IOP Conference Series: Earth and Environmental Science

Prediction of dissolved oxygen content in water based on EEMD-Pearson and LSTM hybrid models

To cite this article: Qihua Li *et al* 2021 *IOP Conf. Ser.: Earth Environ. Sci.* **760** 012012

Prediction of dissolved oxygen content in water based on EEMD-Pearson and LSTM hybrid models

Qihua Li[a], Xin WANG[b], Jiangying Wang, Yun Zhou

College of Mathematics and Computer Science, Guangdong Ocean University,
Zhanjiang 524088, China

[a]e-mail: 756483412@qq.com, [b]email: 2681387129@qq.com

Abstract: Improving the accuracy of dissolved oxygen (DO) prediction and establishing a water body DO prediction model are of great importance in water environment pollution management and planning management. In this paper, we propose a hybrid model (EEMD-Pearson-LSTM) of ensemble empirical modal decomposition-Pearson analysis and long-short memory neural network (LSTM), which firstly uses EEMD to decompose the non-stationary dissolved oxygen data into several sub-series that are easy to analyze, and secondly uses Pearson correlation analysis method to The screened subsequences are input to the LSTM network for training and prediction. By establishing the conventional LSTM model, EEMD-LSTM model, EEMD-BP model, and EEMD-Pearson-BP model for comparison under different time periods, root mean square error (RMSE), mean absolute error (MAE), mean absolute percentage error (MAPE), and coefficient of determination (R2) were used as evaluation indicators. In predicting the first 90 days of data, the RMSE, MAE, MAPE, and R2 of the EEMD-Pearson-LSTM model were 0.2355, 0.1893, 2.4710, and 0.8787, respectively, which were optimized by 37.88%, 35.44%, 37.42%, and 28.15%, respectively, compared with the traditional LSTM model, and the EEMD- LSTM model by 13.74%, 16.46%, 16.82%, and 4.98%, respectively, and the error of EEMD-BP network by 23.93%, 22.70%, and 24.32%, respectively, and its R2 by 11.17%, and the error of EEMD-Pearson-BP network by 18.62%, 14.07%, and 14.44%, and its R2 improved by 7.58%. To further demonstrate the advantages of EEMD-Pearson-LSTM, the prediction models for 30-day and 60-day time periods were selected for comparison, and the results showed that EEMD-Pearson-LSTM outperformed other models for the prediction of dissolved oxygen content in different time periods.

1. Introduction

Dissolved oxygen content is a comprehensive index reflecting the purification capacity of water quality, and the prediction of dissolved oxygen content (Cdo) in water bodies is of great significance in water environment pollution management [1] and planning management.

There have been more studies on the prediction of dissolved oxygen content, Rankovic [2] et al. used the levenberg-marquardt (LM) algorithm for training Feed-Forward Neural Net (FNN) to achieve Cdo prediction by input PH and temperature, due to the using two-stage training mode, the upper limit of model prediction ability is low.Olyaie et al [3] used Support Vector Machines (SVM) model to predict Cdo and the results showed that better model prediction can be obtained using SVM, but a single SVM model has some limitations. Jing Wu et al [4] proposed a combined differential autoregressive moving average (ARIMA) model and genetic algorithm optimized wavelet neural network (GAWNN) Combined model of Cdo prediction method, the method has better results for river Cdo prediction compared to individual models, but Cdo data series generally have non-stationary characteristics, and ARIMA model is not suitable for non-stationary series. Shi Pei et al [5] proposed a Cdo prediction model based on General Regression Neural Network (GRNN), Elman neural network, and the experimental results showed that both networks have higher prediction accuracy and avoid the disadvantage that the BP prediction model is easy to fall into the local maximum-minimum. Zhu Nanyang et al [6] proposed a prediction model to improve its estimation accuracy for the low Cdo case by optimizing the loss function in LSTM backpropagation based on the Long Short-Term Memory (LSTM) model, which adjusts the weights of the network by choosing the sin function to assign different weights to the Cdo at different contents, and by This model is experimentally proven to improve the prediction accuracy of low Cdo. Liang Jian et al [7] proposed a wavelet transform combined with SVM model for Cdo prediction in water bodies, which used wavelet decomposition method to decompose the Cdo time series, and then input the series into the SVM model through phase space reconstruction, and finally superimposed to get the prediction value. Although this method has some improvement on dissolved oxygen prediction, wavelet decomposition needs to artificially select the number of decomposition layers, and thus has some limitations. Yu Chengzhou et al [8] proposed a combination of Ensemble Empirical Mode Decomposition (EEMD) and SVM to model the Cdo data of natural water bodies in the north hot spring control section of Jialing River, and the original Least Square Support Vector Machines (LSSVM) model for comparison and analysis, the accuracy of the model is improved, but the SVM is more sensitive to the quality of the data, and the water body Cdo generally has the characteristics of non-stationary and high complexity, and there is no secondary processing of the sequence after EEMD decomposition. Chen Li et al [9] used EEMD to decompose the Cdo sequences, and the decomposed sequences were reconstructed into high-frequency terms, medium-frequency terms, low-frequency terms and trend terms by correlation analysis, and modeled by Least Square Support Vector Regression (LSSVR) and optimized BP networks, respectively, and then the prediction results were superimposed to derive the model, which had better Cdo prediction performance. Yi-Min Lu [10] proposed the method of time series decomposition and Elman neural network, and the combined model was used to predict the Cdo data at the Nan-Ying site in the Jinjiang River basin, and its RMSE was 0.31, MAE was 0.20, and MAPE was 2.50, which was a large improvement compared with the single model. However, the method did not take into account the effect of low correlation series on the model.

Given that the biggest advantage of EEMD is that it can automatically decompose non-stationary time series data into multiple stationary sub-series, making it possible to decompose highly complex series into simpler ones and facilitate prediction. Then, through Pearson correlation analysis, the sequences with higher correlation are filtered and input into the LSTM model. Due to the characteristics of LSTM [11], which has better applicability to long and complex sequence data and is well suited for problems with highly correlated time series, a hybrid model based on ensemble empirical modal decomposition-Pearson analysis (EEMD-Pearson) and LSTM is proposed in this paper to predict Cdo in water bodies.

2. EEMD decomposition principle

Empirical Mode Decomposition (EMD) was proposed by Huang [12], and EMD decomposition can decompose complex non-smooth data into multiple Intrinsic Mode Functions (IMFs) and a residual term. The construction of IMF mainly follows the following two conditions: first, the difference between the number of over-zero and over-polar points is at most 1, and second, the mean value of the upper and lower envelopes is zero. The expressions are as follows.

$$O(t) = \sum_{i=1}^{n} IMF_i(t) + \lambda(t) \quad , i = 1,2,...,n \tag{1}$$

where $O(t)$ is the original signal, $IMF_i(t)$ denotes the first i inner modal component, and $\lambda(t)$ is the residual term. The decomposed $IMF_i(t)$ components and the residual term can be fitted to the original series, $IMF_i(t)$ reflecting the local fluctuation characteristics. With i the increase of IMF, the frequency of fluctuation decreases and finally decomposes into a smoother time series. When there are high frequency noise and abnormal events in the original signal, the EMD decomposition will have the phenomenon of "modal confusion", which will reduce the accuracy of the IMF components and make it impossible to fit the original time series. To solve the "modal confusion" phenomenon of EMD decomposition, Wu [13] et al. proposed the ensemble empirical modal decomposition method EEMD, which overcomes the modal confusion phenomenon in EMD by adding auxiliary noise and effectively suppresses the influence of high frequency noise and abnormal events. The decomposition steps are as follows.

(1) White noise, which follows a normal distribution, is added to the original signal, starting from 1 and increasing. i

$$O_i(t) = O(t) + N_i(t) \tag{2}$$

where, $O_i(t)$ denotes the original signal with white noise, $O(t)$ denotes the original signal, and $N_i(t)$ denotes the white noise.

(2) The decomposition is performed using the EMD pair $O_i(t)$ to obtain n IMF components.

$$O_i(t) = \sum_{i=1}^{n} IMF_i(t) + \lambda(t) \tag{3}$$

(3) The above two steps (1) and (2) are repeated m times until the components do not exceed two extreme values and stop. Then the mean values of the m sets of IMF components and residual terms obtained by decomposition are calculated separately to eliminate the interference of white noise on the IMF series, and the residual terms $\lambda(t)$ are used as trend terms. The final result is $\theta(t)$ denoted by:

$$\theta(t) = \frac{1}{m}\sum_{i=1}^{n} IMF_i(t) \tag{4}$$

3. LSTM structure and principle

Recurrent Neural Network (RNN) [14] is usually considered as a deep learning network for processing time-series data, which is characterized by the fact that the output of the previous moment can be used as the input of the next moment. In the application of time-series data, historical data from different moments can be passed to the input layer of RNN to predict the next moment's data. The network structure of RNN is shown in Figure 1.

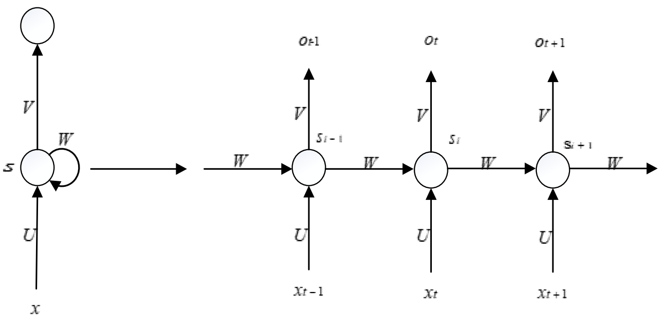


Figure 1. RNN structure diagram

where x, s, and o represent the values of the input layer, hidden layer, and output layer, respectively, and U, V, W and represent the weight matrix of the input layer, the weight matrix of the output layer, and the weight matrix of the hidden layer, respectively. Recurrent neural networks share the weights at different moments in order to reduce the number of computed parameters. In practice, it is found that the current state of the system may be influenced by the state of the system a long time ago, i.e., there is a long-term dependency problem, which cannot be solved by RNN.

Because in RNN algorithm, as the amount of data increases, the computation increases or decreases exponentially, and problems such as gradient explosion or disappearance occur.

To solve the problem of gradient explosion and disappearance of RNN on long-term dependence, Hochreiter et al [15] proposed a long-short memory neural network. The network structure of LSTM is shown in Figure 2.

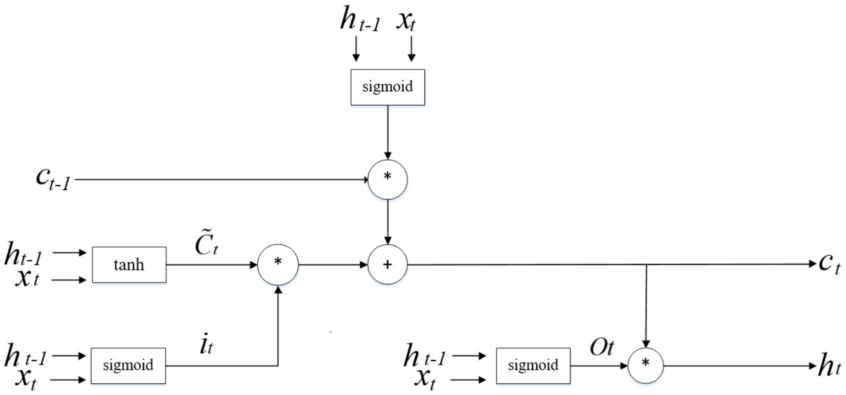


Figure 2. LSTM structure diagram

Analysis of the LSTM structure diagram shown in Figure 2 shows that the historical information x_t and h_{t-1} is decayed and retained under the forgetting gate f_t function; the input gate function i_t suppresses the input information to the hidden layer; $\widetilde{C}_t$ is the input state value, and updates the original matrix information together i_t with the function; the output gate O_t function will select the output part to the output layer. Through the control of forgetting gate, input gate and output gate functions, LSTM overcomes the gradient disappearance problem and mitigates the gradient explosion problem of RNN, whose expressions are as follows.

$$f_t = \sigma(W_f * [h_t - 1, x_t]) + b_f) \tag{5}$$

$$i_t = \sigma(W_i * [h_{t-1}, x_t]) + b_i) \tag{6}$$

$$\widetilde{C}_t = \tanh(W_c * [h_{t-1}, x_t]) + b_c) \tag{7}$$

$$C_t = f_t * C_{t-1} + i_t * \widetilde{C}_t \tag{8}$$

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

$$h_t = o_t * \tanh(C_{t+1}) \tag{10}$$

where W_f, W_c, W_i and W_o are the parameter matrices, σ are the *sigmoid* functions, b_f, b_i, b_o, and b_c are the bias bits, and C_t, h_t are t the outputs at the time.

4. Pearson correlation coefficient

Pearson's correlation coefficient is a quantity to study the degree of linear correlation between variables, and the IMF component is denoted with the raw dissolved oxygen content data as (x_i, y_i) $(i = 1,2,...,n)$, then the Pearson's correlation coefficient equation is as follows.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{11}$$

where $\bar{x}$ and $\bar{y}$ are n the means of the individual data, and r is the correlation coefficient, which indicates the different degrees of correlation of the two series. Its value range is [-1,1], and its grading [16] is shown in Table 1.

Table 1 Correlation coefficient and intensity

Absolute value of correlation coefficient	Related Strength
0.0-0.2	Very weak or no correlation
0.2-0.4	Weak correlation
0.4-0.6	Moderate correlation
0.6-0.8	Strong Related
0.8-1.0	Extremely strong correlation

This study uses Pearson correlation analysis to perform secondary screening of the IMF components to further enhance the temporal and correlation properties of the neural network input data.

5. Dissolved oxygen content prediction model based on EEMD-Pearson-LSTM algorithm

In the prediction of dissolved oxygen content Cdo, due to the characteristics of non-stationary and high complexity of Cdo data, if the model is built directly, the fluctuation degree of Cdo time series cannot be accurately detected, and large errors will inevitably occur in the prediction. the feature of EEMD decomposition is that it can decompose non-stationary series into smoother ones, and when the smooth series is used for prediction modeling, the The prediction error will be significantly reduced, and the decomposed IMF components may be very weakly correlated or uncorrelated with the original Cdo series, leading to a decrease in prediction accuracy. Therefore, this paper uses the EEMD-Pearson method for multimodal decomposition and analysis of Cdo sequences, combines the advantages of LSTM neural network on long and short time series, and proposes a hybrid model of EEMD-Pearson-LSTM, and Figure 3 shows the flow chart of EEMD-Pearson-LSTM prediction.

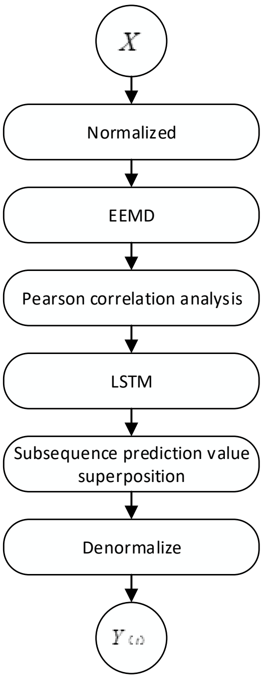

X
Normalized
EEMD
Pearson correlation analysis
LSTM
Subsequence prediction value superposition
Denormalize
Y(n)

Figure 3. EEMD-Pearson-LSTM prediction flow chart

According to Figure 3, the model prediction steps can be seen as follows.

1) The data are X pre-processed and normalized to the range [0,1] using the linear function normalization method, calculated as follows.

$$X_n = \frac{X - X_{\min}}{X_{\max} - X_{\min}} \tag{12}$$

where X_n is the normalized value, and $X_{\max}, X_{\min}$ are the maximum and minimum values of the original data, respectively.

2) The EEMD decomposition is used to obtain several IMF components with a trend term.

3) Secondary screening is performed by Pearson correlation analysis to eliminate the very weakly correlated or uncorrelated components. The training samples, validation samples and test samples are appropriately selected for the screened subsequences, and the LSTM model is used for training, and the IMF subsequences are subsequently predicted separately, and then the model predictions of each subsequence are superimposed to obtain the final results.

4) Inverse normalize the data and reduce the data in the range [0,1] to the original data with the following calculation formula.

$$X = (X_{\max} - X_{\min})*X_n + X_{\min} \tag{13}$$

5) The predicted value $Y(t)$ is the final result.

6. Experiment and analysis

6.1 Data analysis and processing

The data in this paper were obtained from the Open Data Network of Zhejiang Provincial People's Government (http://data. zjzwfw.gov.cn/jdop_front/channal/ data_public.do? deptId=43 &domainId=0), and a total of 1070 dissolved oxygen data were collected from May 21, 2017 to May 4, 2020. A total of 1070 dissolved oxygen data were collected from May 21, 2017 to May 4, 2020, with a data recording interval of 1 day. Among them, anomalous values were detected using the isolated forest method, anomalous values were replaced by the mean replacement method, and missing values were supplemented by the Lagrangian interpolation method. The line graph of the processed dissolved oxygen series is shown in Figure 4, which shows that the Cdo data are volatile and nonlinear. The unit root test (Augmented Dickey-Fuller test, ADF) was applied to the original Cdo data series for the smoothness test, and the test results are shown in Table 2. Where the test obtained 1%, 5%, 10% are the confidence level critical values. From Table 2, the t-value of Cdo is -1.0489, which is greater than the 10% confidence level threshold value of -2.5673, indicating that the original Cdo data series is a non-stationary series.

Table 2 ADF test results of original dissolved oxygen sequence

t-value	1% confidence -level	5% confidence -level	10% confidence -level
-1.0489	-3.4328	-2.8625	-2.5673

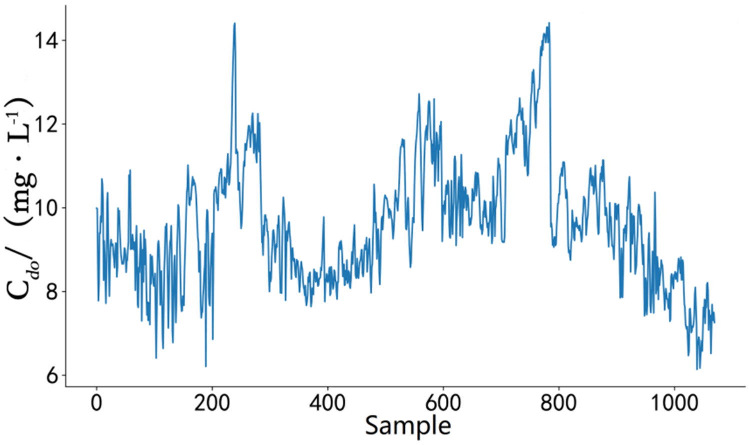


Figure 4. Original Dissolved Oxygen Sequence

EEMD was used to decompose the normalized data into modalities, and the results showed that the Cdo sequence tended to be smooth when the modal number was 7. The Cdo sequence was decomposed into 7 IMF components and 1 residual component (RES), and RES represented the trend of dissolved oxygen. The EEMD method decomposes the Cdo sequence into multiple subsequences, which reduces the complexity of the original sequence and each subsequence includes different scale information, preserving the characteristics of the dissolved oxygen sequence. To further test the stability of the seven IMF components, the same ADF method was used to test the results as shown in Table 3. According to Table 3, the t-values of each IMF component and the remaining components are much smaller than the critical value of -3.4328 at a confidence level of 1%, indicating that the decomposed IMF components are smooth, i.e., the IMF components are smooth series. The experiments demonstrate that the EEMD method can decompose the non-smooth dissolved oxygen series into a smooth series, and show that the IMF component is more suitable to be the input data for LSTM modeling than the original Cdo data.

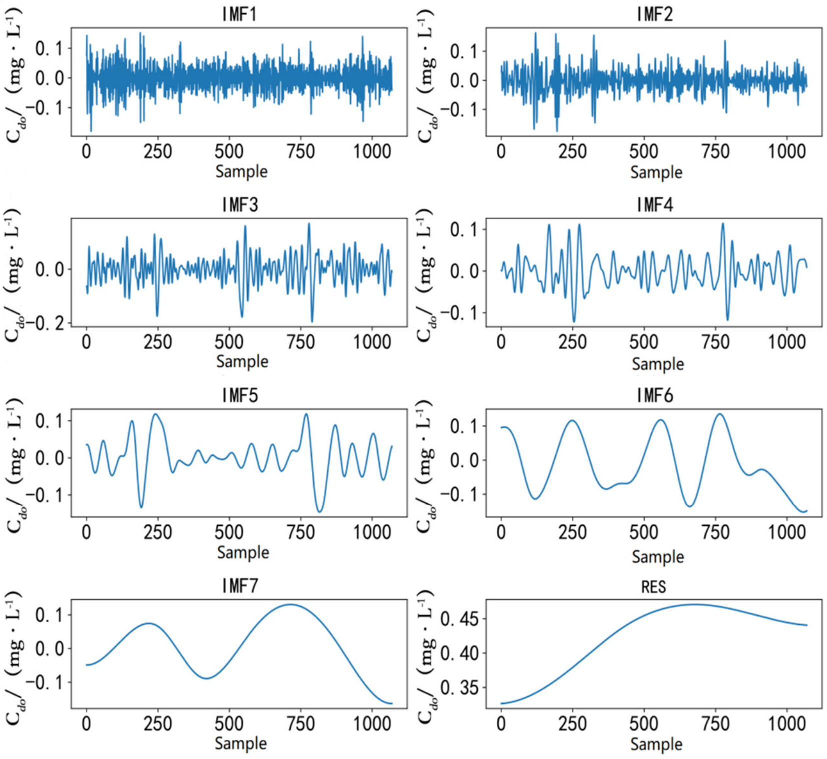


Figure 5. EEMD Modally Decomposed Waves

Table 3 Results of the ADF test for the IMF component

Portion size	IMF1	IMF2	IMF3	IMF4
t	-34.0850	-36.4161	-36.2961	-36.3738
Portion size	IMF5	IMF6	IMF7	RES
t	-44.0961	-26.3743	-92.0751	-9.8257

6.2 Pearson Correlation Analysis

To further analyze the correlation between the IMF component and the original Cdo series, the IMF component derived from the previous section was subjected to Pearson correlation analysis with the original Cdo series, and the results obtained are shown in Table 4.

Table 4 Correlation coefficients of IMF components

IMF Portion	Correlation coefficient
IMF1	0.1515
IMF2	0.2373
IMF3	0.3174
IMF4	0.2826
IMF5	0.3922

IMF6	0.6503
IMF7	0.6929
RES	0.2524

Analysis of Table 4 shows that IMF1 is either very weakly correlated or uncorrelated with the original Cdo series, IMF2, IMF3, IMF4, IMF5, and RES show weak correlation, and IMF6 and IMF7 show strong correlation with the original series. Therefore, in order to reduce the prediction error of the model, this paper chooses to eliminate the very weakly correlated or uncorrelated IMF1 components and select the remaining components for modeling.

6.3 Model Training and Comparative Analysis

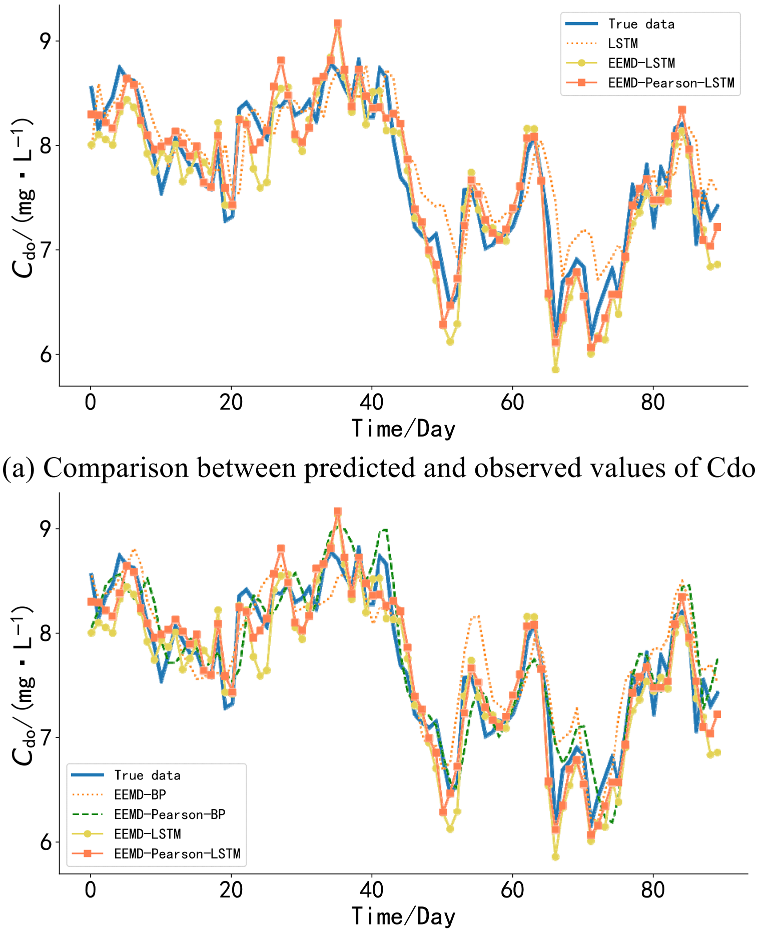
In order to prove the effectiveness of EEMD-LSTM and Pearson hybrid model EEMD-Pearson-LSTM on Cdo prediction, the prediction results are chosen to compare with the conventional LSTM model and EEMD-LSTM model. Since BP network is a classical deep learning network, the model in this paper will compare the prediction results of EEMD-BP and EEMD-Pearson-BP models for comparative analysis to verify the advantages of EEMD-Pearson-LSTM hybrid model in improving the prediction accuracy.

The 1070 Cdo data in the dataset were divided into 3 groups, where the first 80% were used as the training set, 10% as the validation set, and 10% as the test set. The models were trained separately, where the model loss was recorded using Mean Square Error (MSE) for each batch of training samples with the following equation.

$$MSE = \frac{1}{m}\sum_{i=1}^{m}(y-\hat{y})^2 \tag{14}$$

where y is the observed value of the data and $\hat{y}$ is the predicted value. The experimental parameters are set as follows: 1) the model batch_size value is set to 16; 2) the activation function is selected as the relu function; 3) the number of model training rounds (epochs) is set to 100; 4) the Adam optimization algorithm is used for training.

After constructing the five training models, the Cdo for the next 90 days is predicted, and the results are shown in Figure 6. As can be seen from Figure 6(a), when using the conventional LSTM model for prediction, although the LSTM can predict the trend of Cdo, the model shows a certain delay due to the non-stationary and high complexity of Cdo data, which leads to a large error with the original curve. In contrast, the EEMD-LSTM improves the latency problem of the LSTM to some extent, but its prediction for the peaks and valleys of the series shows a large deviation. From Figure 6(b), it can be seen that the trend of the EEMD-BP neural network model can fit the true value, but mostly there is a large error between the predicted and observed values, resulting in a lower accuracy of its prediction. the EEMD-Pearson-BP model has a better curve fit with the original series and achieves a higher prediction accuracy compared to the EEMD-BP model. Compared with other models, the prediction results of EEMD-Pearson-LSTM fit the original data better, and the prediction of peak and trough values is greatly improved compared with other models, so obviously the prediction of Cdo by EEMD-Pearson-LSTM has better accuracy and higher prediction accuracy compared with other models.



(a) Comparison between predicted and observed values of Cdo

(b) Comparison of predicted and observed values of Cdo

Figure 6. Model prediction results for the 90-day period

To further judge the accuracy of the five models, the root mean square error (RMSE), mean absolute error (MAE), mean absolute percentage error, MAPE) and the coefficient of determination (R-Squared, R2) method for the comprehensive evaluation of their merits. The smaller the RMSE, MAE and MAPE, the better the model fit; the larger the R2, the better the model fit. the equations of RMSE, MAE, MAPE and R2 are as follows.

$$RMSE=\sqrt{\frac{1}{m}\sum_{i=1}^{m}(y_i-\widehat{y}_i)^2} \tag{15}$$

$$MAE=\frac{1}{m}\sum_{i=1}^{m}\left|y_i-\widehat{y}_i\right| \tag{16}$$

$$MAPE=\frac{1}{m}\sum_{i=1}^{m}\frac{\left|y_i-\hat{y}_i\right|}{y_i} \tag{17}$$

$$R^2=1-\frac{\sum_{i=1}^{m\text{-}1}(y_i-\hat{y}_i)^2}{\sum_{i=0}^{m-1}(y_i-\overline{y}_i)^2} \tag{18}$$

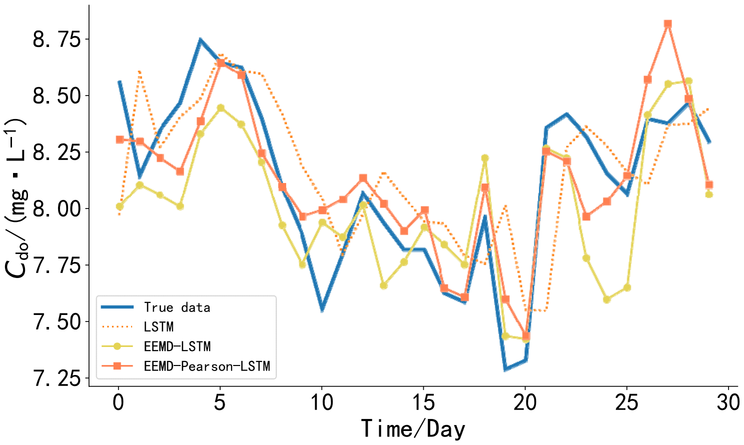
Where, m is the number of samples, y_i is the observed value of Cdo, $\overline{y_i}$ is the sample mean, and $\hat{y}_i$ is the predicted value of Cdo.

The experimental results of the quantitative analysis of model merits by the integrated judging method are shown in Table 5. It can be seen that the values of RMSE, MAE, MAPE, and R2 of the EEMD-Pearson-LSTM model are 0.2355, 0.1893, 2.4710, and 0.8787, respectively, which is a large improvement relative to the LSTM and EEMD-LSTM models. Again By comparing the E EMD-Pearson-BP model, the EEMD-Pearson-LSTM has reduced RMSE by 18.62%, MAE by 14.07%, MAPE by 14.44%, and R2 by 7.58%. Table 5 proves that after Pearson secondary screening, the prediction accuracy of EEMD-BP and EEMD-LSTM models has improved significantly, and all error indicators have been reduced, among which EEMD-Pearson-LSTM has the best performance, which proves that it has a greater advantage in predicting dissolved oxygen content.

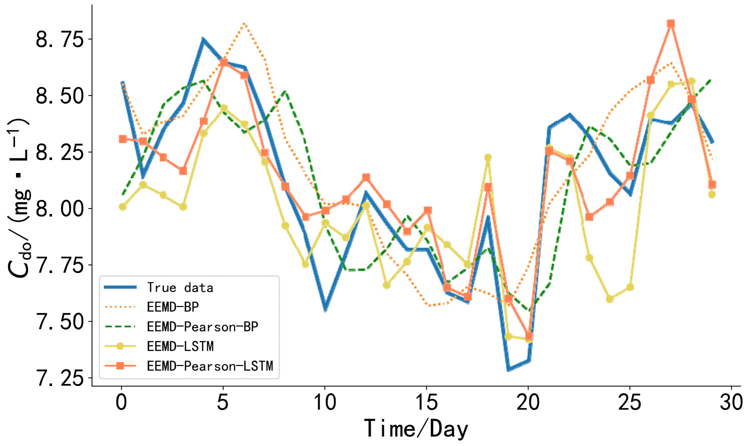
Table 5 Experimental Results Of Model Error

	RMSE	MAE	MAPE	R2
LSTM	0.3791	0.2932	3.9551	0.6857
EEMD-LSTM	0.2730	0.2266	2.9708	0.8370
EEMD-Pearson -LSTM	**0.2355**	**0.1893**	**2.4710**	**0.8787**
EEMD-BP	0.3096	0.2449	3.2649	0.7904
EEMD-Pearson -BP	0.2894	0.2203	2.8881	0.8168

To further compare the prediction adaptation time periods of the above models, two more time periods of 30 days and 60 days were selected for the prediction of dissolved oxygen content, as shown in Figs. 7 and 8. The EEMD-LSTM model also showed a fitting trend, but the peaks and valleys of the prediction curves were steeper than the original series, and the prediction errors were larger. The EEMD-Pearson-LSTM model after secondary screening improves the errors on the troughs and fits better with the original sequence. It can be seen by Figure 8(b) that EEMD-BP has a better fitting trend with lower error index and higher R2 compared to EEMD-LSTM and EEMD-Pearson-BP, but its fit to the original curve is poor compared to EEMD-Pearson-LSTM.



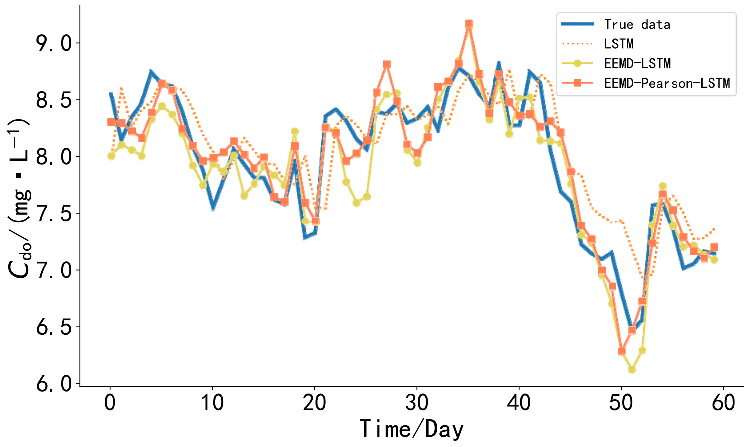
(a) Cdo predicted versus observed values



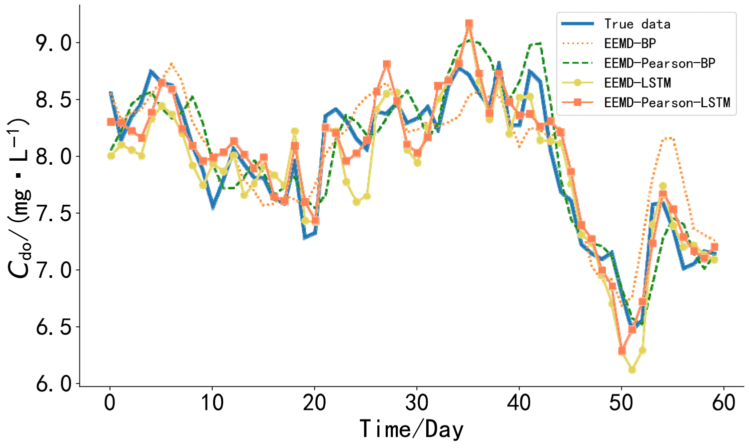
(b) Comparison of predicted and observed values of Cdo

Figure 7. Model prediction results for the 30-day period

The prediction curve of the EEMD-BP network for the 60-day period no longer has an advantage over the EEMD-Pearson-BP network, and its fitting curve shows a large error, while the EEMD-Pearson-LSTM continues to perform optimally.



(a) Comparison between predicted and observed values of Cdo



(b) Cdo predicted values vs. observed values

Figure 8. Model prediction results for the 60-day period

In order to more obviously determine the forecasting effects of the above models in the 30-day and 60-day time periods, the RMSE, MAE, MAPE, and R2 evaluation methods were used to test the forecasting accuracy of the models, respectively. Their evaluation indexes for the 30-day time period and 60-day time period are shown in Table 6. Combining Table 5 and Table 6, it can be seen that the EEMD-Pearson-LSTM has the optimal results for the forecasts in different time periods. The experiment proves that the EEMD-LSTM after secondary screening occupies a greater superiority in predicting the data of different time periods.

Table 6 30-day and 60-day error evaluation index table

Predictive Models	30-day error indicator				60-day error indicator			
	RMSE	MAE	MAPE	R2	RMSE	MAE	MAPE	R2
LSTM	0.3057	0.2305	2.8794	0.3660	0.3376	0.2673	3.4698	0.6876
EEMD-LSTM	0.2760	0.2281	2.7888	0.4834	0.2640	0.2168	2.7308	0.8089
EEMD-Pearson-LSTM	**0.2036**	**0.1615**	**1.9904**	**0.7188**	**0.2318**	**0.1870**	**2.3581**	**0.8527**
EEMD-BP	0.2327	0.1930	2.4117	0.6328	0.2986	0.2358	3.0407	0.7556
EEMD-Pearson-BP	0.2575	0.2023	2.4963	0.5503	0.2590	0.2045	2.5469	0.8161

7. Conclusion

The Cdo time series data of dissolved oxygen content are characterized by non-smoothness and high complexity. In this paper, the EEMD-Pearson-LSTM model is proposed to predict the dissolved oxygen content in three different time periods in the future. It is summarized as follows.

(1) In order to fully exploit the hidden time-series of dissolved oxygen data and improve the prediction accuracy of the model, the non-smooth dissolved oxygen series were first decomposed into smooth IMF components by EEMD, the decomposed subsequences were screened by Pearson's secondary screening, the screened sequences were subjected to LSTM modeling, and finally multiple LSTM model prediction results were derived for superposition to produce the final results. The results show that the EEMD-Pearson-LSTM has a large improvement in prediction results compared with the conventional LSTM and EEMD-LSTM.

(2) To further test the superiority of EEMD-Pearson-LSTM, it is compared with EEMD-Pearson-BP network, which has better fit with the original sequence, smaller prediction error and more accurate curve fitting compared with EEMD-Pearson-BP model.

(3) For the prediction of dissolved oxygen content in 30-day time period, 60-day time period and 90-day time period, EEMD-Pearson-LSTM has the best performance in prediction error index and R2, and has higher prediction accuracy for different time periods compared with other models. The advantages of EEMD-Pearson-LSTM in dissolved oxygen prediction were demonstrated.

In addition to the dissolved oxygen content, the effects of other factors (such as temperature, ammonia nitrogen, rainfall, etc.) on the dissolved oxygen content were not considered, and the inclusion of index data of other factors will be considered in further studies to improve the prediction accuracy of Cdo in the future.

Acknowledgments

This research is financially supported by the National Key Research and Development Program of China (2016YFC1401203), Research start-up projects of Guangdong Ocean University (E16187), and "Innovation Strengthening School Project" Independent Innovation Ability Enhancement Project of Guangdong Ocean University (GDOU2017052501).

References

[1] Xing X F, Xie S Y, Huang M F, Wang Z L,Wu Z L, Sun Z Y. (2018)Characteristics of

Biochemical Parameter and Its Correlation Analysisi in Chudao Island Seawater of WeihaiCity.J, Journal of Ocean Technology, 37(1):54-61.

[2] Rankovi V, Radulovi J, Radojevi I.(2010)Neural network modeling of dissolved oxygen in the Grua reservoir, Serbia.J, Ecological Modelling, 221(8):1239-1244.

[3] Olyaie E, Abyaneh H Z, Mehr A D.(2017)A comparative analysis among computational intelligence techniques for dissolved oxygen prediction in Delaware River.J, Geoscience Frontiers, 8(3): 517-527.

[4] Wu J, Li Z B, Zhu L, Li C. (2017) Hybrid Model of ARIMA Model and GAWNN for Dissolved Oxygen Content Prediction. J, Transactions of the Chinese Society for Agricultural Machinery, 48(S1): 205-210+204.

[5] Shi P,Yuan Y M, Zhang H Y, He Y H.(2017) Application of GRNN and Elman Neural Network in the Prediction of Dissolved Oxygen in Water.J, Jiangsu Agricultural Sciences, 45(23): 217-221.

[6] Zhu N Y, Wu H, Yin D H, Wang Z Q, Jiang Y N, Guo Y.(2019)An improved method for estimating dissolved oxygen in crab ponds based on Long Short-Term Memory.J, Smart Agriculture, 1(03):67-76.

[7] He,T N.(2011)Water Quality Prediction Based On Wavelet Analysis And Support Vector Machine.J, Computer Applications and Software,28(02):83-86.

[8] YU Z C, LI Y,BAI Y.(2018)DO Prediction Based on Ensemble Empirical Mode Decomposition and Support Vector Machine.J, The Administration and Technique of Environmental Monitoring,30(03):27-31.

[9] Chen Li, Zhenbo Li, Jing Wu. (2018) A hybrid model for dissolved oxygen prediction in aquaculture based on multi-scale features. J, Information Processing in Agriculture, 5(1): 11-20.

[10] Yang L,Wu Y X, Wang J L,Liu Y L.(2018)Research on recurrent neural network.J, Journal of Computer Applications,38(S2):1-6+26.

[12] Huang Norden E., Shen Zheng, Long Steven R.(1998)The empirical mode decomposition and the Hilbert spectrumforn on linear and non-stationary time series analysis.J, Proceedings of the Royal Society of London A,454(1971):903-995.

[13] Wu Z,Huang N E.(2009) Ensemble empirical mode decomposition: a noise-assisted data analysis method.J, Advances in Adaptive Data Analysis,(1):1-41.

[14] Cho K, Van Merrienboer B, Gulcehre C.Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation[C]//Proc of Empirical Methods in Natural Language Processing,2014:1724-1734.

[15] Hochreiter S, Schmidhuber J.(1997)Long Short- Term Memory.J, Neural Computation, 9(8): 1735-1780.

[16] ZHOU G R. (2004) theory of statistics.2nd Ed. Nankai University Press, Tianjin.