Water Supply

IWA PUBLISHING

Water Supply Vol 24 No 4, 1282 doi: 10.2166/ws.2024.060

Water quality prediction using ARIMA-SSA-LSTM combination model

Tingyu Wang [a], Wei Chen [b] and Bo Tang [a,b,*]

[a] College of Metrology Measurement and Instrument, China Jiliang University, Hangzhou, China
[b] Ningbo Water Meter (Group) Co., Ltd, Ningbo, China
*Corresponding author. E-mail: tangbo@cjlu.edu.cn

TW, 0009-0001-9828-9636

ABSTRACT

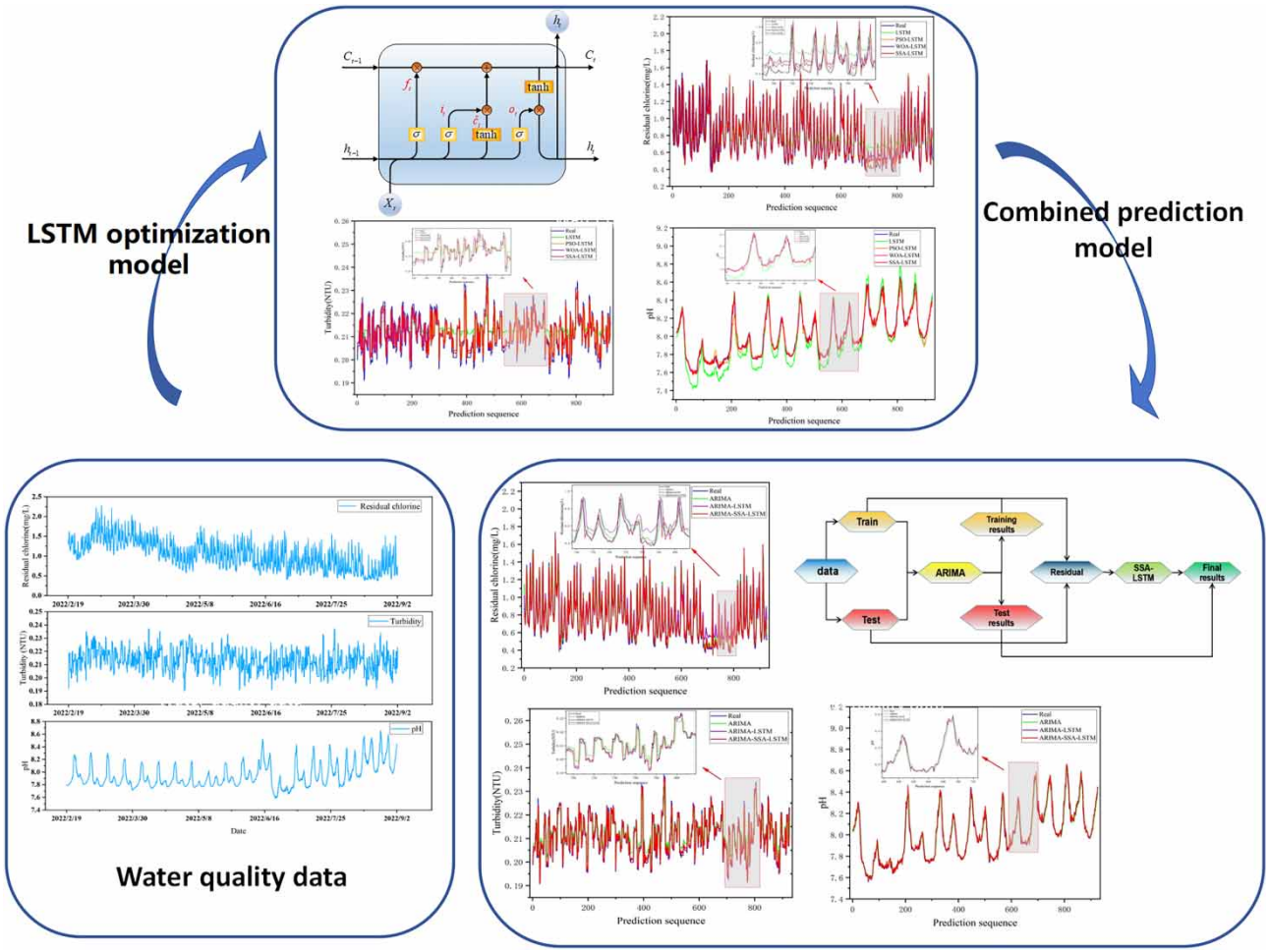
The water quality index model is a popular tool for evaluating drinking water quality. To overcome low precision and significant errors in the traditional single prediction model, a novel autoregressive integrated moving average (ARIMA)-sparrow search algorithm (SSA)-long short-term memory (LSTM) combination model is proposed to accurately predict residual chlorine, turbidity, and pH in drinking water. First, the ARIMA model is used to extract the linear part of water quality data and output the nonlinear residual. Then, the LSTM model is used to predict the residual, and the SSA is used to find the optimal hyperparameters of the LSTM model, which plays an essential role in reducing the error of the model. To prove the superiority of the model developed, the ARIMA-SSA-LSTM model is compared with SSA-LSTM, whale optimization algorithm-LSTM, PSO-LSTM, ARIMA-LSTM, ARIMA, and LSTM. The results show that the coefficient of determination (R^2) of the combination model for residual chlorine, turbidity, and pH are 0.950, 0.990, and 0.998, respectively, which are greater than all comparison models. Therefore, the model is more suitable for the prediction and analysis of water quality data.

Key words: autoregressive integrated moving average, combined model, long short-term memory, sparrow search algorithm, water quality prediction

HIGHLIGHTS

- This article presents a new combined forecasting method to predict the trend of water quality in water distribution networks.
- The autoregressive integrated moving average (ARIMA)-sparrow search algorithm (SSA)-long short-term memory (LSTM) combined model overcomes the limitations of the traditional single model.
- SSA can find more suitable hyperparameters for LSTM model, so as to improve the prediction accuracy of LSTM model.

GRAPHICAL ABSTRACT



1. INTRODUCTION

Urban water distribution networks are responsible for providing people with high-quality and pollution-free drinking water. Drinking water is prone to 'secondary pollution' when flowing through a vast and complex water distribution system, which leads to declining water quality and affects people's lives and health. Residual chlorine, turbidity, pH, and other indicators in water quality monitoring are essential for evaluating drinking water quality. Fast and accurate prediction of these indicators can enable water supply enterprises to quickly find the trend of water quality deterioration and take relevant measures. Therefore, how to scientifically predict the change in water quality data has become increasingly important.

The main methods for water quality prediction are the time series model and the deep learning model. Time series analysis is based on studying a given variable's historical characteristics. Then, a model is built according to its regularity to predict the state or value of the variable in the next period. Due to its flexibility, simplicity, and feasibility, the autoregressive integrated moving average (ARIMA) model has become the most important and widely used time series model (Chen 2007). Wang *et al.* (2019) built a general water quality prediction model by combining the Holt-Winters seasonal model with the time series ARIMA model, taking total phosphorus and total nitrogen, and the eutrophication indicators as parameters. Abdul Wahid & Arunbabu (2022) successfully predicted water quality trends in the Krishnagiri Reservoir in India using a seasonal ARIMA model by integrating *in situ* measurement and remote sensing techniques. However, the ARIMA model cannot analyze and deal with nonlinear time series.

To address this issue, a large number of nonlinear deep learning methods have been widely applied to the analysis and prediction of time series data. The most commonly used deep learning model for the analysis and prediction of time series data is a recurrent neural network (RNN) (Ong *et al.* 2014; Li *et al.* 2019). The long short-term memory (LSTM) is an enhanced RNN model that successfully solves the common problems of gradient disappearance and gradient explosion in the traditional RNN model. It has vital information capture and storage capabilities and has apparent advantages in processing time series data such as water quality data (Pascanu *et al.* 2013). Zhou *et al.* (2018) established a water quality prediction model based on LSTM using water quality features selected by the improved grey association analysis algorithm. They demonstrated the method's effectiveness on two water quality datasets: Taihu Lake and Victoria Bay. Hu *et al.* (2019) used the LSTM model to predict water quality data, such as pH and water temperature in seawater cages, and obtained relatively accurate results. This study provides a reliable tool for flood forecasting and a valuable reference for water transfer in the Three Gorges reservoir area. However, when using the LSTM model in the aforementioned studies, the selection of parameters is usually determined based on the user's experience, and this uncertainty limits the model's applicability (Xu *et al.* 2023). Therefore, it is necessary to find optimal hyperparameters of LSTM.

Currently, swarm intelligence has been widely used to adjust the parameters of neural networks. Bonabeau *et al.* (1999) define it as 'the burst collective intelligence of simple groups of agents.' The method is inspired by the natural foraging behavior of social organisms such as ants, birds, and fish. It uses information exchange and cooperation between groups to achieve optimization through simple and limited interaction between individuals. Particle swarm optimization (PSO) is the most widely used swarm intelligence algorithm. Jia *et al.* (2023) used the PSO algorithm to optimize the parameters of the LSTM model and applied the optimized model to the prediction of crop reference evapotranspiration (ET_o) in Shaanxi Province, China. The results show that the optimized model has good prediction accuracy. Wu *et al.* (2018) used the PSO algorithm to optimize back propagation (BP) neural network parameters, improving the accuracy of predicting dissolved oxygen concentration in water quality. Although the PSO algorithm has a fast convergence speed, its search space is small and cannot solve the high-dimensional problem. Due to the fixed inertia weight used in the PSO algorithm, it often fails to adjust the speed step correctly and quickly falls into the local optimal solution (Nasrollahzadeh *et al.* 2021). To solve the defects of the PSO algorithm, inspired by the hunting behavior of humpback whales, Mirjalili & Lewis (2016) proposed the whale optimization algorithm (WOA), which has an ample search space and a solid ability to jump out of local optimum, but the algorithm has slow convergence speed and low convergence accuracy. Xue & Shen (2020) proposed a new swarm intelligence optimization algorithm: a sparrow search algorithm (SSA) based on sparrow swarms' foraging and antipredation behavior. The algorithm has the advantages of fast convergence speed, high convergence accuracy, few control parameters, and a solid ability to adapt to various complex problems. Once proposed, it has attracted the attention of researchers and has become a new tool in the swarm intelligence optimization algorithms field. It has been widely used in combinatorial optimization, path optimization, image processing, data prediction, and other areas (Wu *et al.* 2021; Fan *et al.* 2023). However, until now, studies have yet to integrate ARIMA and SSA-LSTM into water quality prediction.

Since the water quality data have both linear and nonlinear characteristics, the advantages of the ARIMA model in processing linear time series, the excellent performance of the LSTM model in processing nonlinear time series, and the optimization effect of the SSA on the network structure parameters of LSTM model are considered. An ARIMA-SSA-LSTM combination model is proposed to accurately predict water quality data such as residual chlorine, turbidity, and pH. First, the ARIMA model is used to predict the water quality data of water distribution networks, and the obtained model training residuals are nonlinear time series. Then, the appropriate parameters are selected by the SSA to construct the LSTM residual prediction network to predict the residual values. Finally, the ARIMA-SSA-LSTM model prediction data of the original water quality time series data can be obtained by combining the sequence prediction value calculated by the ARIMA model with the residual prediction value obtained by the LSTM model. To prove that the model proposed has higher prediction accuracy, the water quality data of water distribution networks in a residential area in China are used as case study. The ARIMA-SSA-LSTM model is compared with other models, and the effectiveness and accuracy of the proposed model are comprehensively and systematically evaluated. Aiming at the problems of low accuracy and large error of traditional single prediction model, this article combines the ARIMA model and the SSA-LSTM algorithm into drinking water quality prediction for the first time and verifies the model with actual water quality index data, using mean absolute error (MAE). Root-mean-square error (RMSE) and coefficient of determination (R^2) were used to evaluate the prediction effect of the model. The results show that the model has the smallest error and is more suitable for the prediction and analysis of drinking water quality data.

2. METHODS

2.1. Autoregressive integrated moving average

ARIMA model (Box *et al.* 2015): Based on the autoregressive (AR) model, the difference process and moving average (MA) model are added, which combines the characteristics of the two models (Wang *et al.* 2023). AR model can solve the problem of the relationship between current data and later data, while the MA model can solve the problem of random changes in data. In the ARIMA (p,d,q) model, p is the number of AR terms, q is the number of MA terms, and d is the number of differences taken to make it a stationary time series. Equation (1) shows the ARIMA model:

$$y_t = c + \varphi_1 y_{t-1} + \ldots + \varphi_p y_{t-p} + \theta_1 e_{t-1} + \ldots + \theta_q e_{t-q} \tag{1}$$

where y_t is the water quality time series data, c is a fixed value, φ and θ are the coefficients of AR and MA models, respectively, p and q are the orders of AR and MA models, respectively, and e_t is the representation error of white noise sequence.

For the ARIMA model, it is necessary to determine whether the data are smooth, and if the data are not smooth, the data need to be differenced to make it smooth, and the difference order is d. The size selection of parameters p and q is usually determined by observing the trailing and truncated situations of the autocorrelation function (ACF) and partial ACF of time series data. Still, this method will affect the selection of parameter size due to human subjective factors. Therefore, akaike information criteria (AIC) and Bayesian information criteria (BIC) are now commonly used to determine the values of p and q. Equation (2) presents the equation of AIC, and Equation (3) calculates BIC:

$$A_{\mathrm{AIC}} = 2k - 2\ln(L) \tag{2}$$

$$B_{\mathrm{BIC}} = \ln(n)k - 2\ln(L) \tag{3}$$

where k represents the number of parameters in the ARIMA model, n represents the data length, and L represents the maximum likelihood function value of the model. The flowchart of the ARIMA model is shown in Figure 1.

2.2. Long short-term memory

LSTM is an RNN model commonly used in natural language processing and sequence modeling. The basic idea of the LSTM model is to introduce a variable called 'cell state' to maintain sequence information and introduce the input gate, forget gate, and output gate to control the flow of information (Hochreiter & Schmidhuber 1997). Compared with the RNN model, the LSTM model can deal with long sequence data better and can effectively avoid the problems of gradient disappearing and gradient explosion. Figure 2 shows the construction of LSTM.

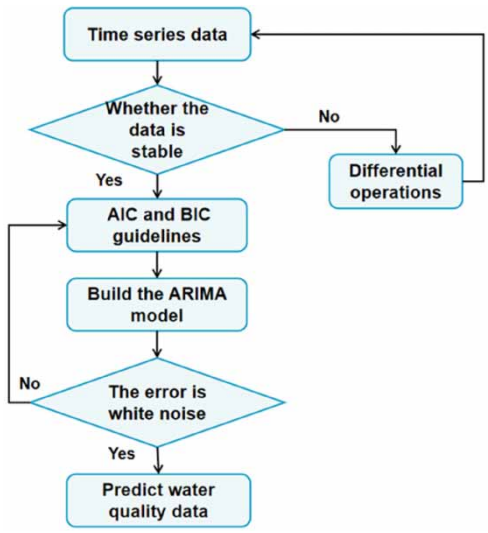


Figure 1 | ARIMA model flowchart.

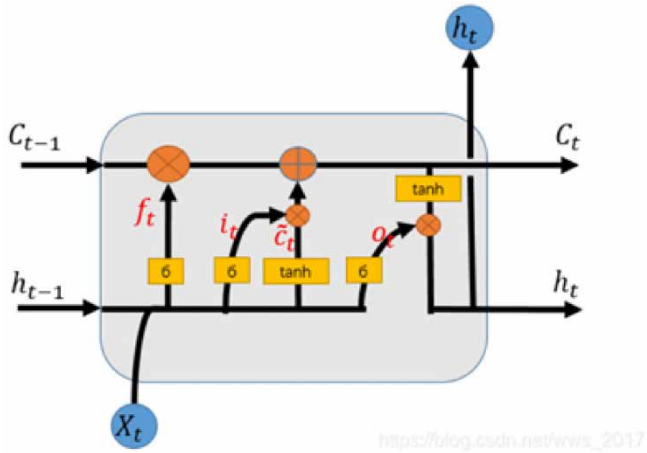


Figure 2 | LSTM structure diagram.

In the LSTM model, the input gate of each memory unit controls what to enter in the memory unit, the forget gate controls what to delete from the memory unit, the output gate controls what to output from the memory unit, and the memory unit is responsible for memorizing long sequences of information. At each time step t, each gate structure receives the input x_t at this time and the hidden state c_{t-1} output by the memory unit at the previous time step t-1. Taking unit state c_t and output h_t at the LSTM layer at t, the computation is as follows.

The forget gate is used to control the forgetting of information from the t-1 time step. The forgetting gate determines what information should be forgotten based on the current input and the hidden state of the previous moment, using Equation (4):

$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f) \tag{4}$$

where x_t is the input of the current cell, h_{t-1} is the output of the previous cell layer, W_f and b_f represent the weight coefficient matrix and bias vector of the neural network, respectively, and $\sigma(x) = 1/(1 + e^{-x})$ represents the sigmoid function.

The input gate controls the state c caused by the new input x. The update equation of the LSTM unit and the control equation of the input gate are shown in Equations (5) and (6):

$$\tilde{C}_t = \tanh(w_c[h_{t-1}, x_t] + b_c) \tag{5}$$

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

where W_c and b_c denote the input gate weight coefficient matrix and bias vector, respectively, determined by tanh, tanh denotes the hyperbolic tangent activation function, and W_i and b_i denote the input gate weight coefficient matrix and bias vector determined by σ:

$$C_t = f_t \times C_{t-1} + i_t \times \tilde{C}_t \tag{7}$$

The output gate controls the output of the hidden layer of the unit, and its control function is defined, as shown in Equation (8):

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \tag{8}$$

where W_o and b_o are the output gate's weight coefficient matrix and bias vector, respectively.

The output characteristic of the final hidden layer is shown in Equation (9):

$$h_t = o_t \times \tanh(C_t) \tag{9}$$

When the LSTM network processes the input sequence, the features of the input sequence are extracted step by step. For each time step input, the aforementioned steps are performed until all elements of the entire sequence are processed, and finally, the final output of the whole series is returned.

2.3. Sparrow search algorithm

Sparrows are birds with social living, strong memory, curiosity, and vigilance, and there is an apparent division of labor when foraging. According to the foraging and antipredation behavior of the sparrow population, researchers summarized three identities: finder, follower, and scout. The finder is responsible for finding food and grasping the foraging area and direction of the population, the follower is responsible for following the foraging, and the scout will timely warn when aware of the danger.

In the SSA, the position of each sparrow corresponds to a feasible solution to the optimization problem. n is the total number of sparrows. $X_n = [x_1{}^n, x_1{}^n, \ldots, x_d^n]$ is the location of the nth sparrow, each position element corresponds to an optimization variable, and d is the dimension of the optimization variable. The fitness value $f(X_n)$ can be used to evaluate the quality of sparrow position X_n. The fitness value reflects the advantages and disadvantages of the feasible solution corresponding to each sparrow's position in the objective optimization problem, determines the sparrow's identity attributes, and obtains different position update rules.

The fitness value determines the finder and the follower dynamically, and the ratio of the two in the total population is constant. The discoverer usually has a higher fitness value in the population and obtains a more comprehensive foraging search range than the follower. The location update rule of the discoverer is shown in Equation (10):

$$x_{i,\mathrm{j}}^{t+1} = \begin{cases} x_{i,j}^{t} \cdot \exp\left(-\dfrac{i}{a \cdot \mathrm{iter}_{\max}}\right) & R_2 < \mathrm{ST} \\ x_{i,j}^{t} + Q \cdot L & R_2 \geq \mathrm{ST} \end{cases} \tag{10}$$

where t represents the current number of iterations, $\mathrm{iter}_{\max}$ represents the maximum number of iterations, $j = 1,2,3\ldots d$, $x_{i,j}^{t}$ represents the value of the jth dimension of the ith sparrow at iteration t, a is the random number with the value (0,1], R_2 is the early warning value with the value range [0, 1], safety threshold (ST) is the safe value with the value range [0.5, 1], and Q is a matrix with element values of 1 and size $1 \times d$.

The follower constantly tracks and monitors the finder's forage and moves in the direction of the finder's position with the optimal fitness value in position iterative update so that its position can obtain a higher fitness value. If the fitness value under the follower's position is higher than that of the finder, it will replace the finder's position. If the follower is in an extreme state of hunger, that is, $i > n/2$, The follower will move to other areas to feed. The follower's position update rule is shown in Equation (11):

$$x_{i,j}^{t+1} = \begin{cases} x_{p}^{t+1} + |x_{i,j}^{t} - x_{p}^{t+1}| \cdot A^{+} \cdot L & i \leq n/2 \\ Q \cdot \exp\left(\dfrac{x_{\mathrm{worst}} - x_{i,j}^{t}}{i^2}\right) & i > n/2 \end{cases} \tag{11}$$

where x_p is the optimal position occupied by the finder, x_{worst} is the global worst position, A is a matrix of size $1 \times d$ with random element values of 1 or -1, and $A^{+} = A^{T}(AA^{T})^{-1}$.

A part of the sparrow population will be randomly selected as the scout. When the scout is aware of the danger, the sparrows in the edge area will move to the safe place, and the sparrows in the center area of the population will move randomly. The position update rule of the scout is shown in Equation (12):

$$x_{i,j}^{t+1} = \begin{cases} x_{\mathrm{best}}^{t} + \beta \cdot |x_{i,j}^{t} - x_{\mathrm{best}}^{t}| & f_i > f_b \\ x_{i,j}^{t} + K \cdot \dfrac{|x_{i,j}^{t} - x_{\mathrm{worst}}^{t}|}{(f_i - f_w) + \varepsilon} & f_i = f_b \end{cases} \tag{12}$$

where x_{best} is the globally optimal position, β is the step control variable with a normal distribution subject to zero mean unit variance, K is a random number in the range of -1 to 1, f_i is the current sparrow fitness value, f_b and f_w are the globally best and globally worst fitness values, respectively, and ε is a small enough positive constant.

To find the optimal hyperparameters of the LSTM, the following steps need to be performed.

(1) Define the hyperparameter space: The hyperparameter space is defined as a total, and each individual is a set of hyperparameter combinations, including hidden layer size, learning rate, batch size, etc.
(2) Initialize the hyperparameter space: The SSA is used to initialize the hyperparameter space and generate the initial population.
(3) Evaluate fitness: Each individual (hyperparameter combination) is evaluated using the fitness function.
(4) The SSA updates the population, selects the hyperparameter combination with the best fitness, and generates a new hyperparameter combination.
(5) Termination condition: Steps (3) and (4) are repeated until a predetermined termination condition is reached, such as a maximum number of iterations or a satisfactory combination of hyperparameters.
(6) The hyperparameter combination with optimal adaptation is taken as the optimal hyperparameter combination output of the LSTM model.

Through the aforementioned steps, you can gradually optimize the hyperparameters of LSTM to improve its performance and effect. This method can better train and adjust the LSTM model's parameters, help find the optimal combination of hyperparameters, and avoid the subjectivity of manual parameter selection.

2.4. ARIMA-SSA-LSTM

Water *quality* data consists of linear and nonlinear components. The linear part refers to the data that the water quality parameters show a linear relationship with the change of time or other factors, such as the linear increase or decrease of the parameter changes with the change of the treatment process. The nonlinear part refers to the data characteristics caused by complex nonlinear interactions or random changes among water quality parameters, which is difficult to be described by a simple linear model. As shown in Equation (13):

$$x_t = L_t + N_t \tag{13}$$

where L_t represents the linear part and N_t represents the nonlinear part.

Considering the advantages of the ARIMA model in processing linear time series and the excellent performance of the LSTM model in processing nonlinear time series (Abebe *et al.* 2022), the ARIMA-SSA-LSTM model developed uses the ARIMA model to predict the water quality data, and the obtained model training residual is a nonlinear time series. Then, the appropriate parameters are selected by the SSA to construct the LSTM residual prediction network to predict the residual values. Finally, by combining the predicted value of the ARIMA model with the residual predicted value of the LSTM model, the predicted data of the ARIMA-SSA-LSTM model of water quality time series data can be obtained. Its flowchart is shown in Figure 3.

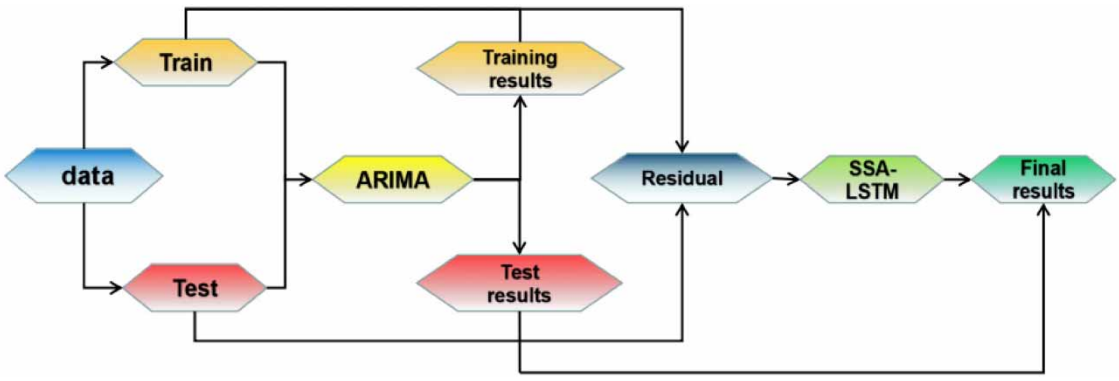


Figure 3 | ARIMA-SSA-LSTM model flowchart.

To improve the model's accuracy, the data processing efficiency, and the model's generalization ability, Equation (14) was used to normalize the water quality data. After model training and prediction are completed, the predicted results are reverse normalized using Equation (15) to obtain the actual values of the data, which are used for subsequent plotting and evaluation of the model prediction effect:

$$x_t = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{14}$$

$$x = (x_{\max} - x_{\min})x_t + x_{\min} \tag{15}$$

where x is the original data, x_t represents the normalized water quality data, $x_{\max}$ is the maximum value in the sequence, and $x_{\min}$ is the minimum value in the series.

Finally, the MAE, RMSE, and coefficient of determination (R^2) are used to evaluate the prediction effect of the combined model and other models. MAE is the sum of absolute errors between the fitted value and the actual value, which can reflect the overall fit of the model, as shown in Equation (16). When the error distribution is Gaussian, RMSE is more suitable to characterize the model performance than MAE (Chai & Draxler 2014), as shown in Equation (17). R^2 is an important statistic reflecting the model's goodness of fit and is not affected by the data scale. It is the ratio of the regression sum of squares to the total sum of squares, and its value is between [0,1]. Generally, a model with R^2 greater than 0.8 can be considered a good prediction model (Xu *et al.* 2022), as shown in Equation (18):

$$MAE = \frac{1}{N}\sum_{i=1}^{N} |\hat{x}_i - x_i| \tag{16}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (\hat{x}_i - x_i)^2} \tag{17}$$

$$R^2 = \frac{\sum_{i=1}^{N} (\hat{x}_i - \bar{x}_i)^2}{\sum_{i=1}^{N} (x_i - \bar{x}_i)^2} \tag{18}$$

where N is the data length, x_i and $\hat{x}_i$ represent the measured value and predicted value of water quality data, respectively, and $\bar{x}_i$ represents the average value of the data.

3. RESULTS AND DISCUSSION

3.1. Study area

The pH has a significant impact on human health and the environment. Too high a pH will make the water alkaline, which may lead to adverse reactions such as sore throat and gastrointestinal discomfort. Too low a pH will make the water acidic, which may hurt the human body and have a corrosive effect on metal materials such as steel. Residual chlorine in urban water distribution networks is unstable. Its concentration will decay with time. When the residual chlorine is lower than 0.05 mg/L, it cannot effectively kill bacteria, viruses, and other microorganisms, deteriorating the water quality and causing severe harm to health. Turbidity refers to the content of small particulate matter in the water, which affects the transparency and cleanliness of the water. High turbidity water will contain more sediment, minerals, and other substances, which will not only affect the beauty of the water but also affect the absorption and utilization of nutrients in the water. High turbidity water can also lead to scale in households, affecting the circulation and use of drinking water. According to the 'Technical Standards for Online Monitoring of Urban Water Supply Quality' (CJJ/T 271-2017) issued by the Ministry of Housing and Urban-Rural Development of China, the online monitoring indicators of water supply quality should include residual chlorine and turbidity, and the size of pH is closely related to residual chlorine, and these three indicators are important indicators for evaluating tap water quality. Monitoring these indicators is helpful to evaluate water quality and ensure the safety and sanitation of residents' drinking water and is a common online monitoring and evaluation index for water supply companies in China. Therefore, this study selected three water quality indicators, residual chlorine, turbidity, and pH, according to relevant standards for online monitoring.

This study uses real-time data of water quality indicators such as residual chlorine, turbidity, and pH value of residents in a community in Ningbo, China, from 19 February 2022, to 2 September 2022. Data were sampled every 2 h for a total of 2,352 data. The collected water quality data are shown in Figure 4. Sixty percent of the dataset was used as the training set and 40% as the test set. The first 118 days of data are used to predict the next 78 days of data.

3.2. Water quality prediction results based on LSTM optimization algorithm

All prediction models in the study are based on the Python language. The LSTM model consists of three hidden layers with 50, 100, and 200 neurons, respectively. The learning rate of the model is set to 0.1, the discard rate to 0.2, the batch length to 64, the number of model iterations to 100, and the time step to 15. RMSE was used as the model's loss function, and the LSTM model was trained using the Adam optimizer.

The population size was set of the SSA to 20 and the maximum number of iterations to 50. The search range of the number of neurons in the three hidden layers of the LSTM model is 10–300, the search range of the learning rate is 0.0001–0.99, and the search range of the optimal batch length is 1–300. The population size, iteration times, and search range of LSTM hyper-parameters of PSO and WOA are the same as those of the SSA to compare the optimization effect of these three optimization algorithms under the same conditions.

Figure 5 shows the residual chlorine prediction result of each optimization model. It can be seen from the figure that each optimization model has higher prediction accuracy than the single LSTM model, among which the SSA-LSTM model has higher prediction accuracy on the later data, which can ensure the prediction accuracy of long-term prediction of residual chlorine.

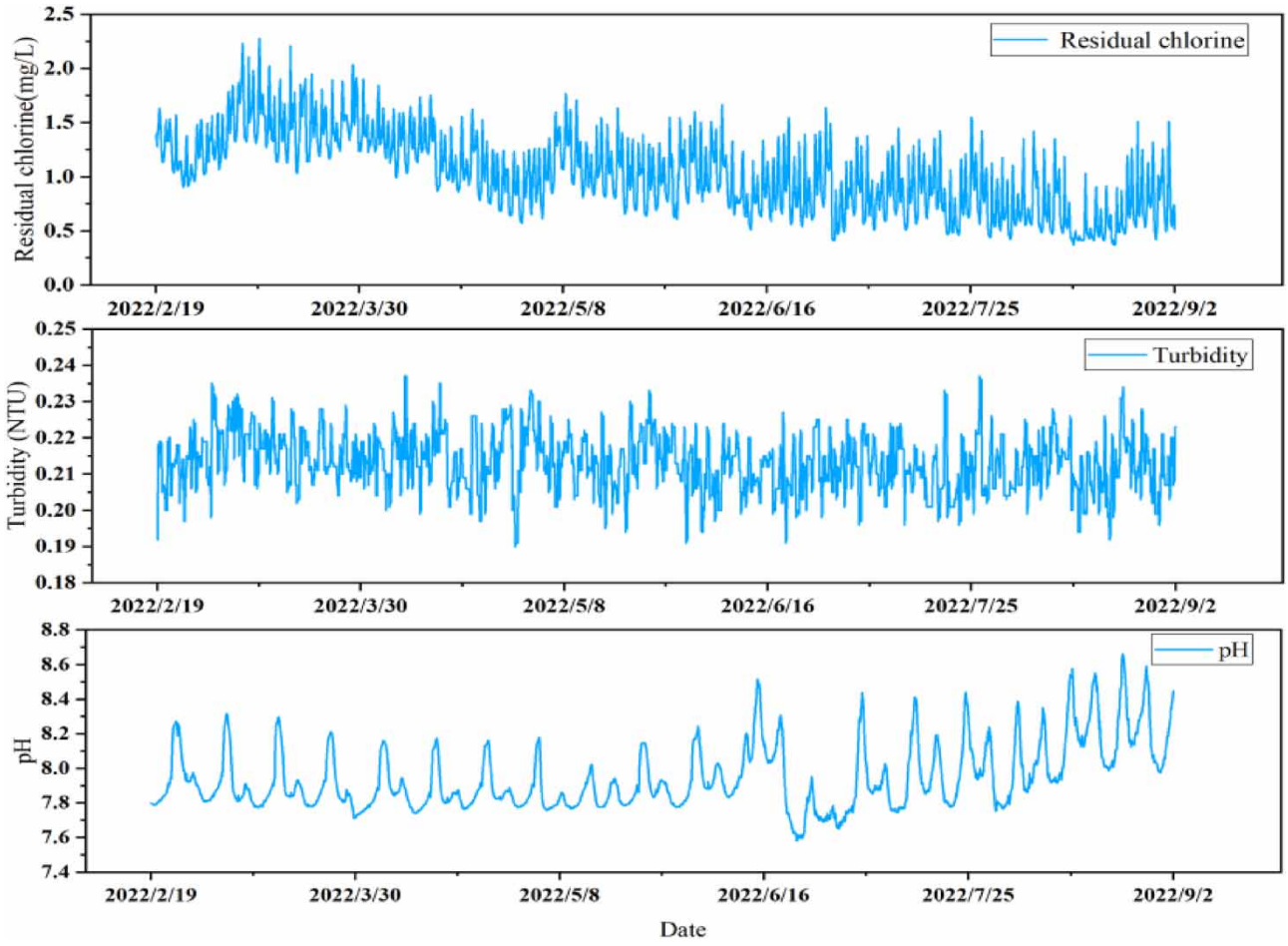


Figure 4 | Water quality data.

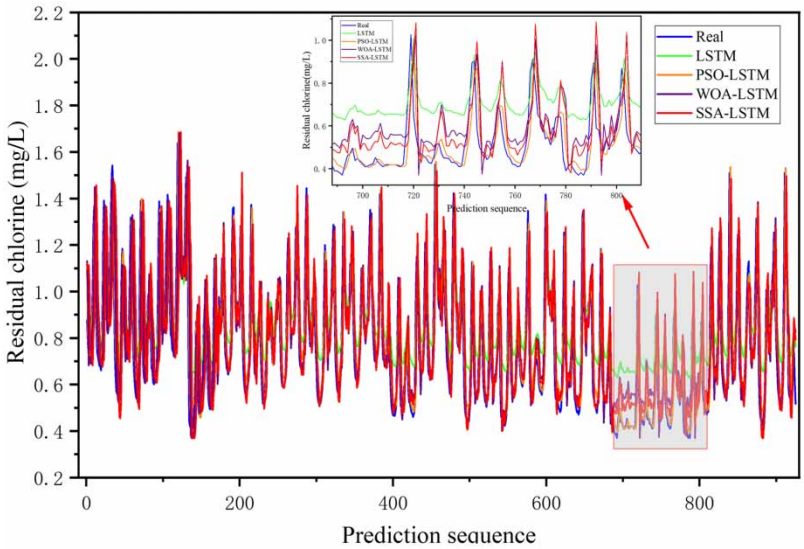


Figure 5 | Residual chlorine prediction based on each LSTM optimization model.

As shown in Table 1, compared with the LSTM, PSO-LSTM, and WOA-LSTM models, the predicted residual chlorine MAE based on the SSA-LSTM model decreased by 36.9, 15.7, and 10.3%, RMSE decreased by 34.0, 17.7, and 10.6%, and R^2 increased by 69.0, 7.7, and 4.3%, respectively.

The turbidity prediction result is shown in Figure 6. The SSA-LSTM model has a better fit on the peak value in the data, which helps to ensure prediction accuracy when the residual chlorine data changes rapidly.

As shown in Table 2, compared with the LSTM, PSO-LSTM, and WOA-LSTM models, the predicted values of MAE for turbidity based on the SSA-LSTM model decreased by 45.5, 19.9, and 14.7%, RMSE decreased by 26.6, 10.3, and 5.90%, and R^2 increased by 130.0, 2.7, and 2.0%, respectively.

The pH prediction results are shown in Figure 7, and it can be seen that all optimization algorithms gave sound predictions, among which the SSA-LSTM model has achieved good results in long-term and peak prediction.

As shown in Table 3, all optimization algorithms have achieved good results in predicting pH, and the coefficient of determination (R^2) is above 0.9, which is related to the periodic characteristics of pH data. Compared with the LSTM, PSO-LSTM, and WOA-LSTM models, the MAE of predicted pH data based on the SSA-LSTM model decreased by 79.6, 67.0, and 50.7, the RMSE decreased by 72.7, 62.0, and 48.6%, and R^2 increased by 15.1, 6.5, and 4.0%, respectively.

In summary, although there are still some inaccuracies in the prediction results of the SSA-LSTM model, the prediction has significantly exceeded that of the single LSTM model, and the prediction is also better than the optimized models such as PSO-LSTM and WOA-LSTM.

3.3. Water quality prediction results based on ARIMA-SSA-LSTM combined model

For the ARIMA model, Augmented DickeyFuller (ADF) and Kwiatkowski–Phillips–SchmidtShin (KPSS) tests are first used to judge data stationarity. Only when both ADF and KPSS tests pass prove the data are stable. Otherwise, the data are processed

Table 1 | Error analysis of residual chlorine prediction models based on LSTM, PSO-LSTM, WOA-LSTM, and SSA-LSTM models

	Evaluation index		
Models	**MAE**	**RMSE**	**R^2**
LSTM	0.111	0.141	0.516
PSO-LSTM	0.083	0.113	0.810
WOA-LSTM	0.078	0.104	0.836
SSA-LSTM	0.070	0.093	0.872

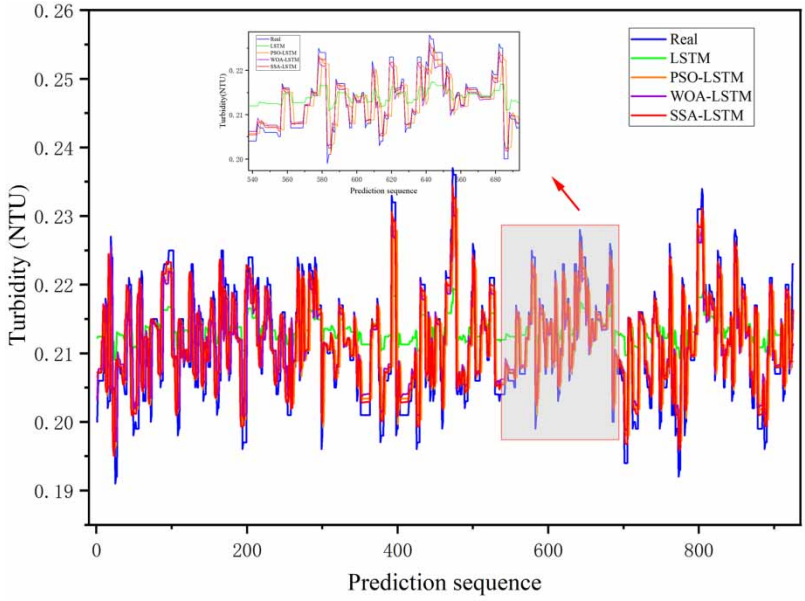


Figure 6 | Turbidity prediction based on each LSTM optimization model.

Table 2 | Error analysis of turbidity prediction models based on LSTM, PSO-LSTM, WOA-LSTM, and SSA-LSTM models

	Evaluation index		
Models	**MAE**	**RMSE**	R^2
LSTM	0.00576	0.00717	0.253
PSO-LSTM	0.00392	0.00587	0.582
WOA-LSTM	0.00368	0.00559	0.586
SSA-LSTM	0.00314	0.00526	0.598

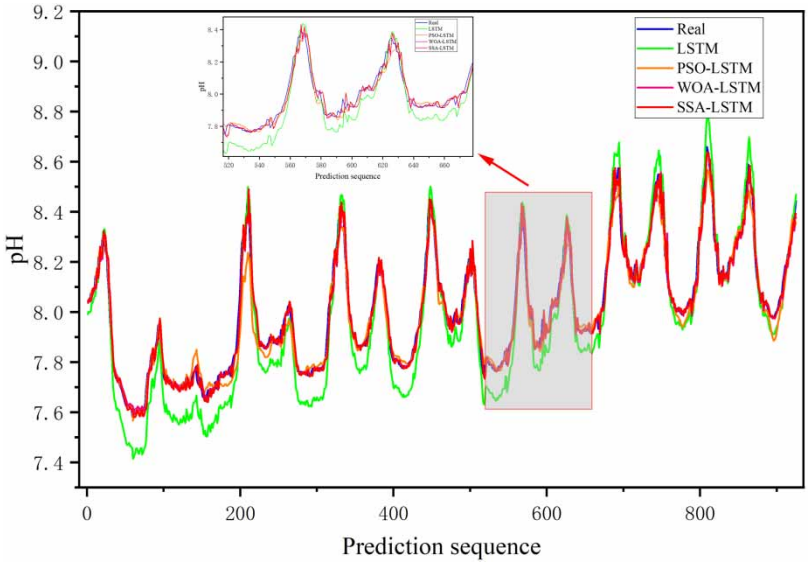


Figure 7 | pH prediction based on each LSTM optimization model.

Table 3 | Error analysis of pH prediction models based on LSTM, PSO-LSTM, WOA-LSTM, and SSA-LSTM models

	Evaluation index		
Models	**MAE**	**RMSE**	R^2
LSTM	0.0872	0.0937	0.859
PSO-LSTM	0.0539	0.0674	0.929
WOA-LSTM	0.0361	0.0498	0.951
SSA-LSTM	0.0178	0.0256	0.989

by difference until the data are stable. The value range of parameters p and q was set to [0,8], and AIC and BIC were calculated according to different combinations of p and q parameters. The parameter combination with the smallest sum of AIC and BIC was selected as the final parameter of the ARIMA model. The ARIMA model parameters of residual chlorine, turbidity, and pH were (7,1,7), (3,1,6), and (2,1,5), respectively. The parameters of ARIMA in other combination models are consistent with the parameters of a single ARIMA model.

In this research, 60% of the residual chlorine, turbidity, and pH water quality datasets were used as the training set and 40% as the test set. The results of the ARIMA-SSA-LSTM combined prediction model, ARIMA-LSTM combined prediction model, and ARIMA single model were compared. Figure 8 shows the residual chlorine prediction results of each model. The ARIMA-SSA-LSTM model gives better predictions on the residual chlorine peak value data, which can ensure good prediction accuracy when the water quality data changes.

As shown in Table 4, compared with the ARIMA and ARIMA-LSTM models, the MAE of the predicted residual chlorine based on the combined model of ARIMA-SSA-LSTM decreased by 55.9 and 42.4%, the RMSE decreased by 54.5 and 38.2%, and R^2 increased by 24.5 and 9.2%, respectively.

The turbidity prediction result of each model is shown in Figure 9. The ARIMA-SSA-LSTM model has the slightest error and has achieved good results in predicting high peak and low peak data, and the predictions are the most stable for long-term prognosis.

As shown in Table 5, compared with the ARIMA and ARIMA-LSTM models, the MAE of the predicted water turbidity based on the combined model of ARIMA-SSA-LSTM decreased by 81.6 and 48.7%, the RMSE decreased by 84.4 and 39.4%, and the R^2 increased by 69.5 and 1.5%, respectively.

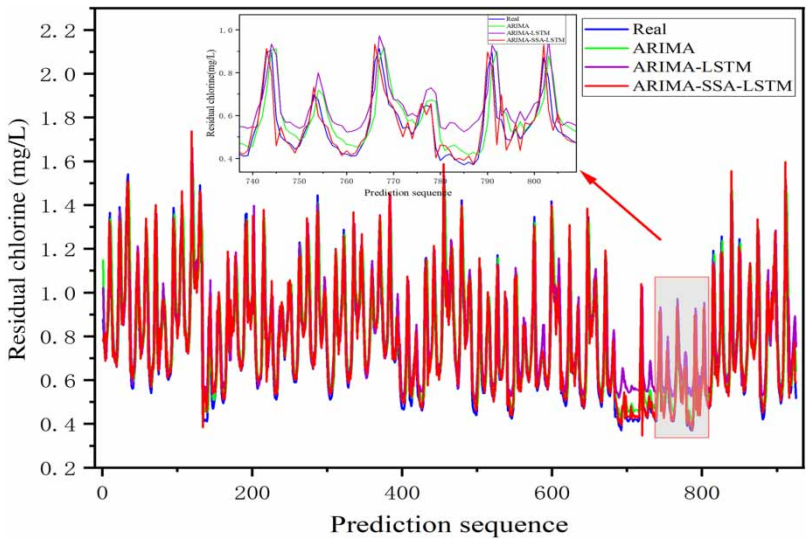


Figure 8 | Prediction of residual chlorine based on three prediction models.

Table 4 | Comparison and analysis of the errors of three residual chlorine prediction models

Models	Evaluation index		
	MAE	RMSE	R^2
ARIMA	0.0941	0.1270	0.763
ARIMA-LSTM	0.0721	0.0936	0.870
ARIMA-SSA-LSTM	0.0415	0.0578	0.950

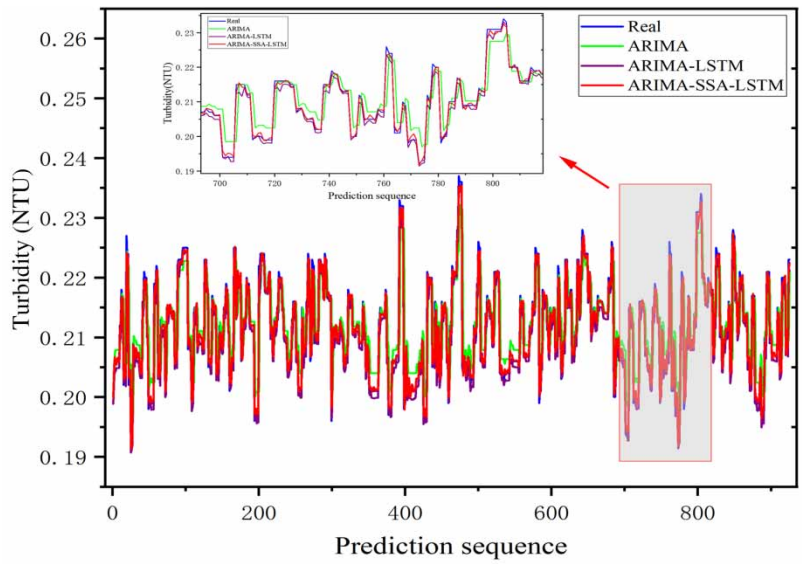


Figure 9 | Prediction of turbidity based on three prediction models.

Table 5 | Comparison and analysis of the errors of three turbidity prediction models

Models	Evaluation index		
	MAE	RMSE	R^2
ARIMA	0.00331	0.00531	0.584
ARIMA-LSTM	0.00119	0.00137	0.975
ARIMA-SSA-LSTM	0.00061	0.00083	0.990

The pH prediction result of each model is shown in Figure 10. Due to the pH periodicity, each model's prediction error is small, and the error of the ARIMA-SSA-LSTM model is the least.

As shown in Table 6, compared with the ARIMA and ARIMA-LSTM models, the MAE of the predicted pH based on the combined model of ARIMA-SSA-LSTM decreased by 70.8 and 46.2%, the RMSE decreased by 70.5 and 49.4%, and the R^2 increased by 1.4 and 0.4%, respectively.

4. CONCLUSIONS

Real-time data from a community water distribution network in China were used as an example in this research to establish the ARIMA-SSA-LSTM water quality prediction model. Comparisons are made between the advantages of SSA, WOA, and

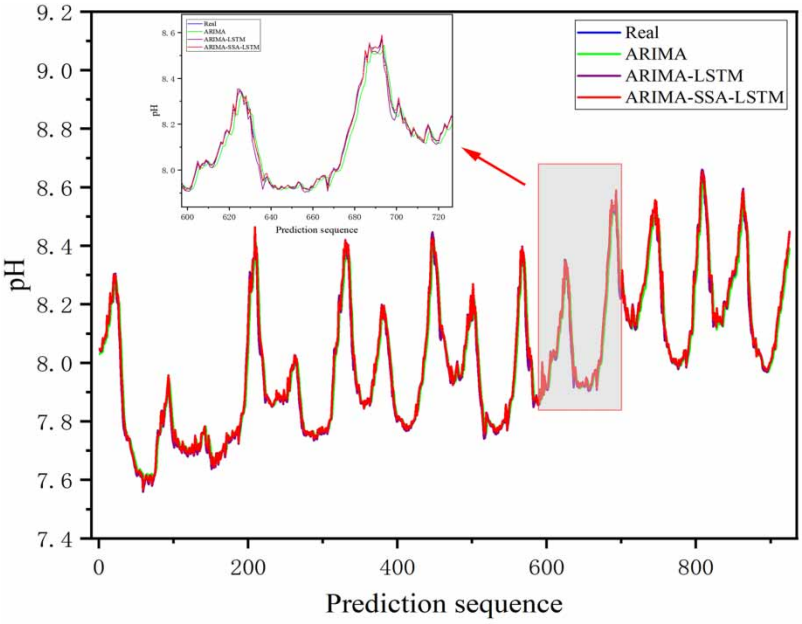


Figure 10 | Prediction curves of pH based on three prediction models.

Table 6 | Comparison and analysis of the errors of three pH prediction models

	Evaluation index		
Models	**MAE**	**RMSE**	R^2
ARIMA	0.0216	0.0295	0.984
ARIMA-LSTM	0.0117	0.0172	0.994
ARIMA-SSA-LSTM	0.0063	0.0087	0.998

PSO algorithms on LSTM model hyperparameter selection, as well as between the benefits of the ARIMA-SSA-LSTM, ARIMA-LSTM, and ARIMA for prediction. The main conclusions are as follows:

(1) The prediction accuracy of the LSTM model can be improved by selecting proper hyperparameters. Compared with WOA-LSTM and PSO-LSTM, the SSA-LSTM has higher prediction accuracy. Therefore, the SSA can find more suitable hyperparameters for the LSTM model.
(2) A new combination prediction model of residual chlorine, turbidity, and pH based on ARIMA, SSA, and LSTM, named ARIMA-SSA-LSTM, is proposed. Experimental results show that the R^2 of this model is maintained above 0.95, and the prediction accuracy is higher than all comparison models.

As a result, the approach developed in this study can provide a new perspective, and residual chlorine, turbidity, and pH can be predicted using the prediction model. When the predicted value of water quality data exceeds the range of normal indicators the state prescribes, relevant departments can take corresponding measures to deal with it.

However, there are still some shortcomings in this study: (1) The number and type of water quality indicators in the dataset are small. The selected water quality indicators, such as residual chlorine, turbidity, and pH, may not be sufficient to fully confirm whether the water quality is clean and hygienic; (2) among many excellent deep learning models, only the LSTM model was selected in this study. These two issues will be addressed in future research by expanding the collection of water quality data to include additional indicators such as *Escherichia coli*, total hardness, dissolved solids, and total

suspended solids to enrich the training set. Furthermore, we plan to explore combining linear time series models with other deep learning models for improved performance.

FUNDING

This study was supported by the Ningbo Key Research and Development Program (No. 2022Z092).

AUTHOR CONTRIBUTION

All authors contributed to the study conception and design. Tingyu Wang and Bo Tang wrote and edited the article; Wei Chen collected the preliminary data. All authors read and approved the final manuscript.

DATA AVAILABILITY STATEMENT

Data cannot be made publicly available; readers should contact the corresponding author for details.

CONFLICT OF INTEREST

The authors declare there is no conflict.

REFERENCES

Abdul Wahid, A. & Arunbabu, E. 2022 Forecasting water quality using seasonal ARIMA model by integrating in-situ measurements and remote sensing techniques in Krishnagiri reservoir, India. *Water Practice & Technology* **17** (5), 1230–1252.

Abebe, M., Noh, Y., Kang, Y. J., Seo, C., Kim, D. & Seo, J. 2022 Ship trajectory planning for collision avoidance using hybrid ARIMA-LSTM models. *Ocean Engineering* **256**, 111527.

Bonabeau, E., Dorigo, M. & Theraulaz, G. 1999 *Swarm Intelligence: From Natural to Artificial Systems (No. 1)*. Oxford University Press, Oxford.

Box, G. E., Jenkins, G. M., Reinsel, G. C. & Ljung, G. M. 2015 *Time Series Analysis: Forecasting and Control*. John Wiley & Sons, Hoboken, New Jersey.

Chai, T. & Draxler, R. R. 2014 Root mean square error (RMSE) or mean absolute error (MAE)?–Arguments against avoiding RMSE in the literature. *Geoscientific Model Development* **7** (3), 1247–1250.

Chen, S. H. 2007 Computationally intelligent agents in economics and finance. *Information Sciences* **177** (5), 1153–1168.

Fan, X., Sun, Z., Tian, E., Yin, Z. & Cao, G. 2023 Medical image contrast enhancement based on improved sparrow search algorithm. *International Journal of Imaging Systems and Technology* **33** (1), 389–402.

Hochreiter, S. & Schmidhuber, J. 1997 Long short-term memory. *Neural Computation* **9** (8), 1735–1780.

Hu, Z., Zhang, Y., Zhao, Y., Xie, M., Zhong, J., Tu, Z. & Liu, J. 2019 A water quality prediction method based on the deep LSTM network considering correlation in smart mariculture. *Sensors* **19** (6), 1420.

Jia, W., Zhang, Y., Wei, Z., Zheng, Z. & Xie, P. 2023 Daily reference evapotranspiration prediction for irrigation scheduling decisions based on the hybrid PSO-LSTM model. *PLoS One* **18** (4), e0281478.

Li, L., Jiang, P., Xu, H., Lin, G., Guo, D. & Wu, H. 2019 Water quality prediction based on recurrent neural network and improved evidence theory: A case study of Qiantang River, China. *Environmental Science and Pollution Research* **26**, 19879–19896.

Mirjalili, S. & Lewis, A. 2016 The whale optimization algorithm. *Advances in Engineering Software* **95**, 51–67.

Nasrollahzadeh, S., Maadani, M. & Pourmina, M. A. 2021 Optimal motion sensor placement in smart homes and intelligent environments using a hybrid WOA-PSO algorithm. *Journal of Reliable Intelligent Environments* **8**, 1–13.

Ong, B. T., Sugiura, K. & Zettsu, K. 2014 Dynamic pre-training of deep recurrent neural networks for predicting environmental monitoring data. In: *2014 IEEE International Conference on Big Data (Big Data)*. IEEE, Washington, pp. 760–765.

Pascanu, R., Mikolov, T. & Bengio, Y. 2013 On the difficulty of training recurrent neural networks. In: *International Conference on Machine Learning*. PMLR, Washington, pp. 1310–1318.

Wang, J., Zhang, L., Zhang, W. & Wang, X. 2019 Reliable model of reservoir water quality prediction based on improved ARIMA method. *Environmental Engineering Science* **36** (9), 1041–1048.

Wang, X., Kang, Y., Hyndman, R. J. & Li, F. 2023 Distributed ARIMA models for ultra-long time series. *International Journal of Forecasting* **39** (3), 1163–1184.

Wu, J., Li, Z., Zhu, L., Li, G., Niu, B. & Peng, F. 2018 Optimized BP neural network for dissolved oxygen prediction. *IFAC-PapersOnLine* **51** (17), 596–601.

Wu, C., Fu, X., Pei, J. & Dong, Z. 2021 A novel sparrow search algorithm for the traveling salesman problem. *IEEE Access* **9**, 153456–153471.

Xu, D., Zhang, Q., Ding, Y. & Zhang, D. 2022 Application of a hybrid ARIMA-LSTM model based on the SPEI for drought forecasting. *Environmental Science and Pollution Research* **29** (3), 4128–4144.

Xu, X., Zhai, X., Ke, A., Lin, Y., Zhang, X., Xie, Z. & Lou, Y. 2023 Prediction of leakage pressure in fractured carbonate reservoirs based on PSO-LSTM neural network. *Processes* **11** (7), 2222.

Xue, J. & Shen, B. 2020 A novel swarm intelligence optimization approach: Sparrow search algorithm. *Systems Science & Control Engineering* **8** (1), 22–34.
Zhou, J., Wang, Y., Xiao, F., Wang, Y. & Sun, L. 2018 Water quality prediction method based on IGRA and LSTM. *Water* **10** (9), 1148.

First received 30 November 2023; accepted in revised form 6 March 2024. Available online 28 March 2024