Hindawi
Computational Intelligence and Neuroscience
Volume 2022, Article ID 1264385, 11 pages
https://doi.org/10.1155/2022/1264385

Research Article

Water Quality Prediction Based on SSA-MIC-SMBO-ESN

Yan Kang,[1] Jinling Song,[1] Zhuo Lin,[1] Liming Huang,[2] Xiaoang Zhai,[1] and Haipeng Feng[3]

[1] *School of Mathematics and Information Science & Technology, Hebei Normal University of Science & Technology, Key Laboratory of Ocean Dynamics and Resources and Environments, Hebei Agricultural Data Intelligent Perception and Application Technology Innovation Center, Qinhuangdao 066000, Hebei, China*
[2] *School of Business Administration, Hebei Normal University of Science & Technology, Qinhuangdao 066000, China*
[3] *CITIC Dicastal Co., Ltd, Qinhuangdao 066000, Hebei, China*

Correspondence should be addressed to Jinling Song; songjinling99@126.com

Received 24 May 2022; Revised 27 June 2022; Accepted 2 July 2022; Published 3 August 2022

Academic Editor: Dalin Zhang

Water pollution threatens the safety of human production and life. To quickly respond to water pollution, it is important for water management staff to predict water quality changes in advance. Drawing on the temporality of water quality data, the leaky integrator echo state network (ESN) was introduced to construct the water quality prediction models for dissolved oxygen (DO), permanganate index (CODMn), and total phosphorus (TP), respectively. First, the missing values were filled by the linear trend method of adjacent points, and the outliers were detected and corrected by the *Z*-score method and the linear trend method. Second, the singular spectrum analysis (SSA) was performed to denoise the original monitoring data, such that the predicted data catch up with the real data, and the model accuracy is not affected by the hidden noise in the data. Third, the correlation between water quality indices was measured by the maximum information coefficient (MIC), and the strongly correlated indices were imported to the prediction model. Finally, according to these strong correlation indicators, the water quality prediction models based on multiple features were constructed, respectively, using the offline and online learning algorithms of the ESN. The hyperparameters of the models were optimized through the sequential model-based optimization (SMBO). Experimental results show that the proposed water quality prediction models, namely, SSA-MIC-SMBO-Offline ESN and SSA-MIC-SMBO-Online ESN, predicted DO, CODMn, and TP accurately, providing suitable tools for practical applications.

1. Introduction

Water is an important resource for human survival and social development. Unfortunately, water shortage becomes an increasingly severe global problem. One of the major causes of water shortage is water pollution. With limited per-capita water resources, China is deeply troubled by water pollution. According to the 2020 Report on the State of the Ecology and Environment in China (https://www.mee.gov.cn/hjzl/sthjzk/zghjzkgb), 16.6% of the monitored surface water areas in China belong to Classes IV and V in terms of water quality. As the country stepped up the protection of water resources, the focus of water pollution control has shifted from posttreatment to preprevention. To effectively reduce water pollution, it is important to predict the future trend of water quality accurately. The early warning would promote the scientific management of water resources, maintain the sustainability of the ecosystem, and protect human health.

In recent years, many water quality prediction models have been developed based on a dazingly array of technologies, namely, multiple linear regression (MLR) [1], regression tree and support vector regression (SVR) [2], nonlinear least squares (NLS) neural network [3, 4], radial basis function (RBF) neural network [5–7], long short-term memory (LSTM) neural network [8–11], and graph neural network (GNN) [12, 13]. These models greatly advance the prediction of water quality. However, the model performance is affected by various factors in the complex and dynamic system of the water environment, including data quality, input features, prediction algorithm, and selected parameters. Therefore, the existing models should be further improved to forecast water quality more accurately.

This paper constructs the prediction models of water quality indices through comprehensive use of multiple techniques: singular spectrum analysis (SSA), maximum information coefficient (MIC), offline and online learning algorithms of the leaky integrator echo state network (ESN) [14], and sequential model-based optimization (SMBO). First, the original data were smoothed and denoised through the SSA. Next, the correlation between water quality indices was measured by the MIC. Finally, the data on relevant indices were collected from river monitoring stations, and imported to the ESN, the hyperparameters of the network were optimized through the SMBO, and the prediction models were established for dissolved oxygen (DO), permanganate index (CODMn), and total phosphorus (TP), respectively. Experimental results show that the proposed water quality prediction models, namely, SSA-MIC-SMBO-Offline ESN and SSA-MIC-SMBO-Online ESN, forecasted each water quality index accurately and practically.

2. Leaky Integrator ESN

The ESN, a novel recurrent neural network (RNN) [14], centers on a large, randomly generated, and sparsely connected reservoir. The only thing that needs to be trained in the network is the output connection weight. Therefore, the ESN boasts the advantages of simple structure and fast training. During the training, the network is not prone to falling to the local optimal trap, a common defect of traditional RNN. In this way, the network weights are always optimized globally.

2.1. Network Structure. As shown in Figure 1, the leaky integrator ESN consists of an input layer; a hidden layer, i.e., the reservoir; and an output layer. There are K nodes in the input layer, L nodes in the output layer, and N nodes in the reservoir. At time t, the input, the reservoir state, and the output are denoted as $u(t)$, $x(t)$, and $y(t)$, respectively. The $N\times K$ connection weight matrix from the input layer to the reservoir can be expressed as Win. The $N\times N$ connection weight matrix of the reservoir from the current moment to the next moment can be expressed as W. The $N\times L$ weight matrix of the feedback connection from the output layer to the reservoir can be expressed as W_{fb}, and the connection is unnecessary. The $L\times(N+K)$ connection weight matrix from the reservoir to the output layer can be expressed as W_{out}. Reservoir, and an output layer. There are K nodes in the input layer, L nodes in the output layer, and N nodes in the reservoir. At time t, the input, the reservoir state, and the output are denoted as $u(t)$, $x(t)$, and $y(t)$, respectively. The $N\times K$ connection weight matrix from the input layer to the reservoir can be expressed as Win. The $N\times N$ connection weight matrix of the reservoir from the cu.

The reservoir state $x(t)$ can be updated based on the current input $u(t)$ and the reservoir state $x(t-1)$ at the previous time, using leaky integrator neurons:

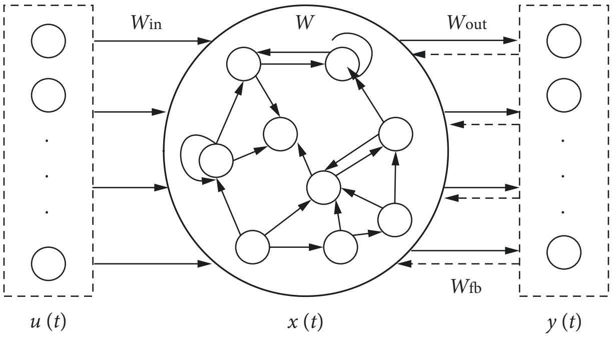

Figure 1: Structure of the leaky integrator ESN.

$$x(t) = (1-\alpha)x(t-1) + \alpha f\left(W_{\text{in}}u(t) + Wx(t-1) + W_{fb}y(t-1)\right), \tag{1}$$

where W_{in}, W, and W_{fb} are randomly initialized and fixed before network training; f is the activation function; α is the leakage rate of the neuron; and $y(t-1)$ is the output at the previous time. In this study, the activation function (tanh) and output state equation of the network can be, respectively, expressed as:

$$f(x) = \frac{e^{x}-e^{-x}}{e^{x}+e^{-x}}, \tag{2}$$

$$y(t) = W_{\text{out}}x(t). \tag{3}$$

2.2. Learning Algorithms. In the ESN, the input connection weight matrix W_{in} and reservoir connection weight matrix W remain unchanged after initialization. Thus, the output connection weight matrix W_{out} is the only model parameter that needs to be trained. In essence, the learning process of the ESN is the solving process of W_{out}. The ESN has two kinds of learning algorithms.

2.2.1. Offline Learning. In the training process of offline learning, to prevent overfitting, W_{out} was solved through ridge regression based on the regularization coefficient [15, 16]. This approach adds a regularization term to the objective function. The output weight matrix W_{out} can be updated by:

$$W_{\text{out}} = \left(X^{T}X + \lambda I_{N}\right)^{-1}X^{T}Y, \tag{4}$$

where X is the state matrix of the reservoir; λ is the ridge parameter; I_N is an N-dimensional identity matrix; and Y is the target output matrix.

2.2.2. Online Learning. Online learning updates the current model with continuously generated new data to make better predictions of future data. The common online learning algorithm is recursive least squares (RLS) [17], which updates W_{out} for each time step to minimize the prediction error. By the RLS, the output weight matrix W_{out} can be updated by:

$$P(t) = P(t-1) - \frac{P(t-1)x(t)x^{\mathrm{T}}(t)P(t-1)}{1 + x^{\mathrm{T}}(t)P(t-1)x(t)}, \quad (5)$$

$$W_{\mathrm{out}}(t) = W_{\mathrm{out}}(t-1) + [d(t) - y(t)]P(t)x(t), \quad (6)$$

where $x(t)$ is the reservoir state; $x^T(t)$ is the transpose of $x(t)$; $d(t)$ is the expected output; $y(t)$ is the actual output; and $P(t)$ is the error covariance matrix at time t.

In this paper, the ESN using offline learning and that using online learning are denoted as Offline ESN and Online ESN, respectively.

2.3. Network Training. According to the principle of the ESN and the input data u, the training steps of Offline ESN and Online ESN can be summarized as follows:

2.3.1. Training Steps of Offline ESN

Step 1: Initialize the reservoir. Configure the relevant parameters; initialize W_{in}, W, W_{fb}, and W_{out}; and set the reservoir state to the zero vector.

Step 2: Update and collect the internal state of the reservoir. Drive the reservoir with the input data u and update the internal state of the reservoir continuously by formula (1). Note that, to overcome the influence of the initial transient, the state of the previous period is generally abandoned, and the internal state is collected from time T.

Step 3: Calculate $\mathrm{W}_{\mathrm{out}}$ of the ESN by formula (4).

Step 4: Calculate the output y of the network by formula (3).

2.3.2. Training Steps of Online ESN

Step 1: initialize the reservoir. Configure the relevant parameters, e.g., reservoir size; initialize W_{in}, W, W_{fb}, and W_{out}; and set the initial reservoir state to the zero vector.

Step 2: initialize the relevant parameters in the RLS. Initialize $P(0) = \psi^{-1}I$, where ψ is generally 1×10^{-8}, and I is the identity matrix.

Step 3: input the sample data u into the network, calculate the internal state matrix $x(t)$ of the ESN by formula (1), update $\mathrm{W}_{\mathrm{out}}$ by formulas (5) and (6), and solve the corresponding output $y(t)$ by formula (3).

Step 4: repeat Step 3 until the input of u is completed.

Each connection weight matrix of the ESN can be initialized as follows: each element of the input connection weight matrix W_{in} is initialized to random number uniformly distributed in [−0.5, 0.5] and adjusted by the input scaling factor η. The reservoir connection weight matrix W is initialized to random numbers uniformly distributed in [−0.5, 0.5], discretized according to the sparsity s, and readjusted by the target spectral radius ρ. The output connection weight matrix W_{out} is initialized as a zero matrix.

Considering the principle of the ESN and the need to initialize the connection weight matrices, several parameters need to be set in advance: reservoir size N, spectral radius ρ, leakage rate α, ridge parameter λ, input scaling factor η, and sparsity s. By optimizing the values of these ESN parameters, the accuracy of the prediction model can be significantly improved.

3. Data Preprocessing

The original data were collected from the records of the automatic monitoring station of Dongzhen Reservoir in the middle reaches of Mulan river. This reservoir is the drinking water source of over 1.5 million people in Chengxiang District, Licheng District, Xiuyu District, Bei'an Development Zone, and Meizhou Island. Therefore, it is of great significance to precisely predict the water quality at this station.

The water quality at the station is monitored in days. A total of 1,095 pieces of data from 2018 to 2020 were collected, including water temperature (WT), pH, DO, conductivity (Con), turbidity (Tur), ammonia nitrogen (NH_3–N), CODMn, TP, total nitrogen (TN), and chlorophyll (aChl). The blue-green algae metrics were discarded for the many missing values.

According to the 2020 Report on the State of the Ecology and Environment in China, DO, TP, and CODMn are the main pollution-monitoring indices of surface water in China. Thus, these three water quality indices were selected as prediction objects. Based on the monitoring data of the station, several prediction models were established for the selected indices. The models can apply to other monitoring stations in the Mulan river basin, facilitating the water quality prediction of the entire basin.

The original data contain missing values, outliers, and noise, which may affect the prediction accuracy. To eliminate the effect, some preprocessing operations were implemented on the original data. Specifically, the missing values were filled by the linear trend method of adjacent points, and the outliers were checked and corrected by the Z-score method and the linear trend method.

In addition, the SSA [18] was employed to extract signals representing different components of the time series and thus reduced to noise in the time series [19, 20]. The noise reduction solves two common problems: the predicted data fall behind the real data, if the time series data have frequent random fluctuations; the prediction accuracy is suppressed by the implicit noise in the data.

During the SSA, a quarter, i.e., 90 days, was taken as the window length. On this basis, a principal component analysis (PCA) was conducted on the decomposed time series data of water quality. The series of the components, whose cumulative explained variance ratio (EVR) reaches 90%, was taken as the real data, and the remaining components were discarded as noise.

For time series reconstruction, the first 35 components were selected for Con; the first 36 components were selected for NH_3–N and aChl; the first 37 components were selected for DO, Tur, and pH; the first 38 components were selected for TP and CODMn; and the first 39 components were selected for WT and TN. Figures 2–4 display the time series of DO, CODMn, and TP before and after reconstruction.

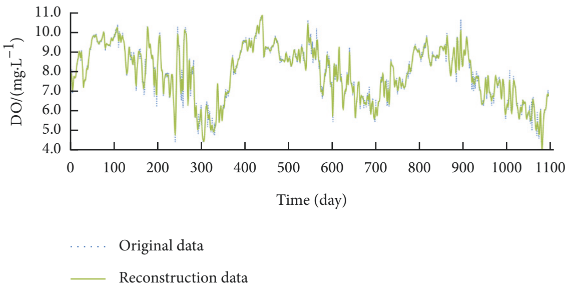

Figure 2: Reconstructed series of DO.

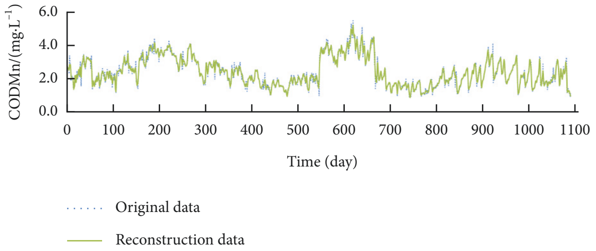

Figure 3: Reconstructed series of CODMn.

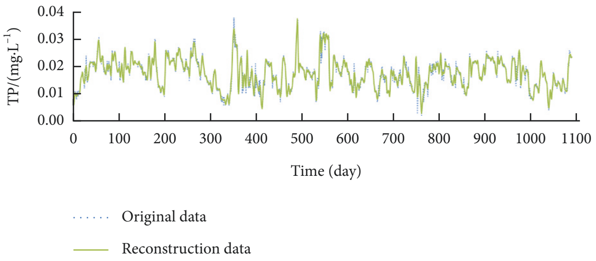

Figure 4: Reconstructed series of TP.

Compared with the original time series, the reconstructed time series contain no extreme values and have obvious cycle and trend. Therefore, the reconstructed data were adopted as the real data for water quality indices.

The denoised water quality data were further normalized through max-min normalization, which maps sample values to the interval [0, 1]:

$$X' = \frac{X - X_{\min}}{X_{\max} - X_{\min}}, \tag{7}$$

where $X_{\min}$ and $X_{\max}$ are the minimum and maximum values in the sample, respectively; X is the original value of the sample.

4. MIC-Based Correlation Analysis of Water Quality Indices

The water environment is a complex dynamic system affected by many factors. The system faces both material changes and energy exchanges. Besides, there may be some correlations between water quality indices. The indices related to the prediction index could be used together as input features. Drawing on the mutual information theory, this section analyzes the correlation between water quality indices.

Table 1: Input features of prediction model.

Predictive evaluation indicators	Input features
DO	DO, pH, NH_3–N, Con
CODMn	CODMn, WT, Tur, aChl
TP	TP, Tur, pH, Con

The mutual information of a random variable pair (X, Y) was adopted to describe the amount of information about variable Y contained in variable X:

$$I(X;Y) = \sum_{x\in X}\sum_{y\in Y} p(x,y)\log\frac{p(x,y)}{p(x)p(y)}, \tag{8}$$

where $p(x, y)$ is the joint distribution of X and Y; $p(x)$ and $p(y)$ are the marginal distributions of X and Y, respectively.

The MIC [21] measures the correlation between variables. The coefficient falls between 0 and 1. The greater the value, the stronger the correlation [22]. MIC can be calculated by:

$$\mathrm{MIC}(x,y) = \max_{ij<B}\frac{I[D(x,y)]}{\log \min(x,y)}, \tag{9}$$

where $ij < B$ is the constraint of the total number of grids; B is set to the power of θ of the size of dataset D. Relevant studies show that the optimal θ is 0.6, when the dataset size is between 1,000 and 2,500 [23].

Unlike mutual information, MIC can solve continuous variables at a high accuracy. With sufficient samples, MIC can detect a wide range of linear and nonlinear relationships between variables, making it possible to mirror the degree of correlation between water quality indices [24]. The specific value of MIC can be determined as follows:

Step 1: For the two given random variables (X, Y), rearrange the data elements of each variable in a certain order, producing an ordered pair set D(X, Y).

Step 2: Mesh the scatterplot of the ordered pair set D(X, Y) into i columns and j rows (the values of i and j are given) and then calculate I(X; Y).

Step 3: Normalize the obtained mutual information.

Step 4: Select different combinations of i and j to divide the random variables, and then find out the maximum mutual information (i.e., MIC) for each division method.

Here, the MIC analysis algorithm is adopted to analyze the correlation between DO, CODMn, and TP of Dongzhen Reservoir; the three target indices; and three other water quality indices (Table 1). The results in Table 2 show that the three other indices are strongly correlated with the three target indices. Thus, these three indices were taken as the common input features of the target indices.

5. SSA-MIC-SMBO-ESN Water Quality Prediction Model

The transmission and diffusion of pollutants in the upstream affect the concentration of pollutants in the downstream.

Table 2: Results of MIC correlation analysis.

	WT	pH	DO	Tur	NH_3–N	CODMn	TP	TN	aChl	Con
DO	0.210	0.497	1	0.219	0.310	0.131	0.192	0.148	0.245	0.283
CODMn	0.363	0.199	0.131	0.361	0.143	1	0.171	0.249	0.316	0.314
TP	0.168	0.223	0.192	0.217	0.192	0.171	1	0.158	0.156	0.232

Considering this effect, the data on water quality indices were collected from four inflow river-monitoring stations (Changtaixi Tukeng Reservoir, Dongtaixi Dongtai Village, Dulixi Xiashan Village, and Juxi Guoxi Village) and added to the ESN as input features. For each target index, a total of eight features were imported to the ESN.

To ensure the timeliness of the forecast, a 3-day forecast model was constructed for each index. The input features are denoted as $t-w+1$ to t, and the outputs (predicted values) are denoted as $t+1$, $t+2$, and $t+3$. Then, the ESN has $K=8w$ input nodes and $L=3$ output nodes. Note that w is a dynamic optimization parameter.

To fully mine the information of each time series, the monitoring data of the first 1,045 days were organized into the training set, and those of the last 50 days into the test set. In addition, the first 45 states were discarded to eliminate the influence of the initial state.

5.1. SMBO-Based Hyperparameter Optimization. According to the network training process, the following hyperparameters of the ESN need to be optimized, namely, reservoir size N, spectral radius ρ, leakage rate α, ridge parameter λ, input scaling factor η, sparsity s, and window length w. These hyperparameters were optimized to build our water quality prediction model. The optimization of hyperparameters essentially aims to improve the loss function in the configuration space. Suppose the optimization problem pursues minimization, then, hyperparameter optimization can be expressed as:

$$x^* = \arg\min_{x\in X} f(x), \quad (10)$$

where $f(x)$ is the objective function to be minimized; x is the set of hyperparameters (N, ρ, α, λ, η, s, and w); X is the hyperparameter domain; and x^* is the set of the smallest hyperparameters. x can take any value in the domain X.

The evaluation of the objective function is generally costly, and the manual or grid search parameter tuning method is time-consuming. By contrast, Bayesian optimization greatly improves the search efficiency, as it approximates the objective function with a low-cost adaptive surrogate model. Bayesian optimization aims to model the objective function, according to the existing N groups of experimental results $H=\{x_n, y_n\}$ ($n\in[1,N]$, where y_n is the observed value of $f(x_n)$), and to calculate $p(y|x,H)$ (i.e., the surrogate model) of y. The surrogate model uses the tree-structured Parzen estimator (TPE). The value of $p(y|x)$ can be calculated by:

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)}. \quad (11)$$

Sufficient sampling is critical to make the surrogate model approach the objective function. To reduce the cost of sample generation, it is favorable to control the sample size being used. Hence, formula (9) was adopted to evaluate the profit brought by a sample to the surrogate model. The larger the profit, the closer the updated surrogate model will be to the objective function. Generally, the profit can be measured by expected improvement (EI):

$$\mathrm{EI}(x,H) = \int_{-\infty}^{\infty} \max(y* - y, 0)p(y\,|\,x,H)\mathrm{d}y, \quad (12)$$

where $y^*=\min\{y_n,\ 1\le n\le N\}$ is the optimal value in the current existing samples. The EI refers to the expectation that the value y of a sample x under the current surrogate model $p(y\,|\,x,\,H)$ exceeds the best result y^*.

The SMBO is a specific algorithm to implement Bayesian optimization [25]. The algorithm 1can be described as follows:

For hyperparameter optimization by the SMBO, the search space and step size of each parameter were configured as shown in Table 3, and the number of iterations LT was set to 5,000.

5.2. Performance Metrics. The predictive power of our water quality prediction models was evaluated by root mean square error (RMSE), mean absolute percentage error (MAPE), and Nash–Sutcliffe efficiency (NSE) [26, 27]. Among them, the NSE specifically verifies the simulation results of the hydrological model. The value of NSE falls between 0 and 1. The larger the value, the stronger the prediction ability of the model. The RMSE, MAPE, and NSE can be, respectively, calculated by:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}, \quad (13)$$

$$\mathrm{MAPE} = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{\widehat{y}_i - y_i}{y_i}\right|, \quad (14)$$

$$\mathrm{NSE} = 1 - \frac{\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}, \quad (15)$$

where y_i is the target output; $\hat{y}_i$ is the network output; and □ is the mean of the target output.

5.3. Construction of the SSA-MIC-SMBO-Offline ESN Model. Offline learning, also known as batch learning, inputs all sample data at once to train the model. During training, the data on the eight input features for each target index were imported to the Offline ESN. The hyperparameters of the

```
    SMBO ($f(x)$, $T$, EI ($x$, $M$))
    Inputs: Objective function $f(x)$, number of iterations LT, and profit function EI($x$, $M$)
    Output: $H$ / * $H$ is the set of hyperparameters and profit function values * /
$H \longleftarrow \emptyset$;
Use the TPE to establish a proxy probability model $M_0$ to obtain the posterior probability $p_M(f(x) \mid x, H)$
for $t \longleftarrow 1$ to LT do
    $x' \longleftarrow \mathrm{argmax}_x$ EI($x$, $M_{t-1}$); / * Use the acquisition function to find the set of parameters * /
    Evaluate $y' = f(x')$; / * Bring the set of parameters into the objective function to evaluate the effect of the surrogate model $M(H)$ * /
    $H \longleftarrow H \cup (x', y')$; / * Update $H$ * /
    Update the surrogate model $M_t$ by $H$, and calculate $p_M(f(x) \mid x, H)$;
end
return $H$;
```

Algorithm 1: SMBO algorithm.

Table 3: Parameter space optimized by SMBO.

Parameter	Range	Step size
N	[50, 400]	10
P	[0.1, 1.5]	0.1
A	[0.01, 0.1]	0.01
H	[0.1, 1]	0.1
s	[0.01, 0.1]	0.01
λ	[1e − 5, 1]	Increase 10 times each time
w	[3, 15]	1

Offline ESN were optimized by SMBO. Table 4 displays the optimization results. On this basis, the authors established the SSA-MIC-SMBO-Offline ESN prediction model for each water quality index. Table 5 shows the index values obtained by verifying each prediction model on the test set. Figures 5–7 compare the predicted value with the actual value of each prediction model.

As shown in Table 5 and Figures 5–7, the SSA-MIC-SMBO-Offline ESN prediction models of DO, CODMn, and TP generally performed well. In particular, the predicted value and actual value basically overlapped in the first 2 days. Thus, the SSA-MIC-SMBO-Offline ESN 3-day prediction model constructed for each water quality index can meet the requirements of practical application. However, the predicted values of the three indices on the second and third days gradually deviated from the actual values, with the deterioration of each evaluation index. This means the prediction effect weakens over the time. Specifically, the prediction indices of all prediction models were good on the first day, where the NSE values were all above 0.98, indicating that the prediction models have a high prediction accuracy for the next day. The NSE values of the prediction indices were all around 0.9 on the second day (only the NSE of the CODMn was slightly low), suggesting that the accuracy on the second day also maintains a high level. The NSE values of the prediction indices of each prediction model reached about 0.8 on the third day, and the values of other evaluation indices were also within the acceptable range. Hence, the prediction effect on the third day can also meet the application requirements.

Table 4: Hyperparameter optimization results of the SSA-MIC-Offline ESN 3-day prediction model.

Evaluation index	Model parameters
DO	$N = 140$, $\rho = 0.63$, $\alpha = 0.23$, $\lambda = 1.E - 04$, $\eta = 0.1$, $s = 0.06$, $w = 5$
CODMn	$N = 170$, $\rho = 0.61$, $\alpha = 0.49$, $\lambda = 1.E - 04$, $\eta = 0.8$, $s = 0.01$, $w = 10$
TP	$N = 70$, $\rho = 1.29$, $\alpha = 0.63$, $\lambda = 1.E - 02$, $\eta = 0.4$, $s = 0.06$, $w = 5$

Table 5: Index values of SSA-MIC-SMBO-Offline ESN 3-day prediction model.

Water quality indicator	Prediction step	RMSE	MAPE	NSE
DO	1	0.0091	0.059	0.992
	2	0.0314	0.175	0.906
	3	0.0608	0.603	0.824
CODMn	1	0.0194	0.092	0.984
	2	0.0607	0.356	0.887
	3	0.0946	0.649	0.753
TP	1	0.0124	0.028	0.991
	2	0.0341	0.080	0.938
	3	0.0556	0.135	0.840

5.4. Construction of the SSA-MIC-SMBO-Online ESN Model. The data on the eight input features for each target index were imported to the Online ESN. The hyperparameters of the Online ESN were optimized by SMBO. Table 6 displays the optimization results. On this basis, the authors established the SSA-MIC-SMBO-Online ESN prediction model for each water quality index. Table 7 shows the index values obtained by verifying each prediction model on the test set. Figures 8–10 compare the predicted value with the actual value of each prediction model.

Experimental results show that the SSA-MIC-SMBO-Online ESN prediction models of DO, CODMn and TP output ideal prediction index values and prediction curves in the first two days. On the first day, the prediction index values of each prediction model were very good. The RMSEs were around 0.01, the MAPEs were within 0.09, and the

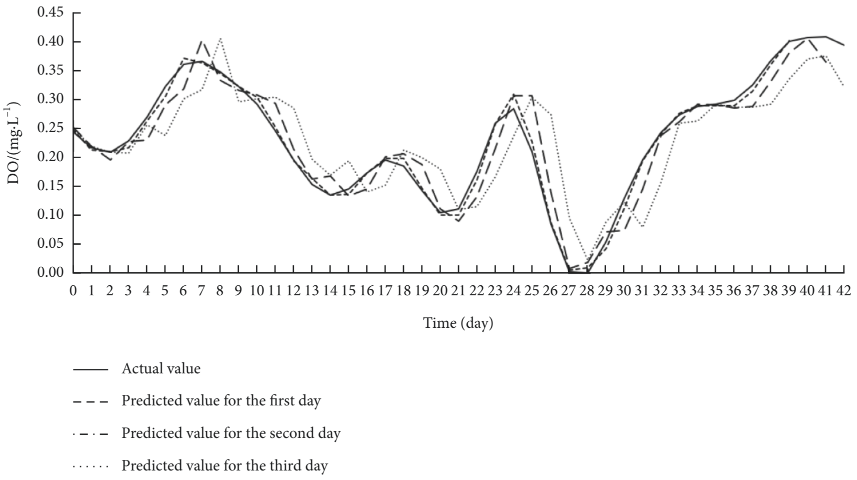

Figure 5: Effect of SSA-MIC-SMBO-Offline ESN 3-day prediction model for DO.

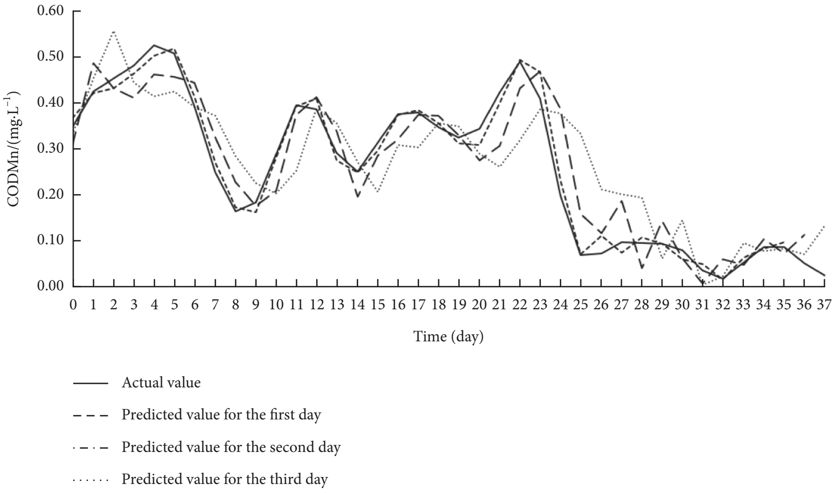

Figure 6: Effect of SSA-MIC-SMBO-Offline ESN 3-day prediction model for CODMn.

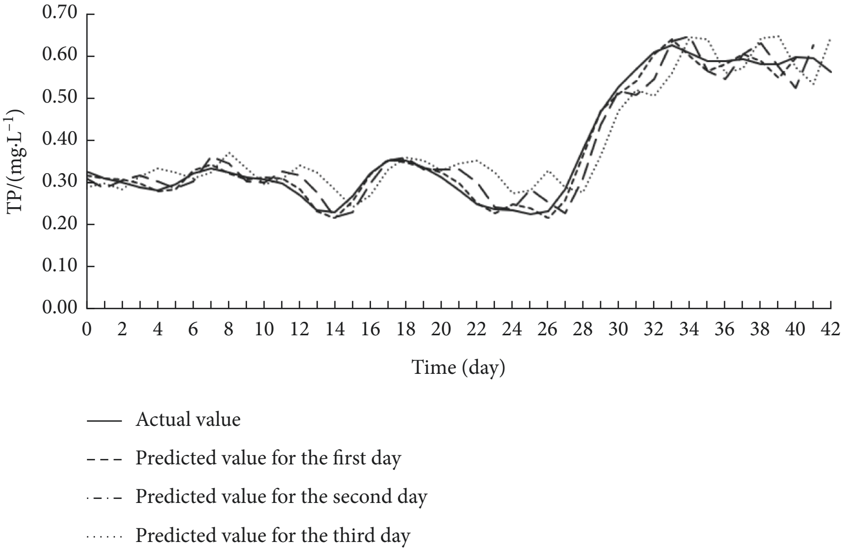

Figure 7: Effect of SSA-MIC-SMBO-Offline ESN 3-day prediction model for TP.

Table 6: Hyperparameter optimization results of the SSA-MIC-Online ESN 3-day prediction model.

Evaluation index	Model parameters
DO	$N=50$, $\rho=0.05$, $\alpha=0.91$, $\eta=0.7$, $s=0.05$, $w=7$
CODMn	$N=130$, $\rho=0.77$, $\alpha=0.31$, $\eta=0.9$, $s=0.04$, $w=14$
TP	$N=50$, $\rho=1.32$, $\alpha=0.98$, $\eta=0.2$, $s=0.05$, $w=4$

Table 7: Index values of the SSA-MIC-SMBO-Online ESN 3-day prediction model.

Water quality indicator	Prediction step	RMSE	MAPE	NSE
DO	1	0.0087	0.058	0.993
	2	0.0299	0.153	0.919
	3	0.0503	0.484	0.846
CODMn	1	0.0187	0.089	0.986
	2	0.0573	0.332	0.897
	3	0.0814	0.506	0.786
TP	1	0.0102	0.022	0.994
	2	0.0328	0.076	0.941
	3	0.0527	0.118	0.867

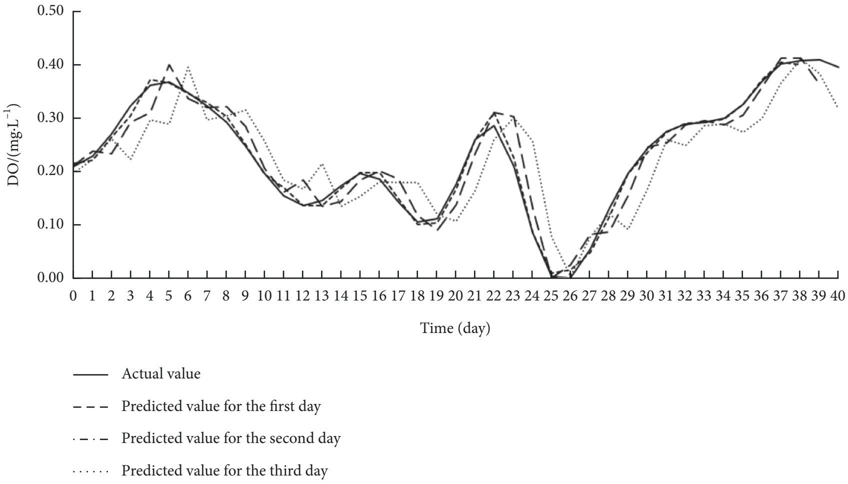

Figure 8: Effect of SSA-MIC-SMBO-Online ESN 3-day prediction model for DO.

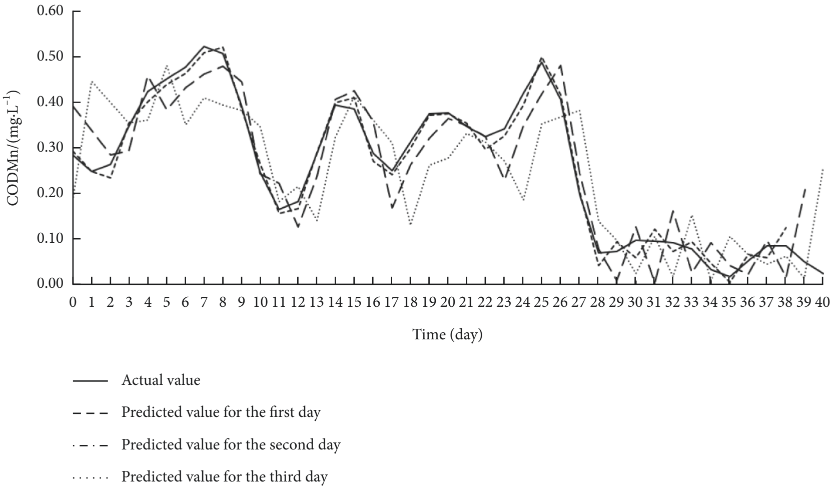

Figure 9: Effect of SSA-MIC-SMBO-Online ESN 3-day prediction model for CODMn.

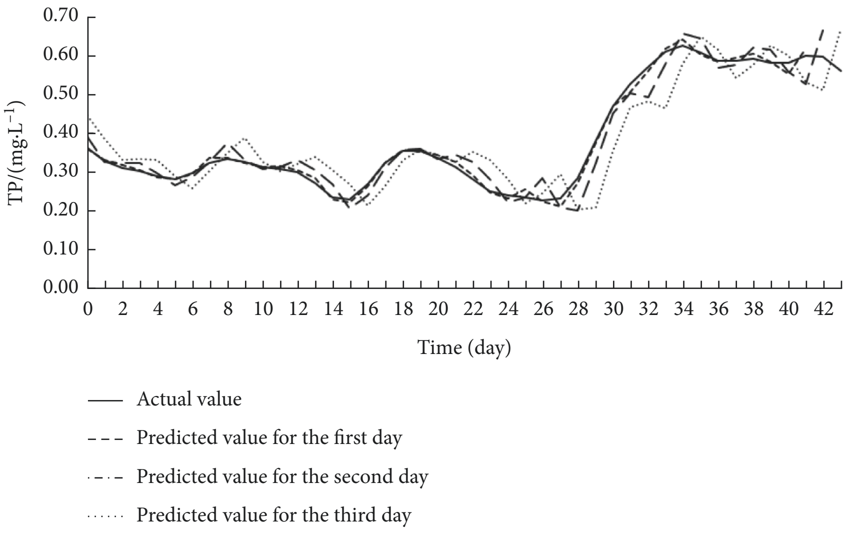

Figure 10: Effect of SSA-MIC-SMBO-Online ESN 3-day prediction model for TP.

NSEs were above 0.985, indicating that the prediction models have a very high prediction accuracy for the next day. On the second day, the RMSEs of the prediction indices were all around 0.04, and the NSE values were around 0.9 (only the NSE value of the CODMn was slightly low), indicating that the accuracy can also reach a high level on the second day. On the third day, the RMSEs of the prediction models reached about 0.06, the NSEs stood at about 0.8, and the MAPEs were within the acceptable range. Overall, the prediction effect on the third day is significantly lower than that on the previous two days, but it can still meet the application requirements. It can also be seen that, on the first day, the prediction curves of the three water quality indices largely overlapped the actual curves. On the second and third days, the prediction curves gradually deviated from the real curves, and exhibited some oscillations. However, the two sets of curves obeyed consistent trends. In general, the SSA-MIC-SMBO-Online ESN 3-day prediction models for DO, CODMn and TP achieved desirable performances, indicating that the model is feasible to predict each water quality index.

5.5. Discussion. From the experimental results, it is easy to find that the SSA-MIC-SMBO-Offline ESN and SSA-MIC-SMBO-Online ESN water quality prediction models can basically forecast the values of DO, CODMn, and TP. Further comparison shows that the SSA-MIC-SMBO-Online ESN model had better evaluation indices than the SSA-MIC-SMBO-Offline ESN model. On the first day, the three water quality indices predicted by SSA-MIC-SMBO-Online ESN were about 0.001 smaller than those predicted by SSA-MIC-SMBO-Offline ESN, while the NSEs were 0.002 larger. On the second day, the RMSE of DO predicted by SSA-MIC-SMBO-Online ESN was 0.0015 smaller than that of DO predicted by SSA-MIC-SMBO-Offline ESN, while the NSEs were 0.013 larger. The other two indices predicted by SSA-MIC-SMBO-Online ESN were also better than those predicted by SSA-MIC-SMBO-Offline ESN in the same period. Overall, the SSA-MIC-SMBO-Online ESN prediction models were more accurate than the SSA-MIC-SMBO-Offline ESN prediction models and more suitable for projecting the dynamics changes in water quality data.

Concerning the commonality of the prediction effects of the SSA-MIC-SMBO-Offline ESN or SSA-MIC-SMBO-Online ESN on the three target indices, the TP prediction model had the best index value than DO and CODMn prediction models in the same period. The DO prediction model outperformed the CODMn prediction model in terms of index value and prediction effect. Meanwhile, the CODMn prediction model had the lowest accuracy. Judging by the performance of the offline and online prediction models of the three indices, it can be observed that the prediction performance is closely related to the data features of the target indices. The prediction performance is good when the index data are clear and not frequently changeable. But the prediction accuracy tends to be low when the index data (e.g., CODMn) fluctuate greatly and frequently, making it difficult for any prediction model to capture the laws and features of the data. This is also the limitation of this prediction method. Therefore, the performance of the prediction models can be further enhanced by inputting more relevant features and expanding the training set.

6. Conclusions

Based on the important water quality indices related to water pollution, this paper separately establishes ESN-based offline and online water quality prediction models by using singular spectrum analysis algorithm, maximum mutual information coefficient, echo state network, and sequential model optimization algorithm. These models achieve good prediction results, and it can be seen from the above experimental results that the SSA-MIC-SMBO-Online ESN prediction models are more accurate than the SSA-MIC-SMBO-Offline ESN prediction models and more suitable for projecting the dynamics changes in water quality data.

The water environment is a complex system affected by many factors. The change of each water quality index is related to various factors rather than the water environment alone. For instance, the DO in the water body depends on climate factors such as water temperature, air pressure, and light. Therefore, the prediction of water quality indices must consider multiple influencing factors, as well as the correlation between water quality indices and meteorological factors/pollution sources. In the future, our research team will improve the proposed prediction models by supplementing various relevant data and further optimize the ESN algorithm from multiple perspectives: improving the reservoir generation approach, improving the initialization of connection weights of the reservoir, optimizing these weights, streamlining the reservoir topology, rationalizing neuron selection, and so on. In addition, LSTM and its improved algorithms will be considered for water quality prediction work in this study.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This paper was supported by the Key R&D Projects in Hebei Province (Grant nos. 21370103D, 21373301D, and 19273301D), the General project of Hebei Natural Science Foundation (Grant no. D2019407046), the National Key R&D Projects (Grant no. 2019YFC1407903), and the Research on Social Sciences Development in Hebei Province (Grant nos. 20210201445 and 20200302075).

References

[1] S. H. Ewaid, S. A. Abed, and S. A. Kadhum, "Predicting the Tigris River water quality within Baghdad, Iraq by using water quality index and regression analysis," *Environmental Technology & Innovation*, vol. 11, pp. 390–398, 2018.

[2] T. Chakraborty, A. K. Chakraborty, and Z. Mansoor, "A hybrid regression model for water quality prediction," *Opsearch*, vol. 56, no. 4, pp. 1167–1178, 2019.

[3] A. Sarkar, S. Mukherjee, and M. Singh, "Determination of the uranium elemental concentration in molten salt fuel using laser-induced breakdown spectroscopy with partial least squares-artificial neural network hybrid models," *Spectrochimica Acta Part B: Atomic Spectroscopy*, vol. 187, Article ID 106329, 2022.

[4] A. K. Kadam, V. M. Wagh, A. A. Muley, B. N. Umrikar, and R. N. Sankhua, "Prediction of water quality index using artificial neural network and multiple linear regression modelling approach in Shivganga River basin, India," *Modeling Earth Systems and Environment*, vol. 5, no. 3, pp. 951–962, 2019.

[5] Y. Deng, X. Zhou, J. Shen et al., "New methods based on back propagation (BP) and radial basis function (RBF) artificial neural networks (ANNs) for predicting the occurrence of haloketones in tap water," *Science of the Total Environment*, vol. 772, Article ID 145534, 2021.

[6] H. Lin, Q. Dai, L. Zheng, H. Hong, W. Deng, and F. Wu, "Radial basis function artificial neural network able to accurately predict disinfection by-product levels in tap water: taking haloacetic acids as a case study," *Chemosphere*, vol. 248, Article ID 125999, 2020.

[7] A. Najah Ahmed, F. Binti Othman, H. Abdulmohsin Afan et al., "Machine learning methods for better water quality prediction," *Journal of Hydrology*, vol. 578, Article ID 124084, 2019.

[8] R. Huang, C. Wei, B. Wang et al., "Well performance prediction based on Long Short-Term Memory (LSTM) neural network," *Journal of Petroleum Science and Engineering*, vol. 208, Article ID 109686, 2022.

[9] J. Hua, Q. K. Xiao, L. Wang, Y. Liu, and X. Ning, "Recognition of electromyographic signal time series on daily hand motions based on long short-term memory network," *Traitement du Signal*, vol. 38, no. 2, pp. 387–394, 2021.

[10] C. Kühnert, N. M. Gonuguntla, H. Krieg, D. Nowak, and J. A. Thomas, "Application of LSTM networks for water demand prediction in optimal pump control," *Water*, vol. 13, no. 5, p. 644, 2021.

[11] Y. M. Zhao, H. Yang, and G. A. Su, "Design and application of a slow feature algorithm coupling visual selectivity and multiple long short-term memory networks," *Traitement du Signal*, vol. 38, no. 5, pp. 1521–1530, 2021.

[12] J. H. Xu, J. L. Wang, and L. Chen, "Surface water quality prediction model based on graph neural network," *Journal of Zhejiang University*, vol. 55, no. 4, pp. 601–607, 2021.

[13] T. Mallick, P. Balaprakash, and E. Rask, "Transfer learning with graph neural networks for short-term highway traffic forecasting," in *Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR)*, pp. 10367–10374, IEEE, Milan Italy, January 2021.

[14] Y. Kang, J. L. Song, K. Q. Li, X. Zhai, and Y. Li, "Research on water quality prediction model based on echo state network," *Journal of Computational Methods in Science and Engineering*, vol. 22, no. 3, pp. 901–910, 2022.

[15] H. Jaeger, "The echo state approach to analysing and training recurrent neural networks," *Bonn, Germany: German National Research Center for Information Technology GMD, Technical Report*, vol. 148, no. 34, p. 13, 2001.

[16] J. Zhao, C. Zhao, F. Zhang, G. Wu, and H. Wang, "Water quality prediction in the waste water treatment process based on ridge regression echo state network," *IOP Conference Series: Materials Science and Engineering*, vol. 435, no. 1, Article ID 012025, 2018.

[17] M. Xu, M. Han, T. Qiu, and H. Lin, "Hybrid regularized echo state network for multivariate chaotic time seri- es prediction," *IEEE Transactions on Cybernetics*, vol. 49, no. 6, pp. 2305–2315, 2019.

[18] C. Yang, J. Qiao, Z. Ahmad, K. Nie, and L. Wang, "Online sequential echo state network with sparse RLS algorithm for time series prediction," *Neural Networks*, vol. 118, pp. 32–42, 2019.

[19] N. Golyandina, "Particularities and commonalities of singular spectrum analysis as a method of time series analysis and signal processing," *WIREs Computational Statistics*, vol. 12, no. 4, Article ID e1487, 2020.

[20] E. Ghaderpour, S. D. Pagiatakis, and Q. K. Hassan, "A survey on change detection and time series analysis with applications," *Applied Sciences*, vol. 11, no. 13, p. 6141, 2021.

[21] A. Martin, M. Parry, A. W. R. Soundy et al., “Improving real-time position estimation using correlated noise models,” *Sensors*, vol. 20, no. 20, p. 5913, 2020.

[22] T. Gu, J. Guo, Z. Li, and S. Mao, “Detecting associations based on the multi-variable maximum information coefficient,” *IEEE Access*, vol. 9, pp. 54912–54922, 2021.

[23] J. Runge, P. Nowack, M. Kretschmer, S. Flaxman, and D. Sejdinovic, “Detecting and quantifying causal associations in large nonlinear time series datasets,” *Science Advances*, vol. 5, no. 11, Article ID eaau4996, 2019.

[24] D. Albanese, S. Riccadonna, C. Donati, and P. Franceschi, “A practical tool for maximal information coefficient analysis,” *GigaScience*, vol. 7, no. 4, pp. 1–8, 2018.

[25] H. Bouzgou and C. A. Gueymard, “Fast short-term global solar irradiance forecasting with wrapper mutual information,” *Renewable Energy*, vol. 133, pp. 1055–1065, 2019.

[26] J. Wu, X. Y. Chen, H. Zhang et al., “Hyperparameter optimization for machine learning models based on Bayesian optimization,” *Journal of Electronic Science and Technology*, vol. 17, no. 1, pp. 26–40, 2019.

[27] Y. Zhou, “Real-time probabilistic forecasting of river water quality under data missing situation: deep learning plus post-processing techniques,” *Journal of Hydrology*, vol. 589, Article ID 125164, 2020.