Water Quality Prediction of Liuxi River based on ICEEMDAN-WOA-LSTM Model

Ke Yong *, Lihui Wang, Lei Hao

Foshan University of Science and Technology, Foshan, China

* Corresponding Author Email: 18946665975@163.com

Abstract. Water quality prediction is a crucial task for ensuring the safety and sustainability of water sources. It supports water resource managers in making informed decisions and taking appropriate actions to prevent water pollution and maintain water quality standards. Considering the non-linearity, latency, and complexity of water quality data, this study proposes a combined model for water quality prediction based on Improved Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (ICEEMDAN) and Long Short-Term Memory (LSTM) networks. The model decomposes and reconstructs variables from time series data to achieve water quality predictions. Comparative experiments with other network models demonstrate superior performance of the proposed model.

Keywords: ICEEMDAN, LSTM, Water Quality Prediction.

1. Introduction

Water resources are crucial for human daily life and various industries, and the water quality condition in a certain area directly relates to the sustainable development of local industrial and agricultural activities. Through effective information analysis and early prediction of water quality, we can address the impacts of water pollution on socio-economics and the environment in a targeted manner. Water quality prediction has become a focal point of attention for governments and research institutions. Therefore, developing new methods and technologies to improve water quality and predict future trends is of paramount importance. In this context, water quality prediction holds significant significance, and appropriate prediction models have positive practical application value and social impact.

In practical research, the physical, chemical, and biological characteristics of water environments are complex and difficult to accurately capture. Traditional methods have certain limitations in dealing with these nonlinear and non-stationary data. In recent years, machine learning methods, especially artificial neural networks, have demonstrated enormous potential in the field of water quality prediction, owing to their ability to capture complex nonlinear relationships in data. For instance, some scholars have introduced Recurrent Neural Networks (RNNs) to predict single water quality data, but RNNs suffer from the vanishing or exploding gradient problem, leading to inaccurate predictions. Long Short-Term Memory (LSTM) neural networks effectively overcome this issue by utilizing forget gates, input gates, and output gates to process single water quality data. However, a single LSTM neural network model exhibits low prediction accuracy when the sample noise fluctuates significantly. To address the noise fluctuation problem, some scholars have applied Empirical Mode Decomposition (EMD) for denoising. However, EMD suffers from a mode mixing problem during the decomposition process, greatly affecting the prediction accuracy of the model. To tackle this issue, some researchers have introduced Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN), which adds paired positive and negative white noises to the original signal and then averages them out to effectively improve decomposition efficiency. However, residual and pseudo-mode problems arise after introducing white noise. ICEEMDAN effectively addresses this issue. In this study, by aggregating multiple mode decomposition results, the influence of residual noise is reduced, and the understanding of the signal's structure is enhanced, leading to more accurate predictions. In this study, we use the ICEEMDAN-WOA-LSTM method to predict the water quality of Liuxi River in Guangzhou, China. The ICEEMDAN-WOA-LSTM method combines Empirical Mode Dec. The ICEEMDAN-WOA-LSTM

method combines Empirical Mode Decomposition (EMD) and Long Short-Term Memory (LSTM) neural networks.

2. Materials and Methods

2.1. Study location and data

Liuxi River is the mother river of Guangzhou. The basin is abundant in water and is an important source of water for residents in Guangzhou. The dissolved oxygen content in the water has a certain guiding effect on the prevention and control of water pollution [14].

Data: This study collected water quality data from the Guangzhou Municipal Water Resources Bureau from April 2018 to August 2022. Nine water quality monitoring indexes were included: temperature (Temp), pH, dissolved oxygen (DO), electrical conductivity (EC), turbidity (TB), high salinity index (CODm), ammonia nitrogen (NH3-N), total phosphorus (TP) and total nitrogen (TN). According to the water quality of Liuxi River in Guangzhou, dissolved oxygen (used to reflect the organic pollution in water quality) was selected in this study. After the data were preprocessed by removing the missing values and outliers, the remaining data was divided into the training set and the test set, with a ratio of 7: 3.

When data collection is missing or abnormal data, interpolation method is used to fill in the missing data; Outliers are detected and eliminated. There are many factors affecting dissolved oxygen. In order to eliminate the error caused by different dimensions, the data is processed by normalization.

2.2. Data preprocessing

Some reasons lead to abnormal or missing data during data collection, which affects the prediction accuracy of the model. In order to reduce the impact of these objective reasons, this study uses two methods to deal with lost data and abnormal data. In view of the absence of original data, linear interpolation method is selected, which can keep the original regularity and trend development of data well. The calculation formula of linear interpolation method is as follows

$$X_k = X_w + (X_r - X_w)\ (\text{k-w})/(\text{r-w}) \tag{1}$$

Where: $\boldsymbol{X_k}$ is the missing value, $\boldsymbol{X_w}$ is the nearest known value before the missing value; $\boldsymbol{X_r}$ is the nearest known value after $\boldsymbol{X_k}$.

To solve the problem of data anomalies, these anomalies are detected and eliminated by using data anomaly monitoring method. The water quality data after processing is non-zero value, and the water quality data of 3300 monitoring points in total. The data set is divided into training set and test set according to the ratio of 7: 3, and the model is trained and its accuracy is verified. The data range of different characteristic data in water quality data is wide. If data is input directly into the model for training, the network will eventually have to learn the distribution of different data at each iteration, which will reduce the training efficiency of the model. Therefore, in order to accelerate the convergence speed and improve the training accuracy, this study selects the range normalization method to normalize the data, and scales all the input values to the interval range [0, 1] to eliminate the impact of data span on the model. The calculation formula is shown as equation (2).

Model training again after the completion of the normalized processing it

$$\text{y} = \frac{x - \min(x)}{\max(x) - \min(x)} \tag{2}$$

Where: y represents the result after normalization of each characteristic variable; x is the initial value of the characteristic variable; min(x) is the minimum value of the feature variable; max(x) is the maximum value of the feature variable.

3. Water quality prediction model based on ICEEMDAN-WOA-LSTM

The ICEEMDAN-WOA-LSTM prediction algorithm first performs ICEEMDAN decomposition on the original water quality data, and establishes the LSTM prediction model for the decomposed IMF sub-sequence. Then the predicted value of each IMF component is summed linearly to obtain the final water quality prediction result. This model is used to predict DO in the water environment of convective river. At the same time, CNN, LSTM, EEMD-LSTM and ICEEMDAN-WOA-LSTM were used as the benchmark models for predicting water quality concentration, and were compared with ICEEMDAN-WOA-LSTM model to verify the prediction effect of the model proposed in this study.

3.1. The result before decomposition

The distribution of dissolved oxygen is shown in Fig 1 below

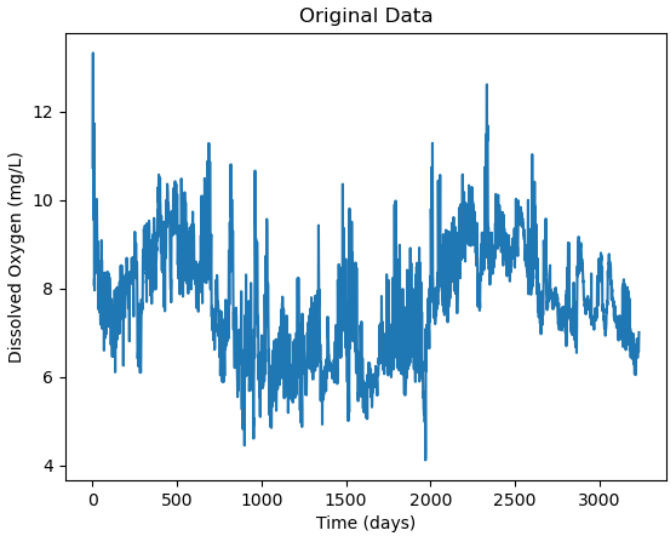


Figure 1. Dissolved oxygen distribution map

Some key trends can be observed in the data distribution map of dissolved oxygen content in the Overflow stream. The data shows that since.

3.2. The result after decomposition

In our experiment, we used the first 3,000 records, 70% of the total number of records, for training the model, and the remaining records were treated as validation sets. We used ICEEMDAN method to decompose the water quality time series data, as shown in Figure 5. As a rule of thumb, the exact number of subsequences is manually set to 9.

3.3. Result analysis

The experimental results of the water quality prediction algorithm based on ICEEMDAN and LSTM are compared with LSTM, EEMD-LSTM andICEEMDAN-LSTM prediction models. Figure 2-4 shows the prediction results.

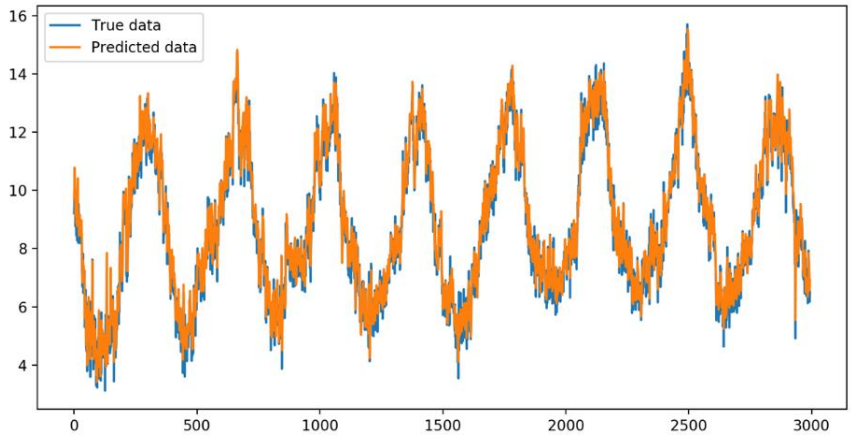


Figure 2. LSTM forecast results

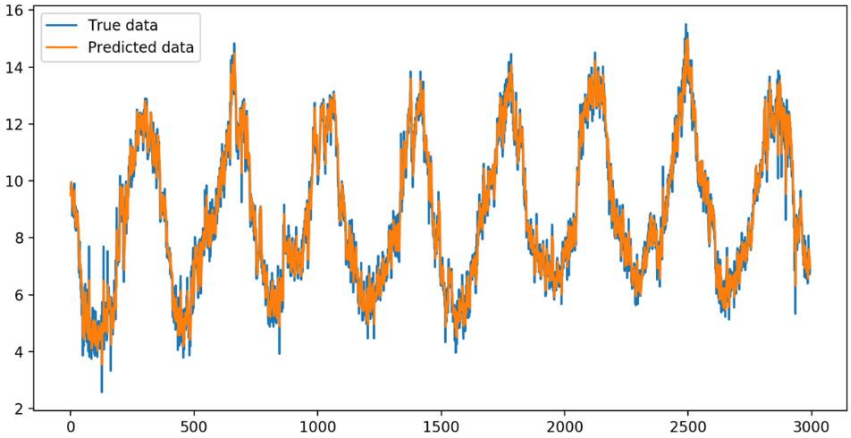


Figure 3. EEMD-LSTM forecast results

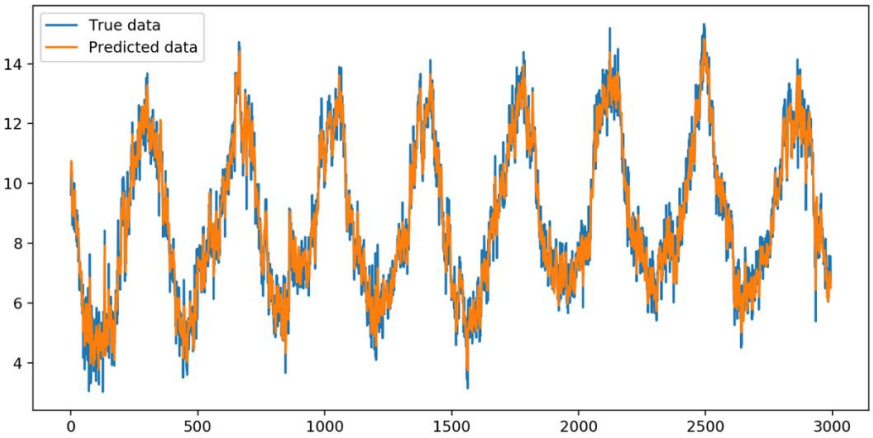


Figure 4. ICEEMDAN-LSTM forecast results

The improved EEMD-LSTM model is highly consistent with the actual data in terms of prediction, especially in dealing with sudden changes in time series. Compared with the case without EEMD method, this model has a smaller prediction error. Compared with the EEMD-CNN model, the proposed method can predict most of the values more accurately and is closer to the real situation. Although the trend of the LSTM model is consistent with the actual value, there is usually a certain delay in the dissolved oxygen prediction, which is shown as translational misalignment on the image. This phenomenon arises from the LSTM network's inability to accurately capture the extent of time series fluctuations, resulting in predictions of previous moments that may appear at the current

moment. After the introduction of EEMD algorithm to enhance the LSTM network, it can be observed that the LSTM delay is improved, making the prediction of extreme values more accurate.

The accuracy evaluation indicators of these different models are shown in Table 2 below. It is evident that the proposed model reduces RMSE by about 46% and MAPE by nearly 47% compared to the LSTM model. There was a 27% reduction in MAPE compared to EEMD-CNN. In summary, the proposed model has the best performance in these four evaluation indicators, demonstrating the superiority and effectiveness of the method.

Table 1. Model evaluation

	LSTM	EEMD-LSTM	ICEEMDAN-LSTM
RMSE	0.0403	0.0221	0.0245
MAE	0.0276	0.0148	0.0211
MAPE	6.4849	4.3267	4.7163
R^2	0.9464	0.9864	0.9558

4. Summary

I Water quality prediction plays an important role in ensuring water quality safety and protection. In this paper, improved ICEEMDAN is used to deal with the nonlinearity and complexity of water quality, whale optimization algorithm (WOA) is used to optimize the model parameters, and long-term memory network is used to predict water quality. Experimental results show the validity of our model.

References

[1] Tejoyadav, M., R. Nayak, and U.C. Pati, Multivariate Water Quality Forecasting of River Ganga Using VAR-LSTM based Hybrid Model, in 2022 IEEE 19th India Council International Conference (INDICON). 2022. p. 1-6.

[2] Moss, B., Water pollution by agriculture. Philosophical Transactions of the Royal Society B: Biological Sciences, 2007. 363 (1491): p. 659-666.

[3] Koklu, R., B. Sengorur, and B. Topal, Water Quality Assessment Using Multivariate Statistical Methods—A Case Study: Melen River System (Turkey). Water Resources Management, 2009. 24 (5): p. 959-978.

[4] Liu, S., L. Xu, and D. Li, Multi-scale prediction of water temperature using empirical mode decomposition with back-propagation neural networks. Computers & Electrical Engineering, 2016. 49: p. 1-8.

[5] Najah Ahmed, A., et al., Machine learning methods for better water quality prediction. Journal of Hydrology, 2019. 578.

[6] Coulibaly, P. and F. Anctil. Real-time short-term natural water inflows forecasting using recurrent neural networks. in IJCNN'99. International Joint Conference on Neural Networks. Proceedings (Cat. No. 99CH36339). 1999. IEEE.

[7] Pascanu, R., T. Mikolov, and Y. Bengio. On the difficulty of training recurrent neural networks. in International conference on machine learning. 2013. Pmlr.

[8] Dheda, D., L. Cheng, and A.M. Abu-Mahfouz, Long Short Term Memory Water Quality Predictive Model Discrepancy Mitigation Through Genetic Algorithm Optimisation and Ensemble Modeling. IEEE Access, 2022. 10: p. 24638-24658.

[9] Nguyen, D.H., J.-B. Kim, and D.-H. Bae, Improving Radar-Based Rainfall Forecasts by Long Short-Term Memory Network in Urban Basins. Water, 2021. 13 (6).

[10] Chen, Z., et al., An Improved Method Based on EEMD-LSTM to Predict Missing Measured Data of Structural Sensors. Applied Sciences, 2022. 12 (18).

[11] Zhang, D., et al. Predicting Water Quality Based on EEMD And LSTM Networks. in 2021 33rd Chinese Control and Decision Conference (CCDC). 2021. IEEE.

[12] Zhang, X., et al., Precipitation forecast based on CEEMD–LSTM coupled model. Water Supply, 2021. 21 (8): p. 4641-4657.

[13] Mohanty, P., Identifying Mode Shapes and Modal Frequencies by Operational Modal Analysis in the Presence of Harmonic Excitation. Experimental Mechanics, 2005. 45 (3): p. 213-220.

[14] Kannel, P.R., et al., Application of Water Quality Indices and Dissolved Oxygen as Indicators for River Water Classification and Urban Impact Assessment. Environmental Monitoring and Assessment, 2007. 132 (1-3): p. 93-110.