scientific reports

Check for updates

OPEN

Application of an improved LSTM model based on FECA and CEEMDAN VMD decomposition in water quality prediction

Jie Long[1], Chong Lu[1]✉, Yiming Lei[1,2], Zhong Yuan Chen[1,2] & Yihan Wang[1]

To address the limitations of existing water quality prediction models in handling non-stationary data and capturing multi-scale features, this study proposes a hybrid model integrating Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN), Variational Mode Decomposition (VMD), Long Short-Term Memory Network (LSTM), and Frequency-Enhanced Channel Attention (FECA). The model aims to improve prediction accuracy and robustness for complex water quality dynamics, which is critical for environmental protection and sustainable water resource management. First, CEEMDAN and Sample Entropy (SE) were used to decompose raw water quality data into interpretable components and filter noise. Then, a VMD-enhanced LSTM architecture embedded with FECA was developed to adaptively prioritize frequency-specific features, thereby improving the model's ability to handle nonlinear patterns. Results show that the model is successful in predicting all six water quality indicators: NH_3-N (ammonia nitrogen), DO (dissolved oxygen), pH, TN (total nitrogen), TP (total phosphorus), and CODMn (chemical oxygen demand, permanganate method). The model achieved Nash-Sutcliffe Efficiency (NSE) values ranging from 0.88 to 0.99. Using dissolved oxygen (DO) as an example, the model reduced the Mean Absolute Percentage Error (MAPE) by 0.12% and increased the coefficient of determination (R2) by 0.20% compared to baseline methods. This work provides a robust framework for real-time water quality monitoring and supports decision making in pollution control and ecosystem management.

Keywords Water quality monitoring, CEEMDAN, VMD, LSTM, FECA, Hybrid model

Water is the foundation of all life on Earth and an indispensable resource for human survival, development, and prosperity. However, with the rapid advancement of industrialization and urbanization, water pollution has become an increasingly critical issue. Moreover, global climate change and the overexploitation of water resources have further exacerbated water stress, heightening the risk of a water crisis[1].

In recent years, machine learning models have gained prominence in water quality evaluation, providing researchers with powerful tools for analysis and prediction[2,3]. Haggerty et al.[4]have noted an significant growth in the use of machine learning in Groundwater Quality Monitoring (GWO) modeling. Jongcheol et al.[5] demonstrated that LSTM networks, as an advanced variant of recurrent neural networks (RNNs), effectively alleviate gradient vanishing, explosion issues and improve prediction accuracy. Comparative studies by Singha et al.[6]revealed that deep learning (DL) models outperform traditional machine learning approaches in groundwater quality prediction. Additionally, Pyo et al.[7]utilized Convolutional Neural Networks (CNN) to predict cyanobacterial concentrations in rivers, demonstrating the feasibility and accuracy of CNNs in water quality monitoring. Sha et al.[8]compared the performance of real-time water quality prediction using various deep learning models and input data preprocessing techniques. This finding suggests that a single model may not fully capture the information embedded in the data, thus limiting predictive performance.

As a result, hybrid models[9–14] have become a promising direction in water quality prediction research. such as CNN-LSTM frameworks[15]and EEMD-LSTM models[16], have further enhanced accuracy by integrating signal decomposition with temporal modeling. For instance, Baek et al.[17] combined CNN and LSTM to simulate water quality indices in river basins, achieving significant improvements in simulation accuracy. Similarly, Luo et al.[16]integrated ensemble empirical mode decomposition (EEMD) with LSTM, the EEMD-LSTM model outperformed the standalone LSTM model (without EEMD preprocessing) in all evaluation metrics.

[1]Present address: School of Information Management, Xinjiang University of Finance and Economics, Urumqi 830012, China. [2]Yiming Lei and Zhong Yuan Chen contributed equally to this work. ✉email: 498841300@qq.com

Despite these advancements, critical challenges persist[18]. Traditional decomposition methods (e.g., EEMD) struggle with non-stationary water quality data, resulting in incomplete feature extraction[19]. Recent studies, such as Hussein et al.[20] on groundwater quality assessment and Bachir et al.[21] on Saf-Saf river modeling, further highlight the limitations of conventional techniques in dealing with high-frequency noise and nonlinear dynamics. Moreover, Complex hybrid architectures (e.g., multistage CNN-GRU) require extensive parameter tuning, hindering real-world deployment[22].Sensitivity to noise and extremes, models often fail under abrupt environmental changes, such as monsoon-induced turbidity in the Selangor River Basin or noisy sensor data[23], compromising reliability.

To address these gaps, this study proposes an innovative hybrid model integrating CEEMDAN, VMD, LSTM, and FECA. The key innovations include adaptive multi-scale decomposition, where CEEMDAN adaptively decomposes nonlinear and non-stationary data into interpretable components, and VMD refines high-frequency signals to minimize information loss—overcoming limitations of traditional methods[24]. Additionally, the FECA module dynamically prioritizes critical frequency components, enhancing robustness to noise and extreme conditions, a significant improvement over static attention mechanisms[25]. Finally, the model achieves computational efficiency without sacrificing accuracy by reducing redundant features and integrating parameter-sharing strategies, addressing bottlenecks in prior hybrid frameworks. These advancements collectively provide a robust and efficient solution for high-frequency water quality prediction.

Materials and methods

Case study

The study area focuses on the Dujiangyan Hydrological Station (103.62°E, 31.01°N) in Sichuan Province, China, a critical monitoring point in the Minjiang River Basin (Fig. 1).

Water quality parameters and data preprocessing

The dataset used in this study is sourced from the National Real-Time Surface Water Quality Inspection and Distribution System (NRWQIS). The input variables for prediction include DO, CODMn, TN, pH, TP, and NH3-N, as shown in Fig. 2. The dataset comprises a total of 1,975 records collected over the period from July 1, 2022, to June 30, 2023, which serve as the basis for the predictive modeling process. In the data preprocessing phase, polynomial interpolation was applied to handle missing values and outliers, which accounted for between 0% and 1.5% of the dataset. To ensure the effectiveness of both model training and validation, the data were split into a training set and a test set. The training set includes 1,875 records spanning from 00:00 on July 1, 2022, to 00:00 on June 14, 2023, used for model training and parameter optimization. The remaining 100 records, covering the period from 04:00 on June 14, 2023, to 16:00 on June 30, 2023, were retained as the test set.

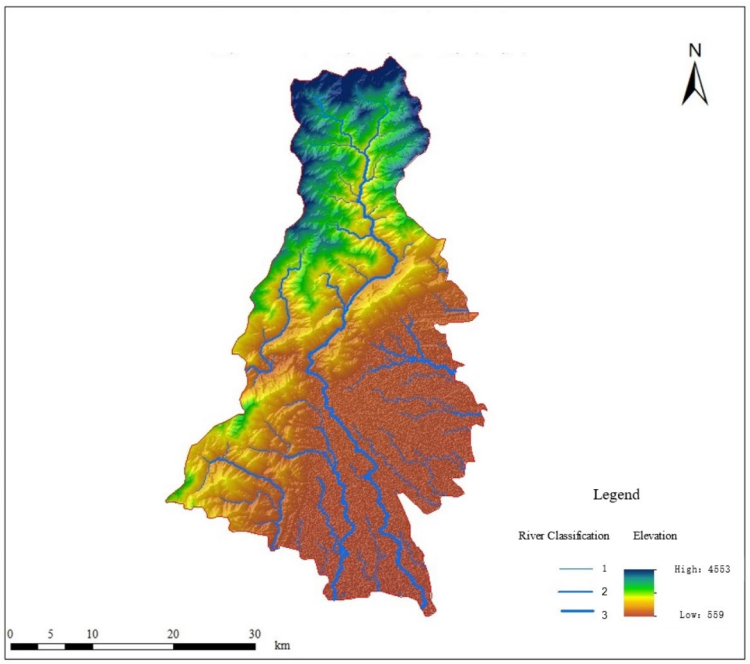

Fig. 1. Location Map of Dujiangyan Hydrological Station(This figure is Generated using ArcGIS Desktop 10.8.URL: https://www.esri.com).

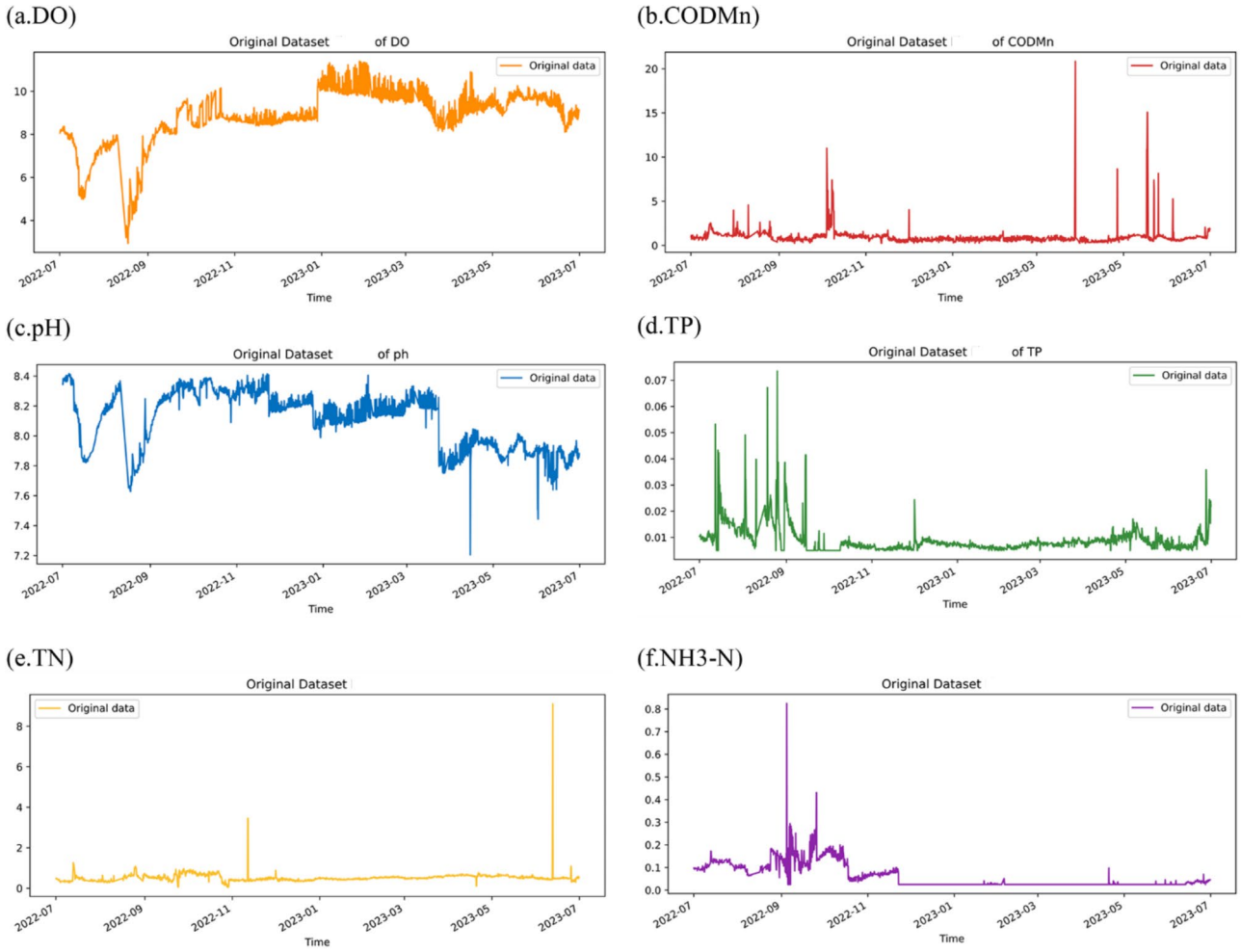

Fig. 2. Historical Distribution Data of Original Water Quality Indicators.

Model architecture and auxiliary approaches

To present more clearly the structure and workflow of the improved LSTM water quality prediction (hereafter abbreviated as CSVLF) model based on FECA and CEEMDAN-VMD decomposition, the design of the CSVLF model architecture is shown in Fig. 3.

CEEMDAN decomposition

CEEMDAN[19] is a novel signal processing method that combines the advantages of Empirical Mode Decomposition (EMD) and EEMD[26]. The specific steps of the CEEMDAN method are as follows:

(1) An initial signal-to-noise ratio α is set and a set of Gaussian white noise sequences $\varepsilon(t)(t = 1, 2, 3, ?, n)$ satisfying a standard normal distribution is generated. Where I is the number of noise sequences and n is the length of the time series.

$$\varepsilon_i(t), \quad t = 1, 2, 3, \ldots, n, \quad i = 1, 2, 3, \ldots, I \tag{1}$$

.

(2)The original time series data were augmented by sequentially adding the generated white noise sequences, resulting in a set of noisy data sequences $f_t\,(t) \,(t = 1,\ 2,\ 3, \ldots\ldots,\ n;\ \ i = 1,\ 2,\ 3, \ldots\ldots\ldots n)$.

$$f_i(t) = f(t) + \alpha \cdot \varepsilon_i(t), \quad t = 1, 2, 3, \ldots, n, \quad i = 1, 2, 3, \ldots, I \tag{2}$$

(3) EMD decomposition is performed on each noisy data sequence to obtain a series of IMFs. The j th IMF is denoted as$E_j^i(t)$:

$$E_j^i(t), \quad t = 1, 2, 3, \ldots, n, \quad j = 1, 2, 3, \ldots, K \tag{3}$$

Where K is the number of IMFs and I denotes the I th noisy data sequence.

(4) Perform ensemble averaging on each IMF to obtain the final one.

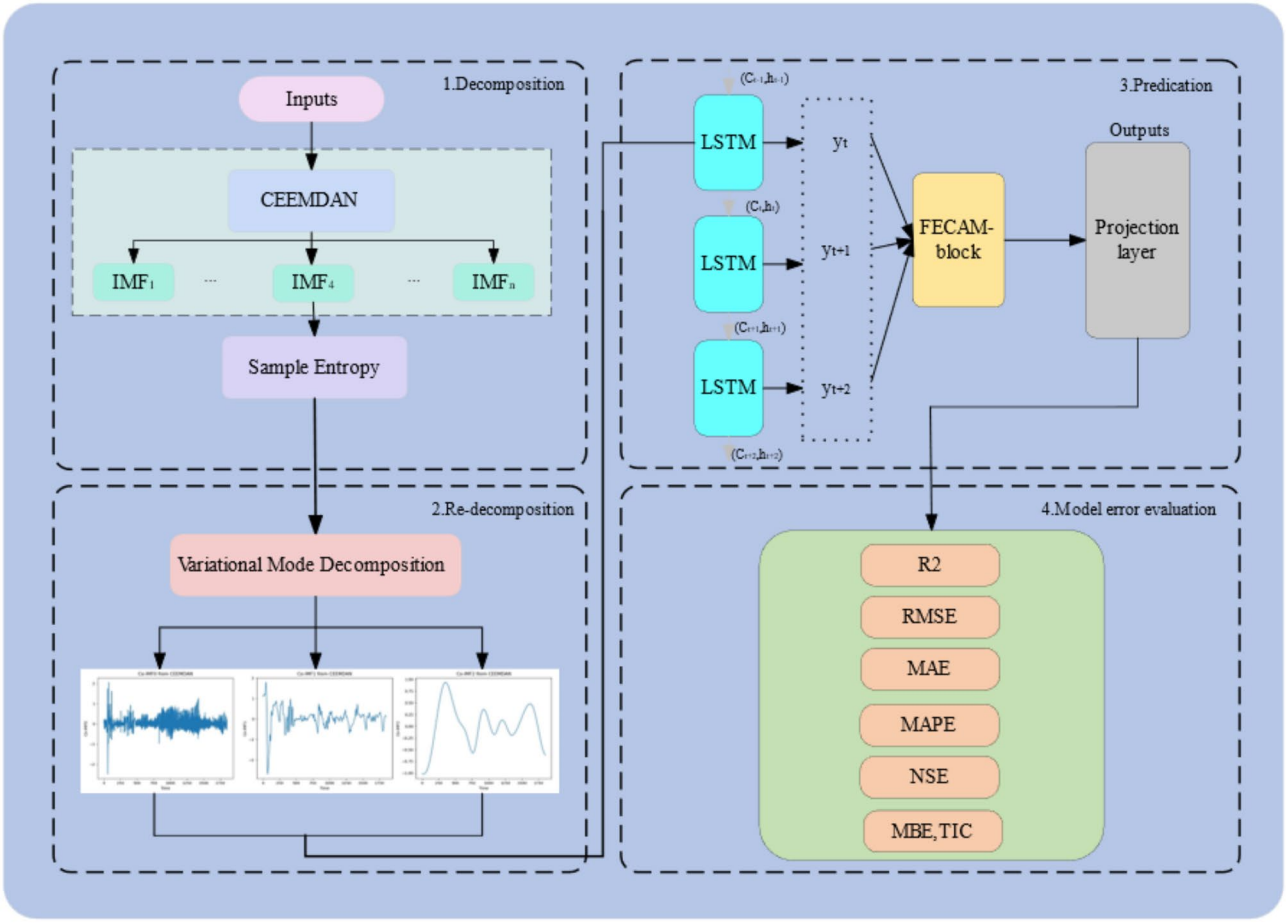

Fig. 3. Architecture Diagram of the CSVLF Model.

$$IMF_j(t) = \frac{1}{I}\sum_{i=1}^{I} E_j^i(t), \quad t = 1, 2, 3, \ldots, n, \quad j = 1, 2, 3, \ldots, K \tag{4}$$

To optimize the decomposition process, the signal-to-noise ratio α can be adjusted based on the specific characteristics of the data. The steps of noise addition, EMD decomposition, and ensemble averaging are repeated iteratively until the desired Intrinsic Mode Functions (IMFs) are obtained.

SE method

Sample Entropy[24]is a tool for measuring the complexity of time series data, which is based on the concept of information entropy and is used to quantitatively assess the stochasticity and predictability of signals. It is defined as:

$$Sample\ Entropy(U) = -lim_{m\to\infty, r\to 0}\frac{1}{N}\sum_{i=1}^{N} logB(m, r) \tag{5}$$

Where U is a discrete-time series, N is the length of the sequence, m is the embedding dimension, r is the fraction of the time series standard deviation corresponding to the time series standard deviation as well as $B(m, r)$ is the ratio of the number of matching template vectors with embedding dimension m and r similarity, to the number of total template vectors.

VMD method

VMD[27]is an advanced signal processing technique that adaptively decomposes a signal into multiple IMFs, with each IMF representing a component of the signal at different frequencies and time scales. VMD aims to minimize the error between the input signal and these IMFs while guaranteeing the diversity and stability of the IMFs. It determines the center frequency of each IMF by minimizing the sum of the estimated bandwidths of the components and using an alternating direction multiplier method.

LSTM model

LSTM is a type of recurrent neural network designed to address the long-term dependency problem in time series data[28].It achieves this through a gated mechanism consisting of three main components: the input gate, forget gate, and output gate. These gates regulate the flow of information, allowing the model to retain relevant long-term dependencies while discarding irrelevant information. In this study, we employ a three-layer LSTM architecture with 128, 64, and 32 units in each layer, respectively. This configuration is optimized to capture both short-term fluctuations and long-term trends in water quality data. The structure and operation of the LSTM model are illustrated in Fig. 4, and the specific formulas for the gates and cell states are as follows:

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f) \tag{6}$$

$$i_t = \sigma(W_i x_i + U_i h_{t-1} + b_i) \tag{7}$$

$$\tilde{c}_t = tanh(W_c x_t + U_c h_{t-1} + b_c), \tag{8}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \tag{9}$$

$$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o) \tag{10}$$

$$h_t = o_t \odot tanh(c_t) \tag{11}$$

Where $x_t$is the water quality input feature vector, $h_t$is the LSTM hidden state vector, and f_t,i_t,$\tilde{c}_t$,c_t,o_t denotes the activation vectors of the forgetting gate, input gate, cell input, cell state, and output gate, respectively. *W* and *U* are weight matrices for input features and previous hidden states respectively, while bare bias terms added before activation. The symbols ⊙denote the Hadamard product.

FECA module

Leverage the advantages of the above VMD-LSTM model, and further introduce the frequency-enhanced channel attention module to improve the model's ability to capture frequency information in time series data. Mathematically, the processing of FECA[25] can be expressed as follows: for an input time series.$X \in R^{C \times T}$., where *C* represents the number of channels and *T* represents the length of the time series, the FECA module first computes the DCT to obtain the frequency spectrum$X\prime\prime$:

$$X\prime = DCT\,(X) \tag{12}$$

.

Subsequently, the channel attention mechanism is applied to obtain the final enhanced frequency spectrum$X\prime\prime$:

$$X\prime\prime = FECA\,(X\prime) = \sigma\,(g\,(X\prime, W)) \odot X\prime \tag{13}$$

Where $g(\cdot)$denotes the attention mechanism with parameter*W*, $\sigma$is the Sigmoid activation function, and ⊙ denotes the element-by-element multiplication.

Benchmark models and evaluation

Evaluation

In evaluating the model, the model employs a variety of criteria to comprehensively measure the model's performance. These criteria include NSE, Mean Bias Error (MBE), Theil's inequality coefficient (TIC), and R^2, Root Mean Square Error (RMSE), Mean Absolute Square Error (MAE), and MAPE. These criteria measure

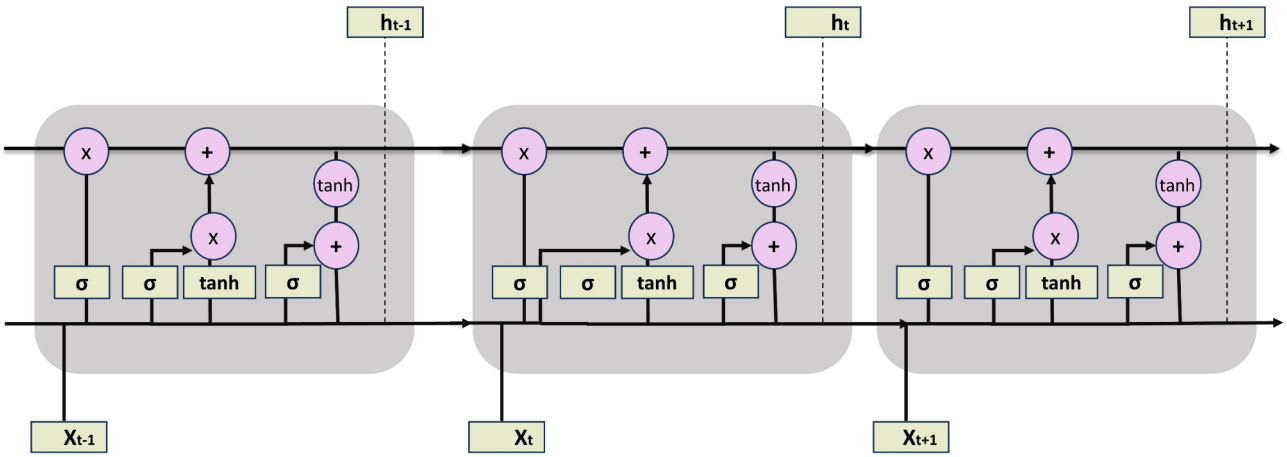

Fig. 4. LSTM Flowchart.

the model's goodness-of-fit, prediction accuracy, and relative error of prediction, respectively, and provide a comprehensive assessment of the model.

$$R^2 = 1 - \frac{\sum_{t=1}^{n} (y_t - \hat{y}_t)^2}{\sum_{t=1}^{n} (y_t - \bar{y})^2}, \tag{14}$$

$$MBE = \frac{1}{n} \sum_{t=1}^{n} (y_t - \hat{y}_t), \tag{15}$$

$$TIC = \sqrt{\frac{\sum_{t=1}^{n} \left(\frac{y_t - \hat{y}_t}{y_t}\right)^2}{\sum_{t=1}^{n} \left(\frac{y_t}{y_t}\right)^2 + \sum_{t=1}^{n} \left(\frac{\hat{y}_t}{y_t}\right)^2}}, \tag{16}$$

$$RMSE(y, \hat{y}) = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \tag{20}$$

,

$$MAE(y, \hat{y}) = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{21}$$

,

$$MAPE(y, \hat{y}) = \frac{100\%}{n} \sum_{i=1}^{n} \frac{|y_i - \hat{y}_i|}{\max(\sigma, |y_i|)} \tag{22}$$

Benchmark models and experimental setup

The proposed model was implemented using Python 3.10, PyTorch 2.1.2, and CUDA 11.8 for GPU acceleration. The model was trained and evaluated on a high performance computing system to ensure efficient processing of the water quality time series data. The input configuration included a dateback of 30 (number of previous time steps used for prediction), periods of 50 (number of future time steps to predict), and a timestep of 1 (interval between consecutive time steps). The LSTM network consisted of three hidden layers of 128, 64, and 32 units, respectively, designed to extract complex features hierarchically. The tanh activation function and a dropout rate of 0.2 were used to prevent overfitting. The model was trained for a minimum of 500 epochs with a batch size of 16 using the Adam optimizer (learning rate = 0.001) and early termination (patience = 50). Training data was randomized before each epoch to improve generalization. Also, to compare the effectiveness of the CSVLF model, the study compares the forecasting performance of five models. The five models are:

Model 1: Single-LSTM: To explore the performance comparison between the hybrid model and the single model.

Model 2: CEEMDAN-VMD-GRU: Explore the role of LSTM by replacing it with GRU.

Model 3: CEEMDAN-ISOS-VMD-LSTM: Replacing SE with ISOS and exploring the role as well as the importance of SE, especially the FECA layer.

Model 4: Respective -LSTM: Explore the role of VMD, SE, and FECA in time series modeling by removing these two components.

Model 5: CEEMDAN-CNN-CBAM-LSTM: Replacing the SE, VMD, and FECA modules with CNN, and CBAM to explore the performance comparison.

Results

Data characteristics and preprocessing results

In the data preprocessing stage, this study analyzed the time series data of several water quality indicators for smoothness using the Augmented Dickey-Fuller (ADF) test. The results showed that the ADF test statistic for DO was − 2.59 with a p-value of 0.09, which exceeded the common significance level of 0.05, indicating that the series was non-stationary. Similarly, the ADF test statistic value for NH3-N was − 1.80 with a p-value of 0.38, which also failed to reject the original hypothesis of non-stationarity. P-value of the ADF test for PH, TP, and CODMn were below 0.05, showing that these series were smooth. In particular, TN had an ADF statistic value of -8.73 with a very low p-value (3.21), clearly indicating a strong smoothness.

In terms of autocorrelation, the results of the Ljung-Box test revealed a p-value of 0.0000 for all series, indicating significant autocorrelation within the data. As presented in the table, the Ljung-Box statistic for DO is 5084.96, and for NH3-N, it is 47466.94. The strong autocorrelation observed in these series necessitates special

Sequence	ADF test statistic	ADF Test *p*-value	Sequence stationarity	Ljung-box test statistic	Ljung-box test *p*-value	Jarque-Bera test statistic	Jarque-Bera test *p*-value	Sequence normality
DO	-2.5974	0.0935	Non-stationary	5084.96	0.0000	2248.2781	0.0000	Non-normal distribution
pH	-3.0976	0.0267	Stable	49129.79146	0.0000	114.2883	1.5227	
TP	-3.8587	0.0024	Stable	13156.4454	0.0000	90135.0773	0.0000	
NH3-N	-1.7997	0.3806	Non-stationary	47466.9427	0.0000	40500.16757	0.0000	
CODMn	-4.9749	2.4882	Stable	2371.5771	0.0000	1461356.0287	0.0000	
TN	-8.7300	3.2102	Stable	1477.6014	0.0000	26095322.947	0.0000	

Table 1. Statistical tests.

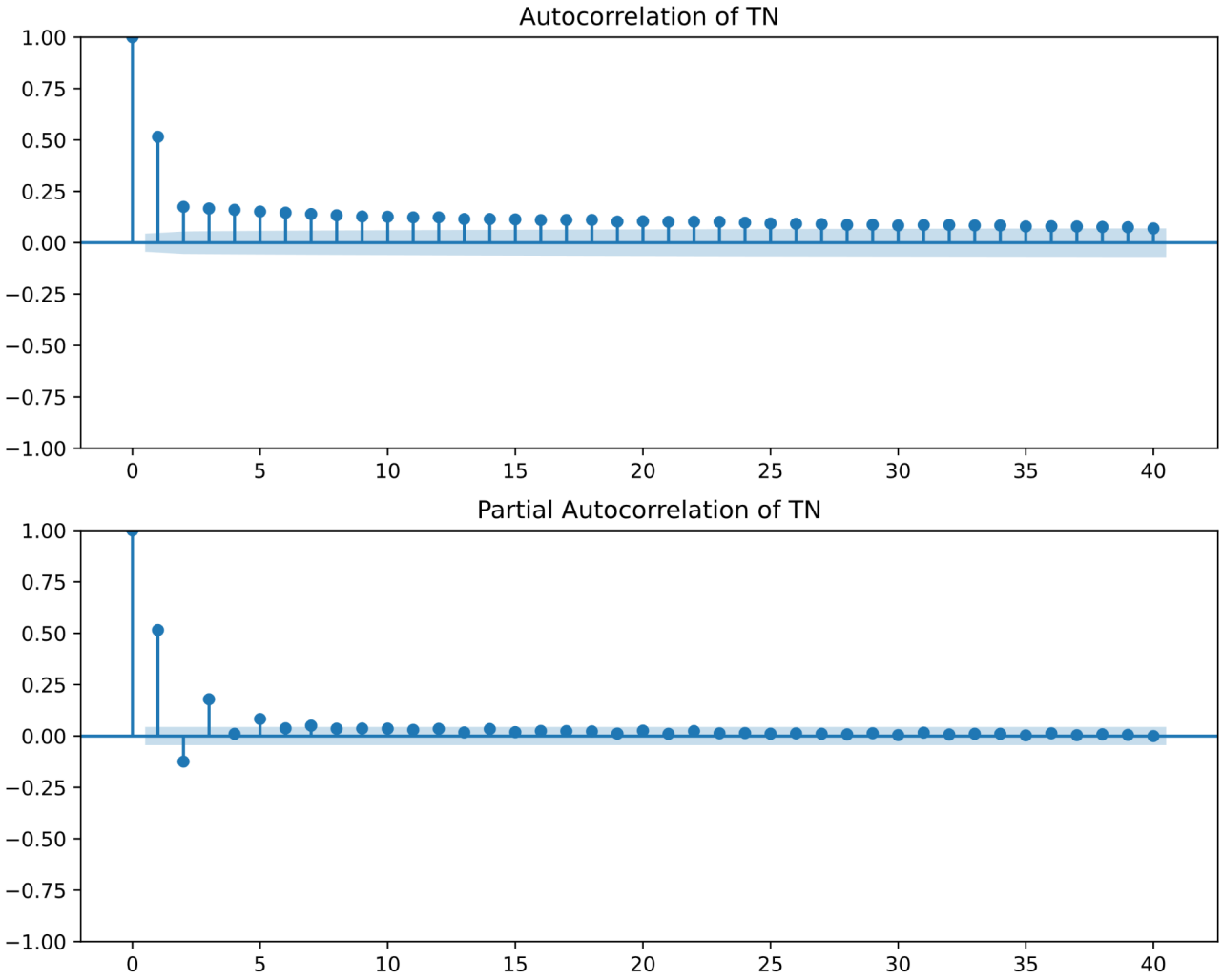

Fig. 5. ACF and PACF Plots.

attention during modeling. To ensure the accuracy of the model, it is often required to mitigate autocorrelation through techniques such as differencing.

Regarding the normality test, the results from the Jarque-Bera test show that all series exhibit a p-value of 0.0000, suggesting that none of the series follow a normal distribution. The non-normality of the distributions may impair the performance of traditional statistical models, which rely on normality assumptions. Consequently, the use of LSTM networks, which are robust to non-normal data distributions, is deemed more suitable for the subsequent analysis[29]. These results are further detailed in Table 1.

In summary, the time series of the water quality indicators exhibit varied characteristics in terms of smoothness, autocorrelation, and distribution normality, as illustrated in Fig. 5. To address the challenges posed by non-stationary and non-normal distributions, it is recommended to employ the CEEMDAN method,

combined with deep learning-based prediction models (such as LSTM), to more effectively capture the complex and dynamic features of the water quality time series.

Initial modal decomposition

The six raw data parameters were decomposed into several IMF components using the CEEMDAN algorithm, TN was decomposed into 8 components (IMF), pH, DO, and TP into 9 components, and both NH3-N and CODMn were decomposed into 10 variables. As shown in Fig. 6, the decomposition process effectively separates the original time series data into distinct frequency components, highlighting the multiscale characteristics of each parameter. For example, the IMFs for DO (Fig. 6c) reveal both high-frequency noise and low-frequency trends, while the integration results (Fig. 6a) and sample entropy values (Fig. 6b) provide insight into the complexity and predictability of the decomposed components. This decomposition provides a robust basis for subsequent analysis and modeling.

In Fig. 6(c), illustrates the decomposition process for DO as a representative example.the raw DO data and the nine IMF decomposition results are shown sequentially from top to bottom. The horizontal axis represents the number of time series and the vertical axis represents the water quality values for each component. Since the water quality dataset contains only about 1975 samples, the data volume is small for the deep learning model. Decomposition using the LSTM model directly would lead to poor prediction results because this study uses a Python sampen module to measure the sample entropy of each IMF, which is used to measure the complexity of water quality data. IMFs with similar sample entropy values can be integrated to appropriately reduce the amount of computation, increase the speed of modeling, and avoid overfitting problems. The sample entropy measurements of all IMFs and residuals for the DO variables are presented in Fig. 6(b), where IMF8 corresponds to the residuals. By looking at Fig. b, it can be noticed that the sample entropy values of the first three IMFs (IMF0, IMF1, and IMF2) are all much higher than the other IMFs, showing a complex, variable but unobtrusive pattern. In contrast, the last two IMFs and the residuals (IMF6, IMF7, and IMF8) have lower sample entropy values, obvious water quality trends, and less complexity and volatility. Therefore, it was decomposed into three parts to be studied see Fig. 6(c), high-frequency CoIMF0 (IMF0, 1, 2), CoIMF1 low frequency (IMF3,4,5), and trend CoIMF2 (IMF6,7,8). pH and TP were reconstructed in the same way as DO, and for the other three parameters, NH3-N as well as CODMn, the two parameters with IMF3-5 were reconstructed as low frequencies and IMF6-9 were reconstructed as trend terms. The same IMF0-1 for TN is reconstructed as high frequency, IMF2-4 as low frequency, and IMF5-7 as trend terms.

2.3 Secondary modal decomposition.

The initial decomposition revealed that the high-frequency component (Co-IMF0) exhibited high complexity, making it particularly challenging to predict. Therefore, VMD secondary modal decomposition was performed on this high-frequency component to reduce its complexity (see Fig. 7(a)). The parameter K was set to 5. The sample entropy measurements, illustrates the decomposition process for DO as a representative example. As shown in Fig. 7(b), indicate that the sample entropy of the high-frequency component decreased to below 0.4 after the secondary decomposition. This reduction in complexity provides the potential for improving the final prediction accuracy.

Prediction results

After performing the first two decomposition operations on the parameters, this study then discusses the two underlying frameworks based on those proposed by Feite Zhou[30], which are the integrated prediction framework and the independent prediction framework. In the integrated prediction framework, they adopt a global approach to integrate the data of all common empirical modal functions (Co-IMFs) into one LSTM model to improve the prediction performance. After CEEMDAN and sample entropy integration, the input data were converted to matrix form to capture the intrinsic correlation of the time series more comprehensively. We

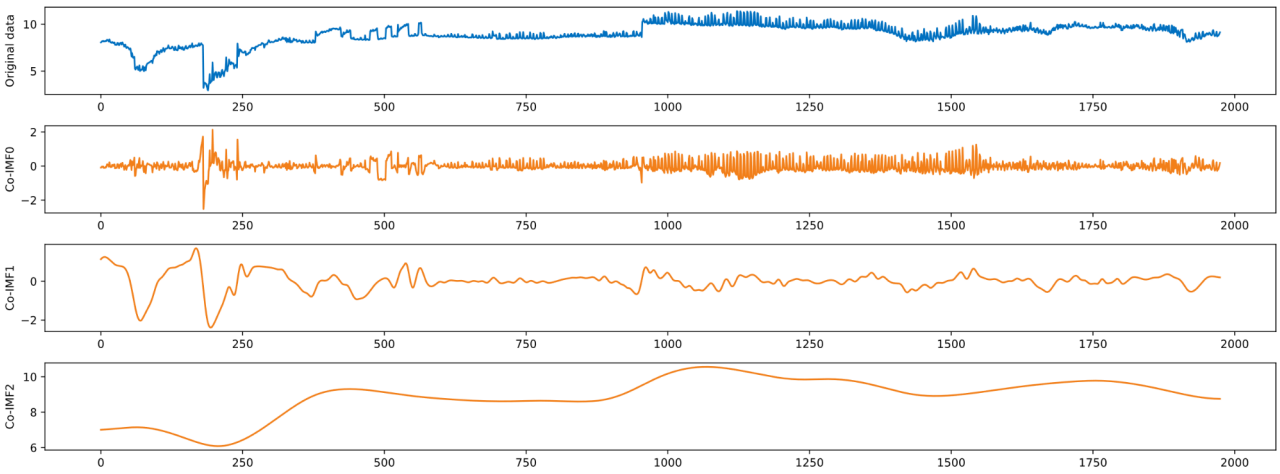

Fig. 6. Diagram of CEEMDAN Decomposition, Sample Entropy, and Integrated Results for DO Parameters.

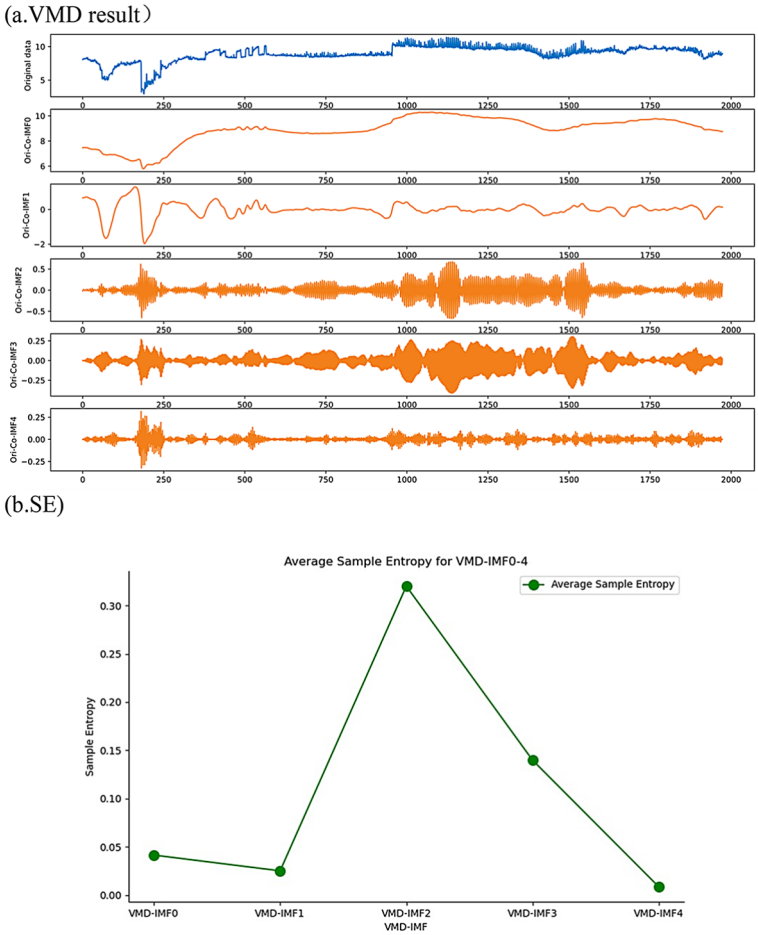

Fig. 7. Sample Entropy Diagram of VMD Decomposition for DO Parameters.

illustrates for DO as a representative example The dissolved oxygen concentration at each time point in the time series was set to correlate with the first 30 time points, and a model structure containing three LSTM layers (128, 64, and 32 cells) was constructed. Although the framework performed well in terms of prediction, there is room for improvement in terms of accuracy in coping with certain inflection points.

On the other hand, the stand-alone prediction framework employs a local prediction strategy that improves the local accuracy of the prediction by modeling each Co-IMF separately and performs more accurately in dealing with the prediction of certain inflection points. Unlike the integrated framework, this framework transforms the input data into vector form for each Co-IMF. The model structure is similar to the integrated framework, which also contains three LSTM layers (128, 64, and 32 cells) and uses the same optimizer and loss function. The final prediction results are obtained by summing the predictions of each Co-IMF to get the integrated prediction of the time series. However, the framework is more complex and takes a longer time for training and prediction.

As a result, the present study has carried out a more in-depth exploration based on the existing research. In the prediction stage of the model study, a composite prediction framework covering the VMD re-decomposition is constructed on the one hand, and on the other hand, the FECA layer is integrated into the LSTM network innovatively. By integrating the FECA layer into the LSTM network, the model can capture and utilize the frequency information in the time series data more efficiently.

In the construction of this prediction framework, the model takes systematic steps to optimize the prediction of water quality indicators. Using the CEEMDAN method, the original time series of the six water quality parameters were decomposed into IMFs and residuals of different frequencies. Through the sample entropy technique, these IMFs and residuals are integrated to form Co-IMFs.For the high-frequency Co-IMF0 which contains more noise, the VMD method is used for further decomposition and combined with the FECA layer for prediction. While the other Co-IMFs apply their respective LSTM prediction methods. Eventually, the prediction results of all Co-IMFs were summarized by the integrated LSTM method to enhance the stability and accuracy of the prediction model.

The model provides an in-depth comparative analysis of prediction models for a variety of water quality indicators including DO, pH, TP, and NH3-N. According to Table 2, the proposed method shows excellent performance in all evaluation indicators, significantly outperforming the existing benchmark models. In predicting DO, the model achieved very low RMSE (0.04) and MAE (0.03), which are much lower than the benchmark models Model1 (RMSE 0.27) and Model2 (RMSE 0.28), showing its high accuracy in DO prediction.

Predict Model		Our method	Model1	Model2	Model3	Model4	Model5
DO	RMSE	0.036	0.2691	0.2823	0.0493	0.2501	0.1838
	MAE	0.0272	0.2143	0.202	0.0345	0.2071	0.4501
	TIC	**0.0028**	0.0023	0.0223	0.0225	0.0033	0.0210
	MBE	-0.0075	-0.0598	-0.0296	**0.0020**	-0.0130	0.0507
	MAPE(%)	0.3044	2.3131	2.2644	0.428	2.299	1.7036
pH	RMSE	0.0125	0.0723	0.0402	0.0514	0.0326	0.0452
	MAE	0.011	0.0432	0.027	0.0358	0.0227	0.0295
	TIC	**0.0011**	0.0034	0.0036	0.0032	0.0023	0.0028
	MBE	-0.0092	**0.0007**	-0.0011	-0.0061	-0.0246	0.0035
	MAPE(%)	0.1398	0.5549	0.3447	0.4447	0.2898	0.3763
TP	RMSE	0.0004	0.0013	2.8996	3.3389	0.0014	0.0008
	MAE	0.0003	0.0009	1.8009	1.4205	0.001	0.0006
	TIC	**0.029**	0.1138	0.1239	0.1041	0.1195	0.0754
	MBE	-0.00002	0.00014	-0.00004	0.00012	**-0.00002**	**-0.00002**
	MAPE(%)	2.879	12.8157	16.8182	14.433	11.9103	8.1522
NH3-N	RMSE	0.0020	0.4493	0.0054	2.2802	0.0045	0.1584
	MAE	0.0017	0.0029	0.0036	0.6895	0.003	0.8182
	TIC	**0.0411**	0.1146	0.1030	0.0843	0.0894	0.0858
	MBE	0.0015	0.0009	0.0009	0.0006	0.0017	**0.0006**
	MAPE(%)	4.9628	6.3156	9.267	10.0633	8.3177	1.3439
CODmn	RMSE	0.1611	0.1491	1.5696	0.923	0.8577	0.6931
	MAE	0.127	0.0894	0.922	0.7068	0.6298	0.5706
	TIC	**0.0130**	0.0611	0.01424	0.0332	0.0153	0.0591
	MBE	0.0009	-0.0215	0.0025	0.0030	**-0.0009**	0.0098
	MAPE(%)	1.5071	11.8286	8.3619	10.36	7.5353	6.1274
TN	RMSE	0.0535	0.4296	0.0761	0.3728	0.2787	0.2626
	MAE	0.0367	0.07749	0.0542	0.2484	0.1992	0.1938
	TIC	**0.0080**	0.0544	0.0114	0.046	0.0095	0.0474
	MBE	-0.0013	-0.0008	**0.0007**	**-0.0007**	-0.00145	0.0011
	MAPE(%)	0.8338	8.1037	1.2286	5.2898	4.6032	4.4313

Table 2. Comparison of various model indicators. Significant values are in bold.

Similarly, the RMSE of the model was only 0.01 in predicting pH, reflecting its accuracy in predicting key water quality parameters.

The model also demonstrated significant results in reducing prediction errors. For example, the MAPE was only 0.30% for DO and 0.14% for pH. In terms of bias, the model also had smaller MBE values, especially in the prediction of NH3-N and TP than most of the comparison models.

Despite the advantages of this composite framework in the presentation of experimental results, a slight underperformance on certain metrics (e.g., TIC and MBE) is also found when compared with other models. Specifically, on TIC, the model showed a slight increase relative to Model1 and Model4, and on MBE, the model changed direction relative to Model3 and Model5. The analysis suggests that this may be due to differences in data preprocessing or model parameter settings. With these measures, the model is expected to show a more balanced and comprehensive performance in future releases.

To further enhance the credibility of the models, this study further employs two intuitive analytical tools, scatterplot, and boxplot, to provide an in-depth and detailed discussion on the performance of different models in predicting water quality indicators. With the introduction of the VMD re-decomposition technique and the FECA layer, the prediction framework exhibits a significant performance improvement. This enhancement was verified in the results after 500 runs, as evidenced by the best average R^2 value of up to 0.96 for the six water quality indicators (see General Fig. 8), which not only highlights the effectiveness of the VMD re-decomposition technique and the FECA layer, but also provides a solid foundation for the subsequent analysis.

The scatterplot further validates the accuracy of the model. The left side of the overall Fig. 8 shows the performance comparison of the six water quality indicators, while the right side visualizes the ability of different models to fit the water quality data through scatter plots. By scrutinizing the scatter distributions of predicted and observed values, it can be seen that the correlation between the predicted and observed values of the model is extremely high for several water quality indicators. Especially in the prediction of TN and pH, the distribution of the scatter plot almost exactly matches the trend of the observed values, which fully demonstrates the high accuracy of the model. In contrast, the scatter distributions of other benchmark models such as Model1 and Model2 are more scattered and deviate from the trend of the observed values, which further highlights the significant advantages of the proposed model in this study for water quality prediction.

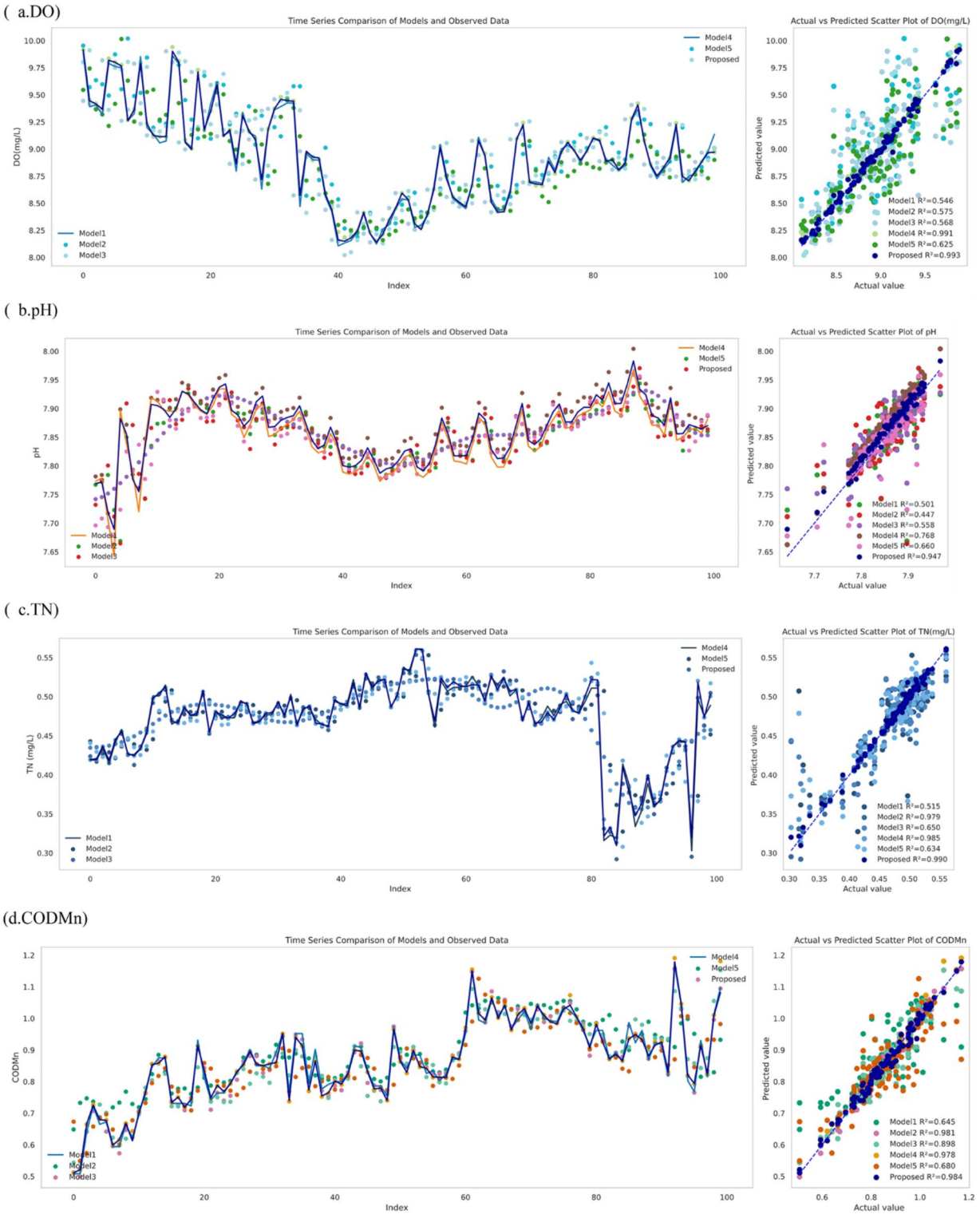

Fig. 8. Comparison Chart of Results from Different Models.

The boxplot comparison in Fig. 9 shows the prediction errors of different models for six water quality parameters, with our model performing particularly well for key parameters such as DO and NH3-N. For example, for the DO parameter, our model has a more concentrated error distribution, with an interquartile range (IQR) of approximately 0.02–0.04 mg/L, which is significantly narrower than other models, such as Model1's range of 0.15–0.25 mg/L. This indicates that our model is more accurate in predicting dissolved oxygen levels. For NH3-N, our model also shows a narrower error range, with an IQR of about 0.001–0.003 mg/L, while

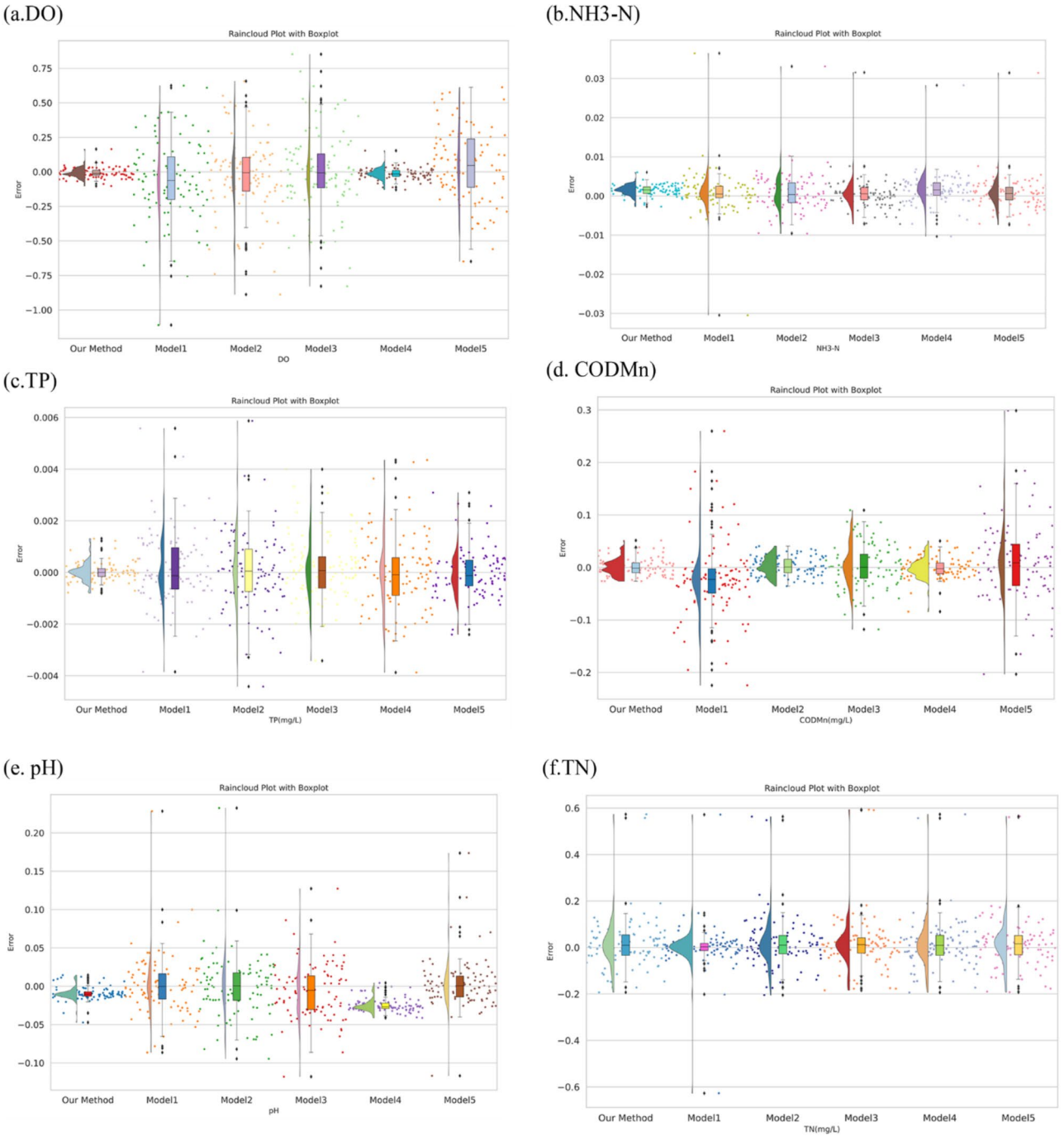

Fig. 9. Boxplot Comparison of Errors.

Model1 and Model2 show wider error ranges of 0.002–0.005 mg/L and 0.003–0.006 mg/L, respectively, further confirming the superiority of our model in handling the dynamics of NH3-N.

Discussion

The CSVLF model proposed in this study demonstrates a significant improvement in water quality prediction accuracy compared to existing methods. The model establishes a robust framework for high-frequency dynamic water quality prediction for multi-parameter systems by synergistically integrating VMD and FECA into the LSTM architecture. Recent advances in hybrid modeling based on signal decomposition, such as EMD with time series decomposition for water quality prediction[31], have highlighted the importance of signal separation in dealing with non-stationary water quality data, but traditional methods are often limited by the problems of modal aliasing and noise interference[7]. The present framework effectively solves this limitation through the

introduction of VMD, whose noise-resistant decomposition capability provides a purer data base for high-frequency component prediction.

The model performance is consistent with the previous research results, and achieves superior performance in most scenarios. For example, Baek et al.[17] showed that the hybrid model of coupled convolutional neural network (CNN) and LSTM can improve the accuracy of water level and water quality prediction, but its DO prediction accuracy only reaches 0.9283, while the CSVLF model improves the index to 0.9930. Luo et al.[16] demonstrated that the integrated empirical mode decomposition-long and short-term memory network (EEMD-LSTM) model outperforms the single LSTM model in all evaluation indicators, and although its research results are of significant value, the CSVLF model further improves the prediction accuracy of the high-frequency component by introducing the VMD and FECA mechanisms, forming a more advantageous prediction framework. The study on the prediction of chlorophyll a concentration in large lakes using the Kolmogorov-Arnold network (KAN)[32]has already emphasized the importance of hierarchical feature extraction and anti-noise decomposition in environmental modeling, while the FECA mechanism innovatively introduced in this model can dynamically optimize the weights of the spectral features, resulting in a significant improvement of the prediction accuracy. This design is in line with the trend of hybrid architecture in environmental science, where the integration of FECA further enhances the adaptability of the model.

This dual-mechanism synergy - the VMD realizes the anti-noise decomposition, and the FECA realizes the frequency-domain feature prioritization - enables the model to perform particularly well in predicting critical water quality parameters. Results show that the RMSE of DO prediction is reduced from 0.269 mg/L to 0.036 mg/L, and the MAPE of TN is improved by a factor of 10 (from 8.10 to 0.83%, compared to model 1). Such performance improvements are consistent with the results of studies that emphasize hybrid architectures that reduce computational complexity while maintaining accuracy[33], but exceed existing benchmarks through targeted optimization of high-frequency components. The dynamic feature selection mechanism of FECA further extends the applicability of the model to complex urban waters, compared to results from river studies where adaptive decomposition techniques have led to improved TN prediction accuracy.

The high accuracy and strong robustness of the CSVLF model in predicting water quality parameters are of great practical significance for ecological environmental monitoring and management. For example, the model's high-precision real-time prediction capability can support the deployment of a dynamic water quality monitoring system that can realize the immediate detection and rapid response to pollution events; the accurate prediction of dissolved oxygen and TN concentrations provides key data support for maintaining the health of aquatic ecosystems; and the model's ability to effectively capture the long-term dependence and dynamic changes in water quality can provide a scientific basis for decision-making on ecological restoration and protection. Despite the excellent performance of the CSVLF model in handling routine water quality parameters, its ability to predict extreme values (e.g., sudden ammonia concentration spikes) is still limited. Therefore, future work can integrate spatially explicit covariates (e.g., industrial emission patterns) or adopt an uncertainty quantification framework to improve robustness and further extend the model's generalization ability in extreme scenarios.

Conclusion

In conclusion, the CSVLF model in this study shows obvious advantages in water quality monitoring. It can effectively extract different frequency components in the water body, which makes the model more adaptable, and the LSTM model can capture the long-term dependence of the water quality time series in modeling, thus predicting the dynamic changes of water quality more accurately. There are some limitations of the model. The model performance may be limited by the data quality and sampling frequency, there is still room for improvement in the handling of certain anomalies, and the construction and tuning of the model may be limited for the adaptability of certain water bodies, which needs to be further verified and optimized in practical applications.

Data availability

The dataset utilized in this study originates from the National Real-Time Surface Water Quality Inspection and Distribution System (NRWQIS). Relevant data are accessible via the designated platform at https://szzdjc.cnemc.cn:8070/GJZ/Business/Publish/, subject to the platform's guidelines and data request protocols. For further assistance or data inquiries, please contact Jie Long at 1444236943@qq.com.

Received: 18 December 2024; Accepted: 1 April 2025
Published online: 14 April 2025

References

1. United Nations. *The United Nations World Water Development Report 2023: Partnerships and Cooperation for Water* (UNESCO, 2023).
2. Jeong, H. et al. Machine learning-based water quality prediction using octennial in-situ daphnia magna biological early warning system data. *J. Hazard. Mater.* **465** (2024).
3. Mondal, I. et al. Assessing intra and interannual variability of water quality in the Sundarban mangrove-dominated estuarine ecosystem using remote sensing and hybrid machine learning models. *J. Clean. Prod.* **442**, 140889 (2024).
4. Haggerty, R., Sun, J., Yu, H. & Li, Y. Application of machine learning in groundwater quality modeling - A comprehensive review. *Water Res.* **233**, 119745 (2023).
5. Pyo, J. et al. Long short-term memory models of water quality in inland water environments. *Water Res. X.* **21**, 100207 (2023).
6. Ye, Q., Yang, X., Chen, C. & Wang, J. IEEE,. River Water Quality Parameters Prediction Method Based on LSTM-RNN Model. In 2019 Chinese Control And Decision Conference (CCDC), 3024–3028 (2019).
7. Sha, J., Li, X., Zhang, M. & Wang, Z-L. Comparison of forecasting models for Real-Time monitoring of water quality parameters based on hybrid deep learning neural networks. *Water* **13**, 1547 (2021).

8. Singha, S., Pasupuleti, S., Singha, S. S., Singh, R. & Kumar, S. Prediction of groundwater quality using efficient machine learning technique. *Chemosphere* **276**, 130265 (2021).
9. Bi, J., Chen, Z., Yuan, H. & Zhang, J. Accurate water quality prediction with attention-based bidirectional LSTM and encoder–decoder. *Expert Syst. Appl.* **238**, 121807 (2024).
10. Wang, K. et al. Hybrid deep learning based prediction for water quality of plain watershed. *Environmental Research* 262, Part 2 (2024).
11. Zhi, W. et al. Deep learning for water quality. *Nat. Water*. **2**, 228–241 (2024).
12. Pyo, J. et al. Using convolutional neural network for predicting cyanobacteria concentrations in river water. *Water Res.* **186**, 116349 (2020).
13. Bui, D. T., Khosravi, K., Tiefenbacher, J., Nguyen, H. & Kazakis, N. Improving prediction of water quality indices using novel hybrid machine-learning algorithms. *Sci. Total Environ.* **721**, 137612 (2020).
14. Bahmani, R., Ouarda T B M & J Groundwater level modeling with hybrid artificial intelligence techniques. *J. Hydrol.* **595**, 125659 (2021).
15. Tan, W. et al. Application of CNN and long short-term memory network in water quality predicting. *Intell. Autom. Soft Comput.* **34**, 1943–1958 (2022).
16. 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 (2023).
17. Baek, S. S., Pyo, J. & Chun, J. A. Prediction of water level and water quality using a CNN-LSTM combined deep learning approach. *Water* **12**, 3399 (2020).
18. Yao, S. et al. Long-Term water quality prediction using integrated water quality indices and advanced deep learning models: A case study of Chaohu lake, China, 2019–2022. *Appl. Sci.* **12**, 11329 (2022).
19. Bennia, F. et al. Comparative study between EMD, EEMD, and CEEMDAN based on De-Noising Bioelectric Signals. In: Proceedings of the 8th International Conference on Image and Signal Processing and their Applications (ISPA), Biskra, Algeria, ; pp. 1–6. (2024).
20. Hussein, E. E. et al. Groundwater quality assessment and irrigation water quality index prediction using machine learning algorithms. *Water* **16**, 2 (2024).
21. Bachir, S. et al. Water quality index modeling using random forest and improved SMO algorithm for support vector machine in Saf-Saf river basin. *Environ. Sci. Pollut. Res. Int.* **29**, 48491–48508 (2022).
22. Ma, Y. et al. Prediction of the dissolved oxygen content in aquaculture based on the CNN-GRU hybrid neural network. *Water* **16**, 24 (2024).
23. Jie, Y. W. et al. Application of artificial intelligence methods for monsoonal river classification in Selangor river basin, Malaysia. *Environ. Monit. Assess.* **193**, 7 (2021).
24. Molina-Picó, A. et al. Comparative study of approximate entropy and sample entropy robustness to spikes. *Artif. Intell. Med.* **53**, 2 (2011).
25. Jiang, M. et al. Frequency enhanced channel attention mechanism for time series forecasting. *Adv. Eng. Inform.* **58** (2023).
26. Zhang, Y. et al. Accurate prediction of water quality in urban drainage network with integrated EMD-LSTM model. *J. Clean. Prod.* **354**, 131724 (2022).
27. Zhang, X. et al. Comparative study of rainfall prediction based on different decomposition methods of VMD. *Sci. Rep.* **13**, 20127 (2023).
28. Zhang, K. et al. Enhancing water quality prediction with advanced machine learning techniques: an extreme gradient boosting model based on long short-term memory and autoencoder. *J. Hydrol.* **644**. (2024).
29. Che, Z., Peng, C. & Yue, C. Optimizing LSTM with multi-strategy improved WOA for robust prediction of high-speed machine tests data. *Chaos Solitons Fractals*. **178**, 114394 (2024).
30. Zhou, F., Huang, Z. & Zhang, C. Carbon price forecasting based on CEEMDAN and LSTM. *Appl. Energy*. **311**, 118601 (2022).
31. Murcia, E. & Guzmán, S. M. Using singular spectrum analysis and empirical mode decomposition to enhance the accuracy of a machine Learning-Based soil moisture forecasting algorithm. *Comput. Electron. Agric.* **224**, 109200 (2024).
32. Saravani, M. J. et al. Predicting Chlorophyll-a concentrations in the world's largest lakes using Kolmogorov-Arnold networks. *Environ. Sci. Technol.* **59**, 1801–1810 (2025).
33. Zhang, K. et al. Enhancing water quality prediction with advanced machine learning techniques: an extreme gradient boosting model based on long Short-Term memory and autoencoder. *J. Hydrol.* **644**, 132115 (2024).

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No.62166039).

Author contributions

Final Manuscript Revision, funding, Supervision: Chong Lu, Jie Long; Study conception and design, analysis, and interpretation of results, methodology development: Jie Long.; Draft manuscript preparation, figures and tables: Yiming Lei, Zhong Yuan Chen, Yihan Wang. All authors reviewed the results and approved the final version of the manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

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

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-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, 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 you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. 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-nc-nd/4.0/.

© The Author(s) 2025