654	0.0859	0.0962	0.1324
	R^2	0.4966	0.6624	0.7363	**0.7142**	0.3945	0.4996	0.7258	**0.7622**
Ours	MSE	**0.2306**	**0.4521**	**0.8685**	**1.1987**	**0.2037**	**0.5051**	**0.7978**	1.2857
	MAE	**0.3477**	**0.4921**	**0.6917**	0.8192	**0.3389**	**0.5011**	**0.6216**	0.8150
	MAPE	**0.0417**	**0.0619**	**0.0845**	0.1021	**0.0406**	**0.0583**	**0.0721**	0.0947
	R^2	**0.9438**	**0.8892**	**0.7859**	0.7028	**0.9504**	**0.8763**	**0.8034**	0.6809

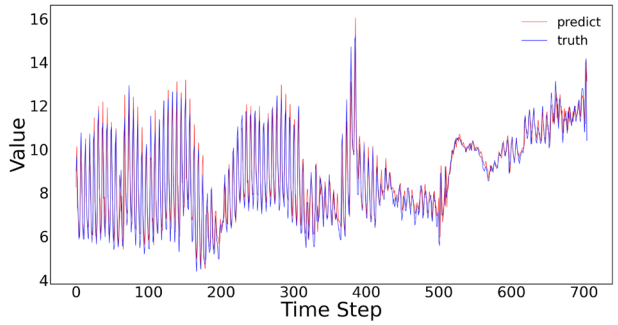
Table 3. Experiment results of model multi-step comparison.



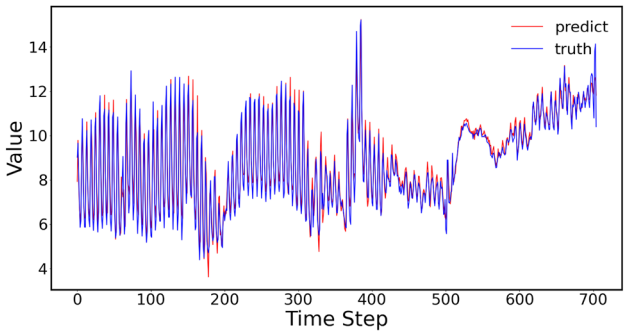
(a) XGboost@1



(b) LSTM@1



(c) GRU@1



(d) Ours@1

Figure 5. Comparison of predicting curves.

It is evident that the prediction accuracy remains relatively consistent even after indicator screening based on correlation analysis. Specifically, Combination 3 achieves the second-best R^2 value in 1-step prediction, while Combination 4 attains the optimal R^2 value in 6-step prediction.

In summary, the selection of indicators that are highly correlated with the target allows for a reduction in data dimensionality without significantly compromising the model's performance. The proposed model, when incorporated with these correlated indicators, continues to deliver robust multi-step dissolved oxygen forecasting.

Combination	Metric	@1	@6	@12	@18
Combination 1	MSE	0.2306	0.4521	0.8685	**1.1987**
	MAE	0.3477	0.4921	0.6917	0.8192
	MAPE	0.0417	0.0619	0.0845	0.1021
	R^2	0.9438	0.8892	0.7859	**0.7028**
Combination 2	MSE	**0.2037**	0.5051	**0.7978**	1.2857
	MAE	0.3389	0.5011	**0.6216**	0.8150
	MAPE	**0.0406**	**0.0583**	**0.0721**	0.0947
	R^2	**0.9504**	0.8763	**0.8034**	0.6809
Combination 3	MSE	0.2224	0.4738	0.9565	1.5275
	MAE	**0.3349**	0.4931	0.7060	0.8699
	MAPE	0.0414	0.0606	0.0859	0.1037
	R^2	0.9458	0.8839	0.7642	0.6217
Combination 4	MSE	0.3048	**0.4473**	0.8466	1.2323
	MAE	0.3956	**0.4859**	0.6535	**0.7924**
	MAPE	0.0481	0.0597	0.0780	**0.0909**
	R^2	0.9257	**0.8904**	0.7912	0.6944

Table 4. Experiment results of correlation analysis.

This approach enables more efficient water quality modeling by utilizing fewer but informative variables, thereby streamlining the modeling process.

(iv) Ablation Experiment: To further substantiate the contributions of individual modules within the proposed model, corresponding ablation experiments have been devised. The results are presented in Table 5, with the optimal value highlighted by bold for clarity.

It is evident that the inclusion of the CNN module enhances prediction performance at step 1. However, its influence diminishes as the step size escalates. Conversely, the introduction of the EEMD decomposition module leads to marked improvements in prediction performance, attaining the second-best results consistently across all step sizes for both Combinations 1 and 2. This underscores that EEMD contributes more significantly towards enhancing predictions compared to the CNN module.

Discussion and conclusion

Given the seasonal, periodic, uncertain, nonlinear, and intricate interdependencies among indicators within water environmental monitoring data, this paper introduces a hybrid CNN-BiLSTM model integrated with EEMD decomposition for water quality data prediction.

The EEMD decomposition technique is highly effective in mitigating noise interference within the data. Additionally, the four resulting modes from this decomposition process augment the data available for model

Method	Metric	Combination 1				Combination2			
		@1	@6	@12	@18	@1	@6	@12	@18
BiLSTM	MSE	0.4275	0.6730	1.0254	1.2852	0.4425	0.7167	1.0086	1.2935
	MAE	0.4538	0.5842	0.7332	**0.8151**	0.4612	0.5952	0.7171	0.8311
	MAPE	0.0536	0.0689	**0.0863**	**0.0946**	0.0542	0.0696	0.0835	0.0967
	R^2	0.8958	0.8349	0.7466	0.6805	0.8922	0.8242	0.7509	0.6785
CNN-BiLSTM	MSE	0.4062	0.8011	1.1218	1.4227	0.4033	0.8278	1.1848	1.3846
	MAE	0.4373	0.6526	0.7709	0.8775	0.4449	0.6552	0.7913	0.8653
	MAPE	0.0518	0.0781	0.0910	0.1044	0.0533	0.0774	0.0930	0.1037
	R^2	0.9010	0.8035	0.7229	0.6463	0.9017	0.7969	0.7072	0.6559
EEMD-BiLSTM	MSE	0.2557	0.5120	0.8751	1.2395	0.2629	0.5400	0.8999	**1.1808**
	MAE	0.3464	0.4881	0.6699	0.8187	0.3756	0.5071	0.6680	**0.7855**
	MAPE	0.0340	0.0577	0.0823	0.0989	0.0451	0.0610	0.0801	**0.0916**
	R^2	0.9377	0.8745	0.7843	0.6925	0.9359	0.8677	0.7783	**0.7072**
Ours	MSE	**0.2306**	**0.4521**	**0.8685**	**1.1987**	**0.2037**	**0.5051**	**0.7978**	1.2857
	MAE	**0.3477**	**0.4921**	**0.6917**	0.8192	**0.3389**	**0.5011**	**0.6216**	0.8150
	MAPE	**0.0417**	**0.0619**	0.0845	0.1021	**0.0406**	**0.0583**	**0.0721**	0.0947
	R^2	**0.9438**	**0.8892**	**0.7859**	**0.7028**	**0.9504**	**0.8763**	**0.8034**	0.6809

Table 5. Experiment results of ablation experiments.

training, thereby enhancing the training efficacy of the model. The incorporation of CNN enables the model to excel in extracting local features, and its integration with BiLSTM facilitates the utilization of bidirectional data and the acquisition of higher-level features, collectively bolstering prediction performance.

Based on prediction experiments conducted on the dissolved oxygen indicator, the proposed model in this paper demonstrates superior prediction performance compared to existing models. This constitutes a valuable exploration of the practical applications of artificial intelligence technology in the realm of water resource protection. In future, the determination of modal quantity in EEMD, data augmentation for water quality data and and the application of Transformers in long-term water quality data prediction would be beneficial research directions.

In conclusion, the proposed hybrid deep learning approach provides an effective solution for precise multi-step water quality forecasting, capable of addressing the intricate attributes of water environment data. The findings underscore the viability of harnessing advanced AI techniques to enhance environmental modeling and conservation efforts.

Data availibility

The datasets used and analyzed during the current study are available from the corresponding author upon reasonable request.

Received: 26 September 2023; Accepted: 11 January 2024
Published online: 19 January 2024

References

1. Xueqing, L. *et al.* Research on regional water quality prediction method based on multi-source data machine learning. *Water Conserv. Hydropower Technol.* **11**, 152–163 (2021).
2. Yan, F. Improvement and application of river water quality evaluation and prediction methods. Master's Dissertation. *Northeast Agricultural University*, China (2017) .
3. Meidan, C., Qi, Y. & Xu, A. WASP water quality model and its research progress. *Water Sci. Technol. Econ.* **07**, 420–426 (2006).
4. Mengchang, H., Xuejun, W. & Lining, S. Review of research progress in water quality model and watershed management model WARMF. *Progress Water Sci.* **02**, 289–294 (2005).
5. Yueling, W. Research progress on comprehensive evaluation and prediction of water quality. *Anhui Agric. Sci.* **02**, 23–26 (2020).
6. Haith, D. A. & Shoemaker, L. L. Generalized watershed loading functions for stream-flow nutrients. *Water Resour. Bull.* **23**, 471–478 (1987).
7. Zhizhen, W. *Application of Grey System and Fuzzy Mathematics in Environmental Protection* (Harbin Institute of Technology Press, 2007).
8. Qiyi, T. & Mingguang, F. *DPS Data Processing System—Experimental Design, Statistical Analysis and Data Mining* (Science Press, 2007).
9. Jiahui, X. *et al.* Surface water quality prediction model based on graph neural network. *J. Zhejiang Univ.* **4**, 601–607 (2021).
10. Hocheriter, S. & Schmidhuber, J. Longshort-term memory. *Neural Comput.* **8**, 1735–1780 (1997).
11. Hu, Z. *et al.* A water quality prediction method based on the deep LSTM network considering correlation in smart mariculture. *Sensors* **6**, 1420 (2019).
12. Yiting, Z. & Tianhong, L. Research on river water quality prediction based on long short-term memory neural network. *Environ. Sci. Technol.* **8**, 163–169 (2021).
13. Mingwei, Z., Zhengquan, L. & Zhihao, F. Water quality prediction model based on CNN-LSTM optimized by quantum particle swarm optimization. *J. China Univ. Metrol.* **3**, 303 (2022).
14. Zhibo, W., Zhongqiu, J. & Tianshu, Z. Study on water quality prediction model of BaiMaho based on CNN-LSTM. *Comput. Knowl. Technol.* **26**, 11–13 (2022).
15. Meixue, Y. *et al.* Seq2Seq water quality prediction model based on wavelet denoising and LSTM. *Comput. Syst. Appl.* **6**, 38–47 (2022).
16. Schafer, B. *et al.* Fluctuations of water quality time series in rivers follow superstatistics. *iScience* **24**, 102881 (2021).
17. Otero, J. F. A. *et al.* EMD-based data augmentation method applied to handwriting data for the diagnosis of essential tremor using LSTM networks. *Sci. Rep.* **12**, 12819 (2022).
18. Weirui, B., Yiqiang, Y. & Xueqin, Z. Water quality prediction model based on VMDLSTNet. *Sci. Technol. Eng.* **22**, 9881–9889 (2022).
19. Zhao-hua, W. & Huang, N. E. Ensemble empirical mode decomposition: A noise assisted data analysis method. *Adv. Adapt. Data Anal.* **1**, 1–41 (2009).
20. Siami-Namini, S., Tavakoli, N., & Namin, A. S. The performance of LSTM and BiLSTM in forecasting time series. in *IEEE International Conference on Big Data (Big Data)* (2019).
21. Weihui, H. *et al.* The dissolved oxygen standard of the United States and its enlightenment to China. *Environ. Sci. Res.* **6**, 1338–1346 (2021).
22. Lim, B. *et al.* Temporal Fusion Transformers for interpretable multi-horizon time series forecasting. *Int. Inst. Forec.* **37**, 1748–1764 (2021).
23. Nair, J. P. & Vijaya, M. S. Temporal fusion transformer: A deep learning approach for modeling and forecasting river water quality index. *Int. J. Intell. Syst. Appl. Eng.* **10**, 277–293 (2023).
24. Zhou, H. *et al.* Informer: Beyond efficient transformer for long sequence time-series forecasting. *AAAI* **35**, 11106–11115 (2021).

Acknowledgements

This work was supported by the university-industry collaboration project "Intelligent Water Environment Monitoring Technology Research",No.HX202109040001.

Author contributions

T.Q. contributed to the study concept, design, data acquisition/analysis and critical revision. Z.W. contributed to data acquisition, experiments, interpretation, drafting. L.D. contributed to design and critical revision. D.S. contributed to the data acquisition.All authors have read and approved the manuscript.

Competing interests
The authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to T.Q.

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.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2024