Mathematical Biosciences and Engineering

http://www.aimspress.com/journal/MBE

MBE, 18(6): 7561–7579.
DOI: 10.3934/mbe.2021374
Received: 17 June 2021
Accepted: 30 August 2021
Published: 02 September 2021

Research article

A prediction model of aquaculture water quality based on multiscale decomposition

Huanhai Yang[1,2] **and Shue Liu**[3,*]

[1] School of Computer Science and Technology, Shandong Technology and Business University, Yantai, China

[2] Co-innovation Center of Shandong Colleges and Universities: Future Intelligent Computing, Shandong Technology and Business University, Yantai, China

[3] Binzhou Medical University, Yantai, China

* **Correspondence:** Email: lse32@bzmc.edu.cn.

Abstract: In the field of intensive aquaculture, the deterioration of water quality is one of the main factors restricting the normal growth of aquatic products. Predicting water quality in real time constitutes the theoretical basis for the evaluation, planning and intelligent regulation of the aquaculture environment. Based on the design principles of decomposition, recombination and integration, this paper constructs a multiscale aquaculture water quality prediction model. First, the complete ensemble empirical mode decomposition with adaptive noise (CEEMDAN) method is used to decompose the different water quality variables at different time scales step by step to generate a series of intrinsic mode function (IMF) components with the same characteristic scale. Then, the sample entropy of each IMF component is calculated, the components with similar sample entropies are combined, and the original data are recombined into several subsequences through the above operations. In this paper, a prediction model based on a long short-term memory (LSTM) neural network is constructed to predict each recombination subsequence, and the Adam optimization algorithm is used to continuously update the weight of neural network to train and optimize the prediction performance. Finally, the predicted value of each subsequence is superimposed to predict the original water quality data. The dissolved oxygen and pH data of an aquaculture base were collected for prediction experiments, the results of which show that the proposed model has a high prediction accuracy and strong generalization performance.

Keywords: complete ensemble empirical mode decomposition with adaptive noise; long short-term memory; sample entropy; water quality prediction; aquaculture

1. Introduction

With the development of the Internet of Things, big data and artificial intelligence technology, aquaculture is increasingly becoming more intensive, precise and intelligent. In the field of high-density intensive aquaculture, predicting the development trend of water quality (that is, predicting the trends of variables such as dissolved oxygen, pH, temperature, and turbidity) in real time is of great significance for preventing the water quality from deteriorating and for avoiding the outbreak of disease.

Existing water quality prediction methods mainly include traditional statistical methods such as regression analysis and time series methods and intelligent calculation methods such as neural networks and support vector machines (SVMs). For instance, Rajaee and Jafari [1] proposed integrating the discrete wavelet transform into artificial neural networks, gene expression planning, and decision trees for the prediction of water quality indicators. Amir Hamzeh Haghiabi et al. [2] studied the application of artificial neural networks (ANNs), the group method of data handling (GMDH) and SVMs to the prediction of water quality. Rahman et al. [3] developed a set of step predictors, each of which predicts a specific timestamp, thereby providing new insights for the long-term prediction of dissolved oxygen. Barzegar et al. [4] studied the wavelet and extreme learning machine (WA-ELM) hybrid model for multi-step-ahead prediction and adopted the boosting integration method. Jafari et al. [5] proposed a water quality prediction model based on hybrid wavelet genetic programming method and Shannon entropy. Rozario and Devarajan [6] employed the fuzzy C-means clustering method and constructed a radial basis function (RBF) neural network to predict the change trend of dissolved oxygen. Kisi et al. [7] proposed Bayesian model averaging (BMA) to estimate the hourly dissolved oxygen. Li et al. [8] established three dissolved oxygen prediction models using a recurrent neural network (RNN) model, a long short-term memory (LSTM) model, and a gated recurrent unit (GRU) model. Dabrowski et al. [9] studied a method to forecast the quality of prawn pond water that introduces mean reversion into multi-step-ahead forecasts of state-space models. Chen et al. [10] established a hybrid three-dimensional dissolved oxygen content prediction model based on an RBF neural network with K-means and subtractive clustering.

An RNN introduces the concept of time series into the network structure, making it more adaptable in time series data prediction and analysis tasks. In contrast, an LSTM neural network [11] solves the gradient disappearance problem and avoids the gradient explosion issue in RNN models. Moreover, LSTM neural networks have a time loop structure that can effectively describe sequence data with temporal and spatial correlations and can solve the problem of long-distance dependence [12]. LSTM adjust the structure of the network on the basis of the simple recurrent neural network, adding a gating mechanism to control the transmission of information in the neural network. As a variant of LSTM, Gated Recurrent Unit (GRU) has made certain changes in the gating mechanism, and also mixed the cell state and hidden state [13, 14]. GRU directly passes the hidden state to the next unit, while LSTM uses memory cell to wrap the hidden state. The performance of GRU and LSTM is similar in many tasks. The GRU structure is simpler and has fewer parameters, so it is easier to converge. In recent years, LSTM and GRU have been considered as one of the effective methods to deal with time series forecasting problems.

Michieletto et al. [15] studied the application of LSTM and phased LSTM (PLSTM) networks to the prediction of dissolved oxygen. Li et al. [16] proposed a water quality prediction model combining a sparse autoencoder with an LSTM network. Barzegar et al. [17] studied the application of

a convolutional neural network (CNN)-LSTM hybrid deep learning model for short-term water quality prediction. Zhou et al. [18] proposed a water quality prediction method based on the improved gray relational analysis (IGRA) algorithm and an LSTM neural network. Zou et al. [19] proposed a water quality prediction method based on a bidirectional LSTM network with multiple time scales.

Aquaculture water quality data are nonlinear and unstable. Hence, if the original data are directly used for prediction, considerable problems such as the impact of noise and a low prediction accuracy will arise. Empirical mode decomposition (EMD) [20], a fully adaptive nonlinear signal processing algorithm, can resolve the nonstationarity of the input data and improve the model prediction accuracy. Accordingly, Fijani et al. [21] proposed a hybrid water quality prediction model that combines complete ensemble empirical mode decomposition with adaptive noise (CEEMDAN) and variational mode decomposition (VMD) algorithms with an extreme learning machine (ELM) and a least-squares SVM (LSSVM). Likewise, Huan et al. [22] studied a hybrid model involving integrated EEMD and an LSSVM for the prediction of dissolved oxygen. Similarly, Eze and Ajmal [23] proposed a combined dissolved oxygen prediction method based on integrated EEMD and an LSTM neural network. Liu et al. [24] constructed a multiscale water temperature prediction model based on EMD and a back-propagation neural network.

For data sets with dynamic and nonlinear characteristics, only relying on information decomposition technology limits the accuracy and efficiency of prediction, and sequence reorganization using sample entropy can decrease the workload and operate more handily. Wei Sun et al. [25] Proposed a hybrid wind speed forecasting model, including fast ensemble empirical mode decomposition, sample entropy, phase space reconstruction and back-propagation neural network with two hidden layers. Jujie Wang et al. [26] proposed a hybrid model composed of complete ensemble empirical mode decomposition (ceemdan), sample entropy (SE), long-term and short-term memory (LSTM) and random forest (RF) to realize the accurate prediction of coal price. Qunli Wu et al. [27] proposed a hybrid air quality index forecasting model using variational mode decomposition (VMD), sample entropy (SE) and long short-term memory (LSTM) neural network. More and more researchers combine sample entropy with decomposition technology. Sample entropy can be used to analyze the complexity of decomposition sequence, reorganize sequence to reduce computational complexity, or determine the number of decomposition layers.

The dissolved oxygen content and pH value are important factors that affect the quality of aquaculture water. When the dissolved oxygen content in aquaculture water falls below 4 mg/L, the intake of food by fish begins to decrease, and dissolved oxygen contents higher than 14.4 mg/L can cause gas bubble disease. In addition, the pH range of aquaculture water suitable for fish is from 7.5 to 8.5; thus, pH values less than 4 or greater than 10 can result in the death of a large number of fish. Dissolved oxygen and pH are important indicators affecting the survival of aquatic organisms. By accurately predicting its development trend, breeders can find abnormal water quality in advance. So as to avoid the death and disease of aquatic organisms and ensure the high-quality development of aquaculture.

By combining the principles of decomposition and reconstruction with deep learning, this paper constructs a water quality prediction model named CEEMDAN-SE-LSTM and conducts research on the prediction of dissolved oxygen and pH to forecast aquaculture water quality. The proposed model first applies CEEMDAN to decompose the dissolved oxygen and pH sequences at multiple scales, thereby obtaining a series of intrinsic mode functions (IMFs) with different characteristic scales and a remainder. Then, the IMF components with similar sample entropy (SE) are recombined to reduce the

input complexity. Finally, the reconstructed sequences are applied to a trained LSTM neural network for single-step prediction, the values of which are integrated to obtain the final prediction result.

The contributions of this paper are listed as follows:

(1) This paper uses CEEMDAN to decompose dissolved oxygen and pH data into subsequences with different time scales. This process can fully determine the characteristics and trends of the water quality series and transform complex single-scale characteristics into simple multiscale characteristics for the ease of prediction.

(2) The SE of each IMF sequence is calculated, merged and recombined into sequences with similar entropy values. Then, the LSTM prediction model is trained for each sequence after the reconstruction, the model structure is optimized for single-step prediction, and finally, the prediction results are integrated.

(3) In this paper, the autocorrelation coefficient is used to measure the degree of correlation between different time points in the water quality series. The autocorrelation coefficient is used as the time step parameter of the model prediction, thus avoiding the redundancy or insufficiency of input information and increasing the efficiency of the constructed prediction model.

The remainder of this paper is structured as follows: the CEEMDAN algorithm, SE and LSTM neural network are described in Section 2. The experimental process of the CEEMDAN-SE-LSTM model and a comparative analysis with other models are discussed in Section 3. The paper is summarized and the directions of future research on water quality prediction are discussed in Section 4.

2. Materials and method

2.1. Complete ensemble empirical mode decomposition with adaptive noise

EMD is based on the variation in the data and can be applied directly without preliminary analyses or research. However, studies have shown that EMD has a limitation regarding mode mixing [28, 29]. To solve the mode mixing problem, an ensemble version of EMD called EEMD [30] was developed, which added white noise on the basis of EMD decomposition, so that the decomposed IMF is a single mode. Although EEMD greatly reduced the possibility of mode mixing, there raised a new problem: a residue noise will be mixed into the original signal after reconstruction. Therefore, the Complementary EEMD (CEEMD) [31] is proposed, which added the white noise to the original data in pairs, which greatly alleviates the residual problem of noise after reconstruction . CEEMD still has some problems, such as incompleteness and large amount of calculation [32]. In recent years, CEEMDAN [33] has been proposed. The CEEMDAN adds a limited amount of adaptive white noise at each stage, which can effectively suppress residuals, increase the reconstruction accuracy, and reduce the number of iterations; moreover, CEEMDAN is more suitable for nonlinear signal analysis than other existing methods[34].

The steps for decomposing the original water quality time series in CEEMDAN are as follows:

(1) Add white noise following a normal distribution. The resulting water quality time series of the i-th experiment is shown in Eq (2.1).

$$X_i(t) = X(t) + \varepsilon_0 v_i(t) \tag{2.1}$$

where $v_i(t)$ is the noise sequence added in the i-th experiment and ε_0 is the noise amplitude.

(2) Perform n-time EMD on the noise-added signal, and obtain the first IMF component through the mean value calculation, as in Eq (2.2).

$$imf_1 t = \frac{1}{n}(\sum_{i=1}^{n} \overline{imf_1^i t})) \tag{2.2}$$

(3) Obtain the remainder from the original data and the first IMF component, as in Eq (2.3).

$$r_1(t) = X(t) - imf_1 t \tag{2.3}$$

(4) Add white noise to the remainder, and continue to implement decomposition to obtain the second IMF component, as in Eq (2.4).

$$imf_2(t) = \frac{1}{n}\sum_{i=1}^{n} E_1(r_1(t) + \varepsilon_1 E_1(v_i(t))) \tag{2.4}$$

where $E_k(\bullet)$ is the kth IMF component produced by the EMD method.

(5) According to the above steps, continue to perform multiple decompositions, and calculate Both the remainder after the kth decomposition and the k+1th IMF component, as in Eqs (2.5) and (2.6).

$$r_k(t) = r_{k-1}(t) - imf_k(t) \tag{2.5}$$

$$imf_{k+1}(t) = \frac{1}{n}\sum_{i=1}^{n} E_1(r_k(t) + \varepsilon_k E_k(v_i(t))) \tag{2.6}$$

(6) Repeat step 5 until the extremum points of the margin do not exceed two; the satisfaction of this condition terminates the decomposition. Assuming that m IMFs are obtained, the final remainder R(t) is described in Eq (2.7).

$$R(t) = X(t) - \sum_{i=1}^{m} imf_i(t) \tag{2.7}$$

After the above steps, the original data are finally decomposed into several IMF components and a remainder. A better decomposition effect can be obtained by adjusting various parameters, such as the noise standard deviation (Nstd), number of realizations (NR), and maximum number of iterations (MaxIter).

2.2. Sample entropy

SE [35] can be used to quantify the regularity of time series fluctuations. If the SE difference between two time series is small, the two series are highly similar.

The SE algorithm is expressed as follows:

(1) For an aquaculture water quality data sequence obtained by sampling at equal time intervals $\{X_i\ (i = 1, 2 \cdots n)\}$, using m as the time window length, divide the original sequence into n-m+1 subsequences, as in Eq (2.8).

$$X_i(t) = \{x_i(t), x_{i+1}(t), \cdots x_{i+m-1}(t)\} \tag{2.8}$$

(2) Define the distance $d[X_m(i), X_m(i+1)]$ between the vectors $X_m(i)$ and $X_m(i+1)$ as the absolute value of the maximum difference between the two corresponding elements, and calculate the distance between each sequence, as in Eq (2.9).

$$d[X_m(i), X_m(i+1)] = \max |x_{i+k}(t) - x_{i+1+k}(t)| \quad (k = 0, 1 \cdots m-1) \tag{2.9}$$

(3) Define the threshold $F = r * std$, where std is the standard deviation of the original sequence and r takes a value between 0.1 and 0.25 according to the application scenario. Count the ratio of the number of distances greater than F to all sample values that do not include itself, denote the ratio as $C_i^m(t)$, and calculate the average value $\Phi^m(t)$ following Eq (2.10).

$$\Phi^m(t) = \frac{1}{n-m} \sum_{i=1}^{m} C_i^m(t) \tag{2.10}$$

(4) Taking the length of the time window as m+1, repeat the above steps to obtain the SE of each subsequence, as in Eq (2.11):

$$SampEn(t) = \ln(\Phi^m(t) - \Phi^{m+1}(t)) \tag{2.11}$$

The SE calculation does not depend on the amount of data; moreover, the calculation speed is fast, and the anti-interference ability is strong [36]. Additionally, the SE is very sensitive to time series changes and thus has been widely used to measure the complexity of various time series.

2.3. Long short-term memory neural network

The cell state of an LSTM network is composed of two activation functions, which are composed of three gating units: a forget gate, an input gate and an output gate [37]. Each gate in an LSTM model has a unique function. The forget gate controls whether the previous cell state is forgotten with a certain probability, while the input gate and output gate control the direction of data flow [38].

The structure of a single neuron in an LSTM network is illustrated in Figure 1, in which X_t and h_t denote the input and output of the neuron at time t, respectively, and C_t is the neuron cell state at time t.

In an LSTM model, the forget gate uses the sigmoid activation function to determine what information can pass through the cell state. The output gate generates a value from 0 to 1 based on the output h_{t-1} at the previous moment and the current input X_t to determine whether to completely or partially pass the information C_{t-1} learned at the previous moment. The output formula of the forget gate is shown in Eq (2.12).

$$f_t = \sigma(W_{xf}X_t + W_{hf}h_{t-1} + W_{cf}C_{t-1} + b_f) \tag{2.12}$$

where W_{xf},W_{hf} and W_{cf} are the relevant connection weights, b_f is the bias matrix, and σ is the sigmoid activation function, the mathematical formula of which is described in Eq (2.13).

$$\sigma(x) = \frac{1}{1+e^{-x}} \tag{2.13}$$

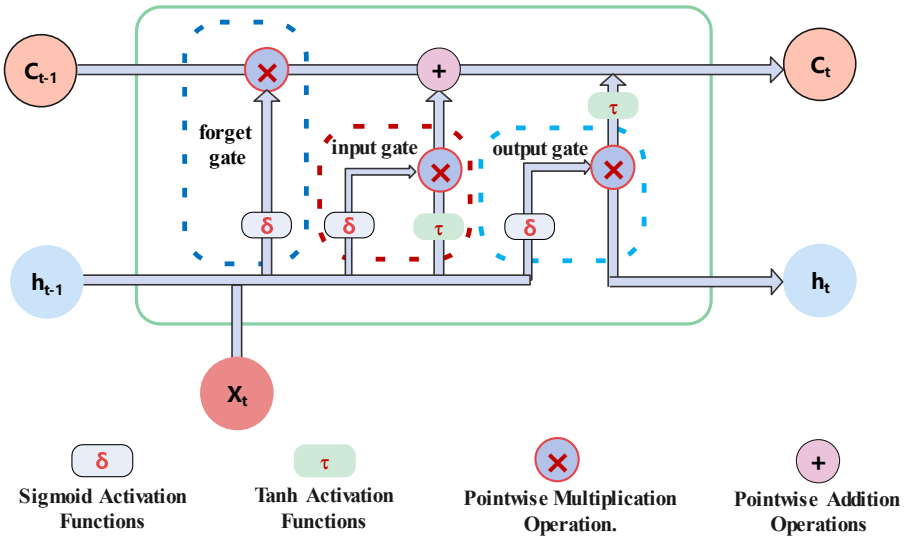

Figure 1. Diagram of the neuron structure in an LSTM network.

The input gate determines which new information needs to be received and consists of two parts. The first part uses the sigmoid activation function to determine which values to update, and the second part applies the tanh activation function to generate a new candidate value $\tilde{C}_t$, as in Eqs (2.14) and (2.15).

$$i_t = \sigma(W_{ii}X_t + W_{hi}h_{t-1} + b_i) \tag{2.14}$$

$$\tilde{C}_t = \tanh(b_c + W_{ci}X_t + W_{ch}h_{t-1}) \tag{2.15}$$

In the above formulas, $W_{ii,}W_{ci}$ and W_{hi}, W_{ch} are the corresponding weights, and b_i and b_c are bias matrices.

The tanh function is a hyperbolic tangent function whose output range is between -1 and 1, and its mathematical formula is shown in Eq (2.16).

$$\tanh(x) = \frac{(e^x - e^{-x})}{(e^x + e^{-x})} \tag{2.16}$$

The cell state C_t exists throughout the entire LSTM chain system and is updated through the input and forget gates, as in Eq (2.17).

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \tag{2.17}$$

In the above formula, the value of C_t is determined by the cell state of the previous neuron C_{t-1} and by the input gate i_t and output gate f_t.

The output gate determines the output of the model. First, an initial output is obtained through the sigmoid activation function, and then the value of C_t is scaled to between -1 and 1 using the tanh activation function. Finally, the output obtained by C_t and the sigmoid activation function is multiplied pairwise to obtain the output of the model, as in Eqs (2.18) and (2.19).

$$O_t = \sigma(W_{ox}X_t + W_{oh}h_{t-1} + b_o) \tag{2.18}$$

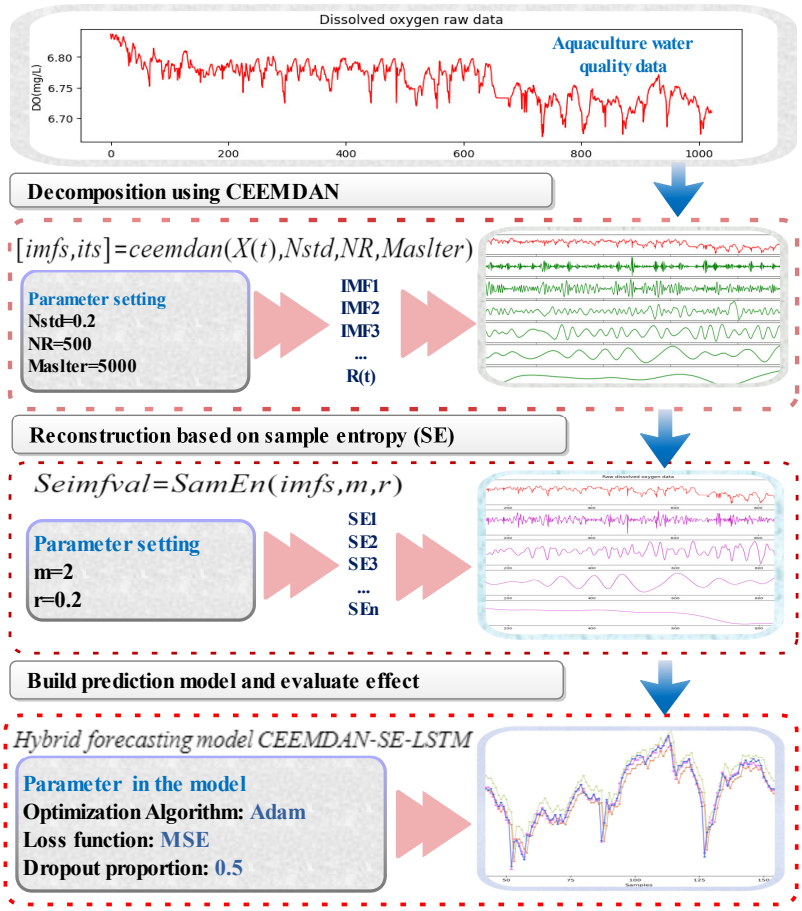

Figure 2. Flow chart of the aquaculture water quality data prediction model.

$$h_t = O_t * \tanh(C_t) \tag{2.19}$$

where W_{ox} and W_{oh} are the relevant connection weights and b_o is a bias matrix. LSTM can selectively retain or forget information when this information flows in each neuron through the gate structure. This structure can effectively solve the problem of long-distance dependence and is suitable for the prediction of aquaculture water quality time series data.

2.4. CEEMDAN-SE-LSTM hybrid prediction model

Aquaculture water quality data (such as dissolved oxygen and pH) are nonlinear and nonstationary and are easily affected by many factors, such as the water temperature, weather, and aquaculture density. This paper proposes a hybrid prediction model named CEEMDAN-SE-LSTM. The model first uses CEEMDAN to decompose the water quality sequence data and then uses the SE to reconstruct similar sequences. Finally, an LSTM network is used for the single-step prediction of each sequence before integrating to obtain the final prediction result. A flow chart of the prediction model is presented in Figure 2.

The CEEMDAN-SE-LSTM prediction model proposed in this paper mainly consists of four parts.

(1) Decomposition of water quality data: The CEEMDAN method is used to decompose the original water quality series into IMF function components with different frequencies, so as to reduce the

influence of the non-stationarity of the original series on the prediction accuracy.

(2) Combination based on SE: Calculate the sample entropy value of each IMF separately, and recombine the IMF with the approximate sample entropy value into a new sequence, which can effectively reduce the amount of calculation and avoid inaccurate information extraction caused by over-decomposition.

(3) Prediction of each sub-sequence: According to the data characteristics of each sub-sequence, the hyperparameters of the LSTM neural network are optimized for individual prediction.

(4) Integration: The prediction results of each recombination sequence are added to obtain the prediction results of the final water quality data.

3. Results

3.1. Sources of water quality data

This paper selected the Shandong Yantai aquaculture base as the experimental area. This aquaculture base is equipped with modern fishery equipment such as dissolved oxygen sensors, pH sensors, aeration pumps, and wireless monitoring systems. Dissolved oxygen data, which fluctuate considerably, are collected every 10 minutes. During the 9 days from August 25 to September 2, 2019, after data preprocessing, a total of 1024 valid data points were retained. The pH of the aquaculture water was relatively stable and was measured once an hour, yielding a total of 634 valid data points. Eighty percent of the data are selected to train the prediction model, and the remaining twenty percent of the data are used for testing.

3.2. Multiscale decomposition of water quality data

The CEEMDAN algorithm is used to decompose the dissolved oxygen and pH data at the marine aquaculture base in Laishan, Yantai, Shandong, and to identify and separate several IMF components and one residual component step by step. The results are shown in Figure 3.

Through the CEEMDAN algorithm, the original dissolved oxygen sequence is decomposed into seven IMF components with different characteristics and a residual signal. Likewise, the original pH sequence is divided into six IMF components and a residual signal. The results demonstrate that the features at different scales in the original data sequences are decomposed well.

3.3. Reconstruction of the IMF components based on the sample entropy

Considering the large number of IMF components obtained by CEEMDAN, direct prediction will increase the computational cost. Therefore, this paper uses the SE to evaluate the complexity of each IMF component and then reconstruct the decomposed components based on the differences in the SE among the components.

An experimental verification suggests that, when calculating the SE of the data samples in this paper, the time window length parameter m = 2 and the threshold parameter F = 0.2 * $std(IMF(i))$ can best reflect the different complexity of each component. The SE of each IMF component decomposed from the above dissolved oxygen and pH data sequences is plotted in Figure 4. Adjacent IMF components whose SE difference is less than 0.1 are similar (that is, their complexity and regularity are similar), and thus, these components can be recombined into a single new component. This recombination of IMF

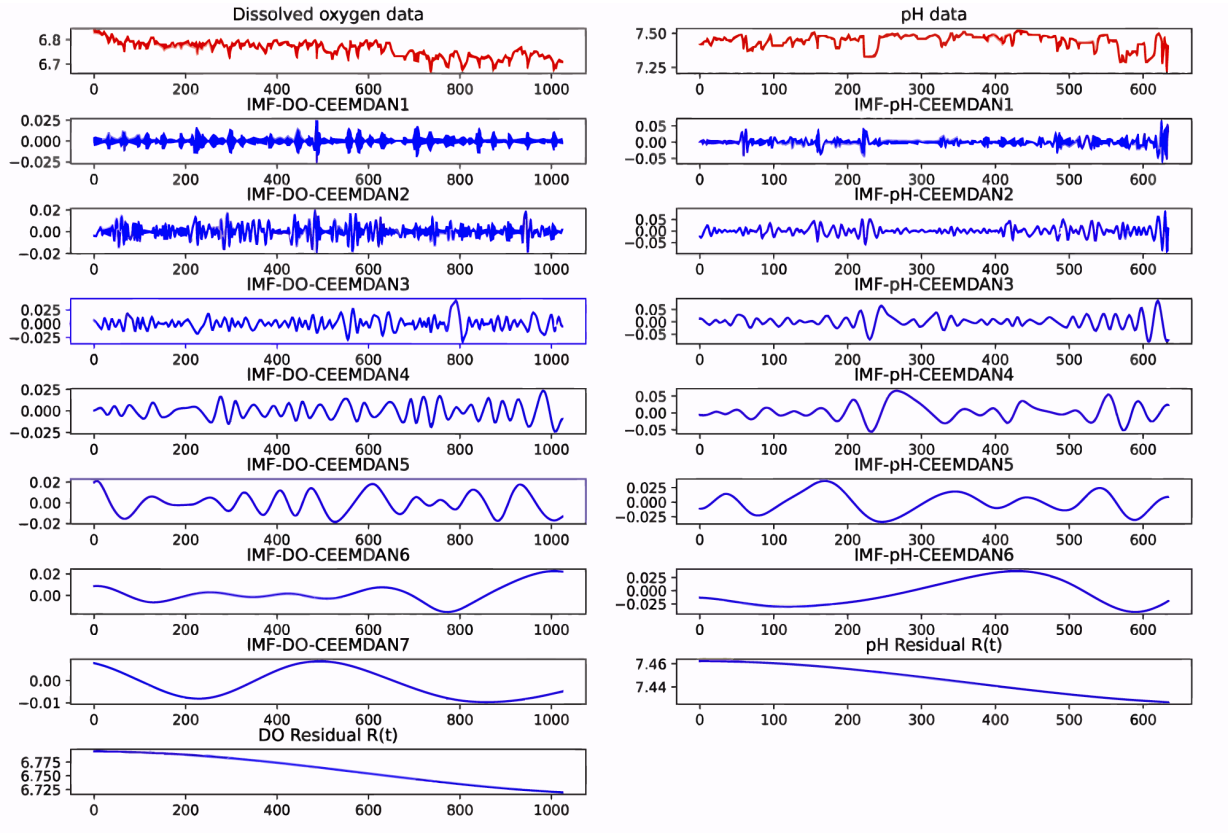

Figure 3. Multiscale decomposition of dissolved oxygen (DO) and pH using CEEMDAN.

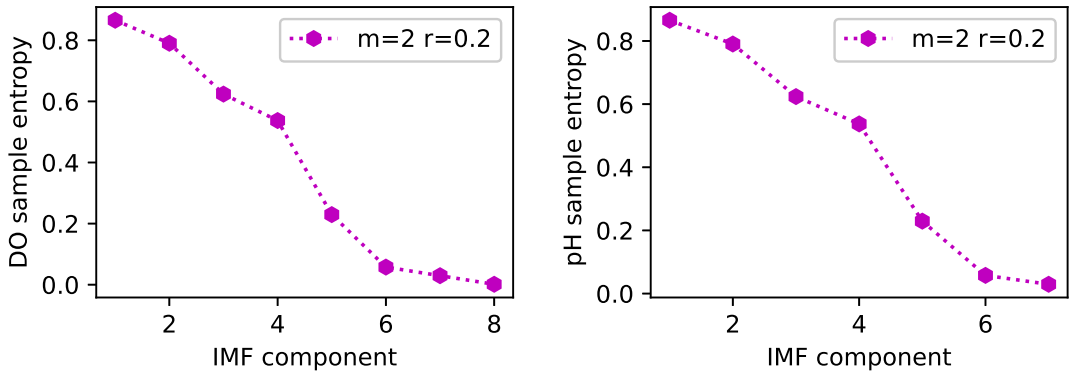

Figure 4. SE of each IMF component.

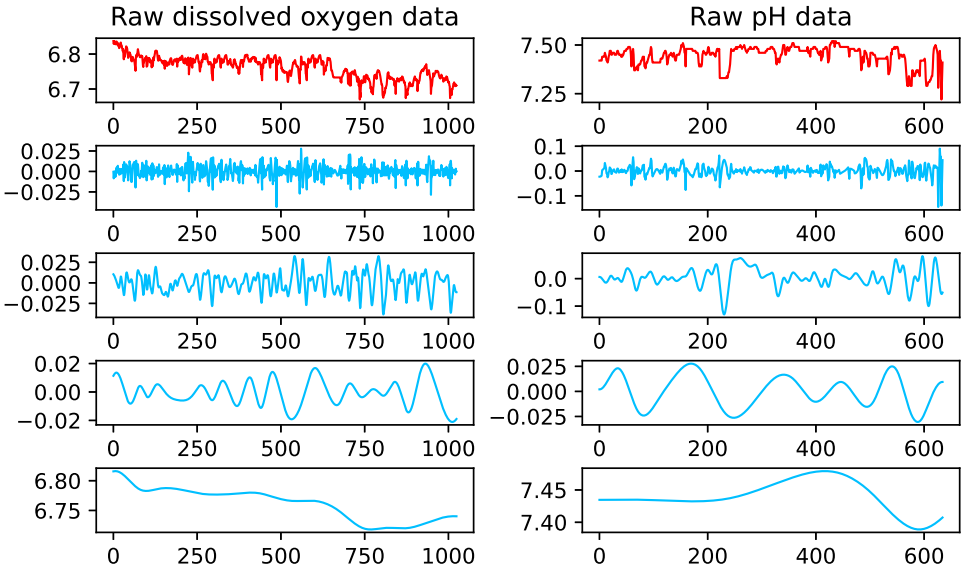

Figure 5. Diagram of the decomposed and recombined dissolved oxygen and pH sequences.

components can reduce the computational complexity of the prediction model and prevent the extraction of inaccurate information caused by overdecomposition. The IMF components of the dissolved oxygen and pH data sequences in this paper can be recombined into several subsequences, as shown in Table 1.

Table 1. Recombination of IMF components based on the SE.

Water quality factors	SE1	SE2	SE3	SE4
Dissolved oxygen	IMF1 + IMF2	IMF3 + IMF4	IMF5	IMF6 + IMF7 + R(t)
PH	IMF1 + IMF2	IMF3 + IMF4	IMF5	IMF6 + R(t)

According to the recombination scheme described in Table 1, the dissolved oxygen and pH data sequences were recombined separately, and the experimental results as shown in Figure 5.

3.4. Construction of the CEEMDAN-SE-LSTM hybrid forecasting model

This paper uses the Keras deep learning library in Python based on the TensorFlow framework, adopts a sequential model structure, and combines an LSTM network with the dense layer to build a prediction model. To train the model, the mean square error (MSE) is selected as the loss function [39], adaptive moment estimation (Adam) is used as the parameter optimizer [40], and the dropout [41] method is used to prevent overfitting. This model is optimized and trained for the input time series window length, learning rate and number of iterations and other parameters to improve the iteration convergence speed and prediction accuracy.

This paper uses the autocorrelation coefficient to determine the length of the input time window. The autocorrelation coefficient [42, 43] measures the correlation degree of the time series with itself at

different time points. The mathematical formula of the autocorrelation coefficient is expressed in Eq (3.1).

$$R_k = \frac{\sum_{i=1}^{n-k}(x_i - u)(x_{i+k} - u)}{\sum_{i=1}^{n}(x_i - u)^2} \tag{3.1}$$

where k is the lag order of the time series $X = \{x_1, x_2 \cdots x_n\}$ and u is the sample mean of the series. The value of R_k is usually between -1 and 1. When the absolute value of R_k is greater than 0.8, the k-th data in the sequence are strongly correlated with the first (k-1) data.

The four subsequences formed by the decomposition and recombination of the abovementioned dissolved oxygen sequences and their respective autocorrelation coefficients are shown in Figure 6.

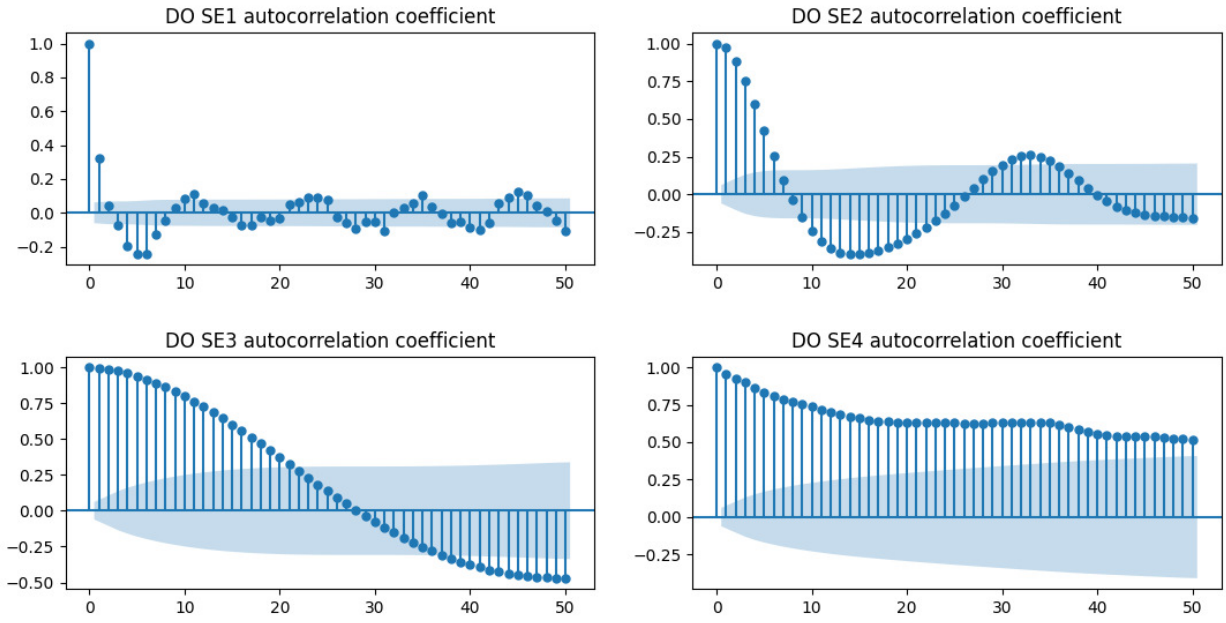

Figure 6. Dissolved oxygen subsequence autocorrelation coefficients

Likewise, the autocorrelation coefficients of the four pH subsequences are shown in Figure 7.

Taking the second dissolved oxygen subsequence (SE2) as an example, the correlation values at the first three lag orders are all greater than 0.8, indicating a strong correlation. Therefore, the length of the input time window is selected as 3; that is, the value of every 3 time points in SE2 is used to predict the value of the next time point. In the same way, the time window lengths of the other subsequences are determined by their autocorrelation coefficients.

Eighty percent of the data are extracted from each sequence to train the model. After training and verification, the settings of the parameters for each single-step prediction model, such as the learning rate, number of iterations, and batch size, are optimized. Then, the data of each subsequence test sample are input into the model for prediction; after the prediction result of each sequence is obtained, each single-step prediction value is superimposed to obtain the final predicted values of dissolved oxygen and pH.

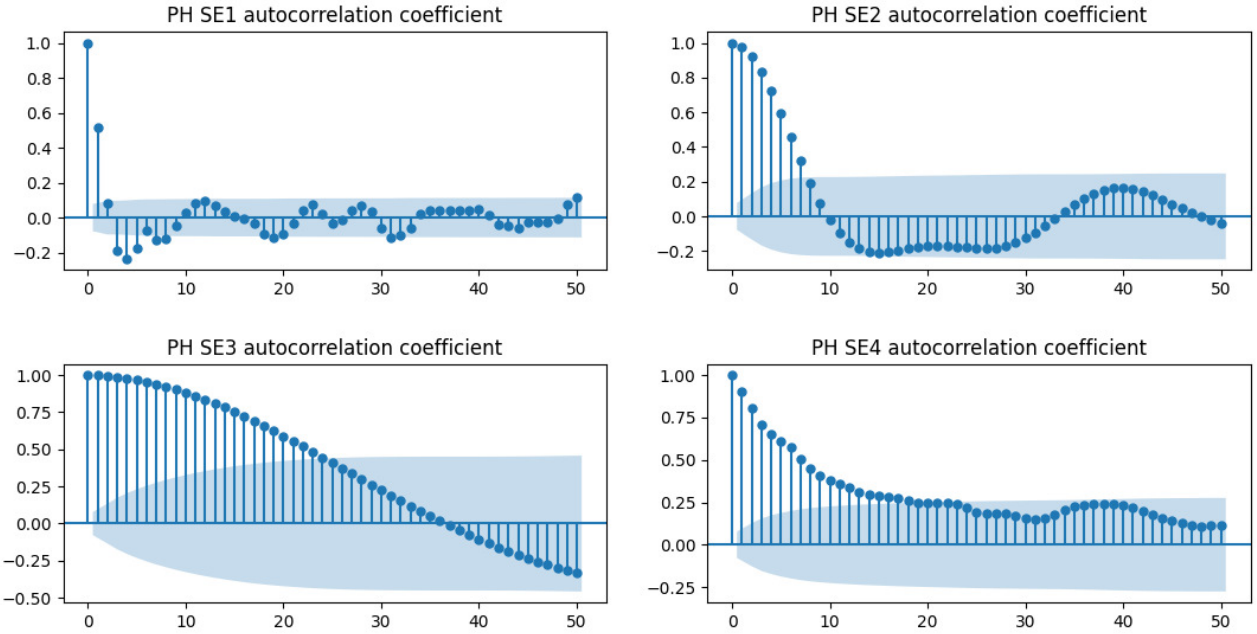

Figure 7. pH subsequence autocorrelation coefficients.

3.5. Model evaluation and comparative analysis

To verify the prediction performance of the CEEMDAN-SE-LSTM model proposed in this paper, a variety of evaluation indicators [44, 45] are used to evaluate the prediction effect of the model.

(1) The mean absolute error (MAE), the average value of the absolute error, can better reflect the actual situation of the error in the predicted value. The MAE is calculated using Eq (3.2).

$$MAE = \frac{1}{N}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{3.2}$$

(2) The root mean squared error (RMSE) is used to measure the deviation between the predicted value and the true value following Eq (3.3).

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{3.3}$$

(3) The mean absolute percentage error (MAPE) is inversely proportional to the accuracy: the smaller the MAPE is, the more accurate the prediction. The MAPE is expressed in Eq (3.4).

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\frac{|y_i - \hat{y}_i|}{y_i} \tag{3.4}$$

In the above formulas, y_i represents the true value, $\hat{y}_i$ represents the predicted value, and N is the number of samples.

Table 2. Forecast model accuracy comparison.

Prediction model	Dissolved oxygen			PH		
	MAE	RMSE	MAPE	MAE	RMSE	MAPE
RBF	0.2356	0.3136	0.0327	0.0478	0.0633	0.0065
RNN	0.2328	0.3011	0.0328	0.0450	0.0580	0.0062
GRU	0.2319	0.2783	0.0312	0.0329	0.0414	0.0042
LSTM	0.1719	0.2283	0.0289	0.0293	0.0307	0.0031
CEEMDAN-SE-RBF	0.2110	0.2854	0.0297	0.0447	0.0576	0.0062
CEEMDAN-SE-RNN	0.1970	0.2635	0.0289	0.0304	0.0225	0.0053
CEEMDAN-SE-GRU	0.1283	0.1899	0.0211	0.0237	0.0276	0.0031
CEEMDAN-SE-LSTM	**0.1026**	**0.1210**	**0.0149**	**0.0189**	**0.0107**	**0.0012**
VMD-LSTM	0.1415	0.2192	0.0218	0.0281	0.0195	0.0043
WT-LSTM	0.1498	0.2287	0.0232	0.0323	0.0414	0.0051

We implemented other three (RBF, RNN and GRU) prediction models using python language programming. We used CEEMDAN and SE to decompose and recombine the time series of dissolved oxygen and pH in aquaculture water quality, and completed other three hybrid prediction models: CEEMDAN-SE-RBF, CEEMDAN-SE-RNN and CEEMDAN-SE-GRU. With reference to related literature, we simulated the hybrid prediction model of variational model decomposition and LSTM (VMD-LSTM) [46] and the hybrid prediction model of wavelet transform and LSTM (WT-LSTM) [47]. Using the same data samples, the above models are compared with the CEEMDAN-SE-LSTM prediction model proposed in this paper. The prediction errors of each model for dissolved oxygen and pH are shown in Table 2.

The experimental results confirm that an LSTM neural network has a long-term memory function, allowing certain advantages in the prediction of water quality data time series, and the prediction accuracy is higher than that of both the RBF, RNN and GRU. Compared with the single prediction models, the hybrid prediction models based on the principles of decomposition and recombination achieve better prediction effects. Compared with the other hybrid models, the CEEMDAN-SE-LSTM prediction model proposed in this paper has the lowest prediction error and the best performance in the prediction of the dissolved oxygen and pH of the aquaculture water quality. The prediction effect of each model on pH is shown in Figure 8.

In order to test the predictive performance of the model on long-period data, we applied 14598 dissolved oxygen data for 111 days from June 2 to September 21, 2020 to conduct simulation experiments. Eighty percent of the data (11679) are selected to train the prediction model, and the remaining twenty percent of the data (2919) are used for testing. The prediction effect of each model on dissolved oxygen is shown in Figure 9.

The results of this simulation experiment demonstrate that the prediction curve of the model constructed in this paper is closer to the original water quality data curve than are those of the other prediction models. The CEEMDAN-SE-LSTM model can quickly track changes in the mutating data, and the agreement between the prediction curve and the original data curve is better than that of the other two hybrid models. Therefore, the prediction error of this model is smaller, and the fitting effect is better. Consequently, the proposed model is suitable for predicting aquaculture water quality data.

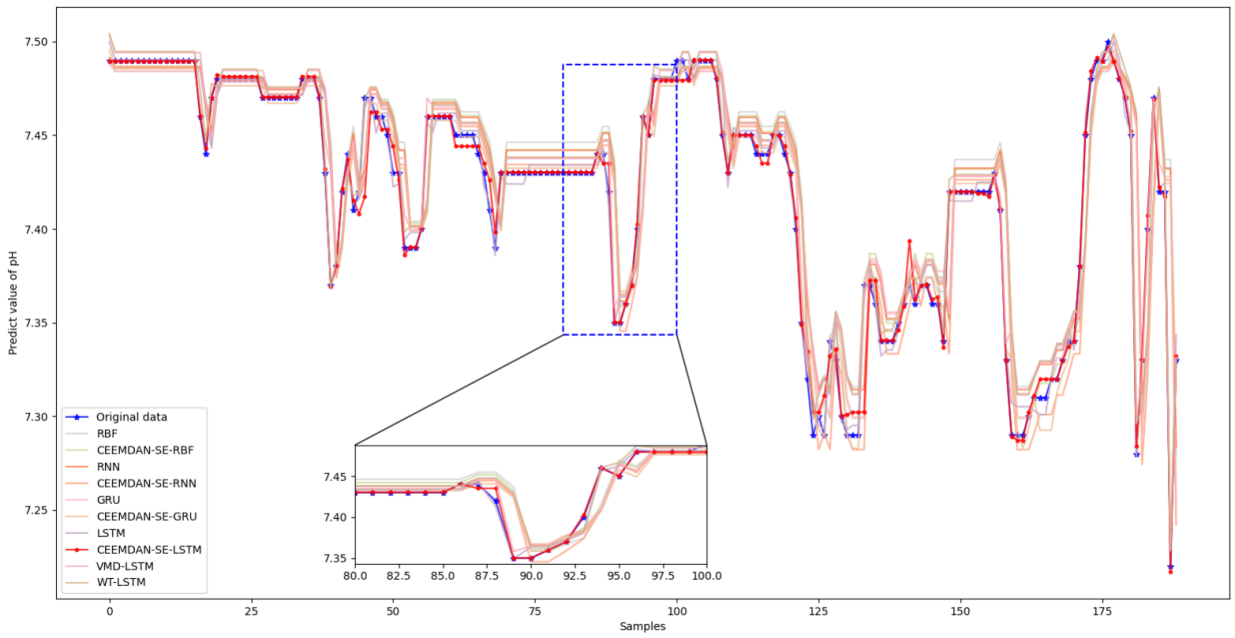

Figure 8. Experimental comparison of the prediction effect on pH.

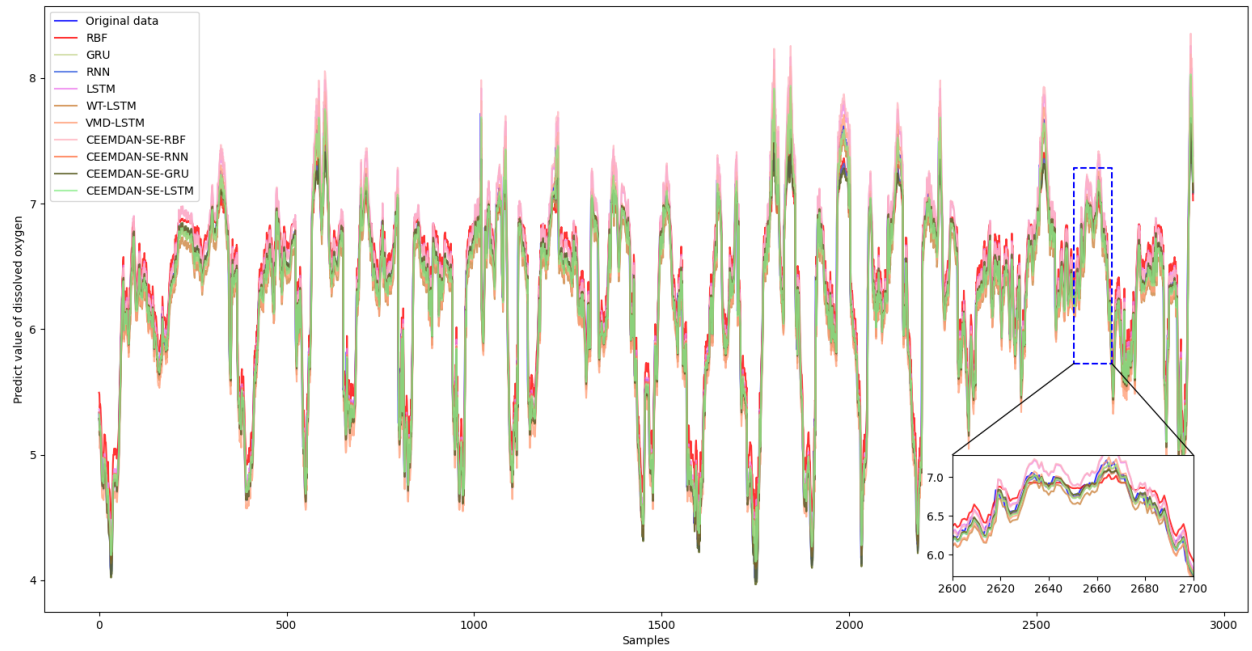

Figure 9. Experimental comparison of the prediction effect on dissolved oxygen.

4. Conclusions

The quality of aquaculture water has a tremendous impact on the growth of aquatic organisms and thus is a key factor that determines the intensive and intelligent development of aquaculture. Therefore, accurately predicting water quality has always been a key issue to be resolved in the aquaculture field. This paper focuses on this problem from two perspectives, namely, multiscale decomposition and LSTM neural network optimization, and the CEEMDAN-SE-LSTM hybrid prediction model is proposed.

The CEEMDAN algorithm does not need to set the basis function in advance and can automatically perform decomposition step by step according to the characteristics of the sequence. The individual IMF components obtained after decomposition reflect the fluctuating characteristics of the time series on different time scales. The original water quality factors are separately predicted after the sequences are decomposed, and finally, the prediction results are integrated, which can improve the prediction accuracy compared with the direct prediction of the original sequences. The LSTM neural network solves the problem of short-term memory by adding gates on the basis of a cyclic neural network model. Compared with other neural networks, LSTM has a better efficiency and higher accuracy in the prediction of time series sequences.

The prediction model proposed in this paper provides a scientific basis for accurately predicting aquaculture water quality and has important guiding significance both for the intelligent regulation and management of water quality and for ensuring the stable and efficient operation of aquaculture. However, the single-step prediction of each subsequence obtained by decomposition and recombination has a certain prediction error. Hence, the simple superposition of the single-step prediction results will increase the overall error. In the future, in-depth research will be conducted on an integrated stacking method to perform single-step prediction and further improve the prediction accuracy of the model.

Acknowledgments

This work was supported by the CERNET Innovation Project (NGII20180319); the Yantai Science and Technology Innovation Development Project (2020YT06000970, 2021XDHZ062); the Key R&D Program of Shandong Province (Soft Science Project) (2020RKB01555); and the Key R&D Program of Shandong Province (2019GGX101069). No additional external funding was received for this study.

Conflict of interest

The authors declare that there is no conflict of interests regarding the publication of this paper.

References

1. T. Rajaee, H. Jafari, Utilization of WGEP and WDT Models by Wavelet Denoising to Predict Water Quality Parameters in Rivers, *J. Hydrol. Eng.*, **23** (2018), 04018054.
2. A. H. Haghiabi, A. H. Nasrolahi, A. Parsaie, Water quality prediction using machine learning methods, *Water Qual. Res. J.*, **53** (2018), 3–13.
3. A. Rahman, J. Dabrowski, J. McCulloch, Dissolved oxygen prediction in prawn ponds from a group of one step predictors, *Inf. Process. Agric.*, **7** (2020), 307–317.

4. R. Barzegar, A. A. Moghaddam, J. Adamowski, B. Ozga-Zielinski, Multi-step water quality forecasting using a boosting ensemble multi-wavelet extreme learning machine model, *Stochastic Environ. Res. Risk. Assess.*, **32** (2018), 799–813.
5. H. Jafari, T. Rajaee, O. Kisi, Improved Water Quality Prediction with Hybrid Wavelet-Genetic Programming Model and Shannon Entropy, *Nat. Resour. Res.*, **29** (2020), 3819–3840.
6. A. P. Rozario, N. Devarajan, Monitoring the quality of water in shrimp ponds and forecasting of dissolved oxygen using Fuzzy C means clustering based radial basis function neural networks, *J. Ambient Intell. Humanized Comput.*, **11** (2020), 1–8.
7. O. Kisi, M. Alizamir, A. R. D. Gorgij, Dissolved oxygen prediction using a new ensemble method, *Environ. Sci. Pollut. Res.*, **27** (2020), 1–15.
8. W. Li, H. Wu, N. Zhu, Y. Jiang, J. Tan, Y. Guo, Prediction of dissolved oxygen in a fishery pond based on gated recurrent unit (GRU), *Inf. Process. Agric.*, **8** (2021), 185–193.
9. J. J. Dabrowski, A. Rahman, D. E. Pagendam, A. George, Enforcing mean reversion in state space models for prawn pond water quality forecasting, *Comput. Electron. Agric.*, **168** (2020), 105120.
10. Y. Chen, H. Yu, Y. Cheng, Q. Cheng, D. Li, A hybrid intelligent method for three-dimensional short-term prediction of dissolved oxygen content in aquaculture, *PLOS ONE*, **13** (2018), 1–17.
11. S. Hochreiter, J. Schmidhuber, Long Short-Term Memory, *Neural Comput.*, **9** (1997), 1735–1780.
12. X. B. Jin, R. J. RobertJeremiah, T. L. Su, Y. T. Bai, J. L. Kong, The new trend of state estimation: from model-driven to hybrid-driven methods, *Sensors*, **21** (2021), 2085.
13. D. Zhang, M. R. Kabuka, Combining weather condition data to predict traffic flow: a GRU-based deep learning approach, *IET Intell. Trans. Syst.*, **12** (2018), 578–585.
14. X. B. Jin, N. X. Yang, X. Y. Wang, Y. T. Bai, T. L. Su, J. L. Kong, Hybrid deep learning predictor for smart agriculture sensing based on empirical mode decomposition and gated recurrent unit group model, *Sensors*, **20** (2020), 1334.
15. L. Michieletto, B. Ouyang, P. S. Wills, Investigation of water quality using transfer learning, phased LSTM and correntropy loss, *Big Data II: Learning, Analytics, and Applications. International Society for Optics and Photonics*, (2020), 73–85.
16. Z. Li, F. Peng, B. Niu, G. Li, J. Wu, Z. Miao, Water quality prediction model combining sparse auto-encoder and LSTM network, *IFAC-PapersOnLine*, **51** (2018), 831–836.
17. R. Barzegar, M. T. Aalami, J. Adamowski, Short-term water quality variable prediction using a hybrid CNN–LSTM deep learning model, *Stochastic Environ. Res. Risk Assess.*, **27** (2020), 1–19.
18. J. Zhou, Y. Wang, F. Xiao, Y. Wang, L. Sun, Water quality prediction method based on IGRA and LSTM, *Water*, **10** (2018), 1148.
19. Q. Zou, Q. Xiong, Q. Li, H. Yi, Y. Yu, C. Wu, A water quality prediction method based on the multi-time scale bidirectional long short-term memory network, *Environ. Sci. Pollut. Res.*, **27** (2020), 16853–16864.
20. N. E. Huang, Z. Shen, S. R. Long, M. C. Wu, H. H. Shih, Q. Zheng, et al., The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis, *Proc. R. Soc. London, Ser. A*, **454** (1998), 903–995.

21. E. Fijani, R. Barzegar, R. Deo, E. Tziritis, K. Skordas, Design and implementation of a hybrid model based on two-layer decomposition method coupled with extreme learning machines to support real-time environmental monitoring of water quality parameters, *Sci. Total Environ.*, **648** (2019), 839–853.
22. J. Huan, W. Cao, Y. Qin, Prediction of dissolved oxygen in aquaculture based on EEMD and LSSVM optimized by the Bayesian evidence framework, *Comput. Electron. Agric.*, **150** (2018), 257–265.
23. E. Eze, T. Ajmal , Dissolved oxygen forecasting in aquaculture: a hybrid model approach, *Appl. Sci.*, **10** (2020), 7079.
24. S. Liu, L. Xu, D. Li, Multi-scale prediction of water temperature using empirical mode decomposition with back-propagation neural networks, *Comput. Electr. Eng.*, **49** (2016), 1–8.
25. W. Sun, Y. Wang, Short-term wind speed forecasting based on fast ensemble empirical mode decomposition, phase space reconstruction, sample entropy and improved back-propagation neural network, *Energy Convers. Manage.*, **157** (2018), 1–12.
26. J. Wang, X. Sun, Q. Cheng, Q. Cui, An innovative random forest-based nonlinear ensemble paradigm of improved feature extraction and deep learning for carbon price forecasting, *Sci. Total Environ.*, **762** (2021), 143099.
27. Q. Wu, H. Lin, Daily urban air quality index forecasting based on variational mode decomposition, sample entropy and LSTM neural network, *Sustainable Cities Soc.*, **50** (2019), 101657.
28. M. Rezaie-Balf, N. Maleki, S. Kim, A. Ashrafian, F. Babaie-Miri, N. W. Kim, et al., Forecasting daily solar radiation using CEEMDAN decomposition-based MARS model trained by crow search algorithm, *Energies*, **12** (2019), 1416.
29. X. B. Jin, N. X. Yang, X. Y. Wang, Y. T. Bai, T. L. Su, J. L. Kong, Deep hybrid model based on EMD with classification by frequency characteristics for long-term air quality prediction, *Mathematics*, **8** (2020), 214.
30. Z. Wu, N. E. Huang, Ensemble empirical mode decomposition: a noise-assisted data analysis method, *Adv. Adapt. Data Anal.*, **1** (2009), 1–41.
31. R. Ye, P. N. Suganthan, N. Srikanth, A comparative study of empirical mode decomposition-based short-term wind speed forecasting methods, *IEEE Trans. Sustainable Energy*, **6** (2015), 236–244.
32. J. R. Yeh, J. S. Shieh, N. E. Huang, Complementary ensemble empirical mode decomposition: A novel noise enhanced data analysis method, *Adv. Adapt. Data Anal.*, **2** (2010), 135–156.
33. A. A. Mousavi, C. Zhang, S. F. Masri, G. Gholipour, Structural damage localization and quantification based on a CEEMDAN Hilbert transform neural network approach: a model steel truss bridge case study, *Sensors*, **5** (2020), 1271.
34. M. E. Torres, M. A. Colominas, G. Schlotthauer, P. Flandrin, A complete ensemble empirical mode decomposition with adaptive noise, *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (2011), 4144–4147.
35. J. S. Richman, J. R. Moorman, Physiological time-series analysis using approximate entropy and sample entropy, *Am. J. Physiol. Heart Circ. Physiol.*, **278** (2000), H2039–H2049.

36. A. Sherstinsky, Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network, *Phys. D Nonlinear Phenom.*, **404** (2020), 132306.
37. Z. Wang, L. Yao, Y. Cai, Rolling bearing fault diagnosis using generalized refined composite multiscale sample entropy and optimized support vector machine, *Measurement*, **156** (2020), 107574.
38. J. Kumar, R. Goomer, A. K. Singh, Long short term memory recurrent neural network (LSTM-RNN) based workload forecasting model for cloud datacenters, *Procedia Comput. Sci.*, **125** (2018), 676–682.
39. Z. Chang, Y. Zhang, W. Chen, Electricity price prediction based on hybrid model of adam optimized LSTM neural network and wavelet transform, *Energy*, **187** (2019), 115804.
40. P. Nystrup, E. Lindström, P. Pinson, H. Madsen, Temporal hierarchies with autocorrelation for load forecasting, *Eur. J. Oper. Res.*, **280** (2020), 876–888.
41. A. Labach, H. Salehinejad, S. Valaee, Survey of dropout methods for deep neural networks, preprint, arXiv:1904.13310.
42. S. Afyouni, S. M. Smith, T. E. Nichols, Effective degrees of freedom of the Pearson's correlation coefficient under autocorrelation, *NeuroImage*, **199** (2019), 609–625.
43. Z. Liang, R. Zou, X. Chen, T. Ren, H. Su, Y. Liu, Simulate the forecast capacity of a complicated water quality model using the long short-term memory approach, *J. Hydrol.*, **581** (2020), 124432.
44. Ü. B. Filik, T. Filik, Wind speed prediction using artificial neural networks based on multiple local measurements in Eskisehir, *Energy Procedia*, **107** (2017), 264–269.
45. M. V. Shcherbakov, A. Brebels, A. Tyukov, A survey of forecast error measures, *World Appl. Sci. J.*, **24** (2013), 171–176.
46. H. Niu, K. Xu, W. Wang, A hybrid stock price index forecasting model based on variational mode decomposition and LSTM network, *Appl. Intell.*, **50** (2020), 4296–4309.
47. Z. Chang, Y. Zhang, W. Chen, Electricity price prediction based on hybrid model of adam optimized LSTM neural network and wavelet transform, *Energy*, **187** (2019), 115804.

AIMS Press