RESEARCH ARTICLE

A watershed water quality prediction model based on attention mechanism and Bi-LSTM

Qiang Zhang[1] · Ruiqi Wang[1] · Ying Qi[1] · Fei Wen[2]

Abstract

Accurate prediction of water quality contributes to the intelligent management and control of watershed ecology. Water Quality data has time series characteristics, but the existing models only focus on the forward time series when LSTM is introduced and do not consider the effect of the reverse time series on the model. Also did not take into account the different contributions of water quality sequences to the model at different moments. In order to solve this problem, this paper proposes a watershed water quality prediction model called AT-BILSTM. The model mainly contains a Bi-LSTM layer and a temporal attention layer and introduces an attention mechanism after bidirectional feature extraction of water quality time series data to highlight the data series that have a critical impact on the prediction results. The effectiveness of the method was verified with actual datasets from four monitoring stations in Lanzhou section of the Yellow River basin in China. After comparing with the reference model, the results show that the proposed model combines the bidirectional nonlinear mapping capability of Bi-LSTM and the feature weighting feature of the attention mechanism. Taking Fuhe Bridge as an example, compared with the original LSTM model, the RMSE and MAE of the model are reduced to 0.101 and 0.059, respectively, and the R2 is improved to 0.970, which has the best prediction performance among the four cross-sections and can provide a decision basis for the comprehensive water quality management and pollutant control in the basin.

Keywords Bi-LSTM · Neural networks · Attention mechanism · The time series · Water quality prediction · Hybrid model

Introduction

Water is the material basis for the survival, growth, and development of humans and other living things. Currently, with rapid industrialization and urbanization, the load on local water resources is also increasing. Many watersheds around the world are experiencing eutrophication and excessive phosphorus and manganese levels (González et al. 2008). It affects the survival of plants and animals, poses a threat to the ecosystem, and has many negative effects on the social life along the watershed. Therefore, resource conservation and pollution control in river basins have become a hot and important topic in the world (Evans et al. 2012).

Water quality safety is the focus of water pollution management in the basin, but also the key to the efficient and rational use of water resources. Previous water quality measurements are measured from a biochemical point of view, although the accuracy is high, but the detection cycle is long, for some sudden water, pollution cannot be dealt with in a timely manner (Jouanneau et al. 2014). Therefore, it is necessary to carry out the prediction of non-mechanical water quality parameters to help water environment managers keep abreast of water quality conditions and their changing trends. It can also provide early warning for the ecological health of the watershed and targeted prevention and control management of pollution point sources along the watershed.

✉ Ying Qi
   qiying@nwnu.edu.cn

   Qiang Zhang
   zhangq@nwnu.edu.cn

   Ruiqi Wang
   ruiqiwangxy@163.com

   Fei Wen
   wenfei258@126.com

[1] Department of Computer Science and Engineering, Northwest Normal University, Lanzhou, Gansu Province, China

[2] Gansu Academy of Eco-Environmental Sciences, Lanzhou, Gansu Province, China

In recent years, with the establishment of a large number of water quality monitoring sites, more and more data-driven models and methods for predicting the quality of the water environment have been proposed. In contrast to physical models, data-driven approaches build predictive models by analyzing historical data based on time series data of water quality and correlations between individual factors. Scholars have proposed many efficient algorithms and models for data-driven.

In this paper, the key parameters of water quality evaluation dissolved oxygen (DO) (Zhu et al. 2010) are selected as an example for model construction and testing. In aquatic environments, the right amount of dissolved oxygen ensures the growth of aquatic plants and animals. Currently, there are many traditional methods of water quality prediction, such as statistical models (Huang et al. 2014), multiple linear regression (MLR) (Gholizadeh et al. 2016), and autoregressive integrated moving average (ARIMA) (Faruk 2010). Huang et al. (2020) developed a multivariate adaptive regression spline model for estimating river component concentrations (MARS-EC) to guide water conservation practices and environmental decision-making. However, MARS-EC is a statistically based model that does not have a good understanding of the nonlinear relationships of water quality parameters. Jaynes (1982) presents the classic differential ARIMA, which uses the ARIMA to predict the time series. The main drawback of ARIMA is the pre-defined linear model, which must be checked for the stability of the time series data during the model identification phase. Kisi and Parmar (2016) studied the potassium permanganate index of the Yamuna River in India using LSSVR and ARIMA, and the improved model improved the prediction accuracy of SVM in most cases. However, traditional statistical methods often fail to capture many of the underlying characteristic relationships. In fact, due to the complexity of water quality data, the traditional method cannot capture the non-linearity and randomness of water quality series data very well (Liu et al. 2013).

In addition to traditional statistical and machine learning methods, more and more research in recent years has combined traditional water quality prediction algorithms with neural networks to solve the nonlinear problem of water quality time series prediction. Noori et al. (2020) have proposed a hybrid model by combining SWAT and ANN to optimize the water quality prediction model by using a complex process of water quality change. Wang et al. (2013) proposed a new ARIMA-ANN-based model that combines artificial neural nNetwork (ANN) with linear methods to obtain higher prediction accuracy. However, these hybrid models still do not fully consider the correlation in the time series of water quality data, which still affects the accuracy of prediction.

Water quality data is time series data, the water quality parameters of the basin have the characteristics of serial correlation (Hirsch et al. 1982). Specifically, there may be events in the sequence that have long intervals or delays, but have a significant impact on the value of the next moment. It is difficult for traditional neural networks to obtain critical information about such time series. Recursive neural network (RNN) is a deep learning method that stacks multiple layers of neural networks and relies on stochastic optimization to perform machine learning tasks (Cho et al. 2014). RNNs can take into account the time correlation on time series data. In theory, it can use historical information of any length and model the time series more completely. However, RNNs have gradient disappearance and gradient explosion problems in training, and they are not capable of sequence correlation. The long short-term memory network (LSTM) (Hochreiter and Schmidhuber 1997) is an improvement on the RNN, which overcomes the gradient disappearance and gradient explosion of the RNN (Pulver and Lyu 2017). LSTM adds three-door structures compared to RNNs with only one hidden state. LSTM can capture long-term dependencies from a time series, and LSTM has been successfully applied in the field of water quality prediction (Hu et al. 2019). Feng et al. (2020) use LSTM for short-term runoff forecasting and further improve the accuracy of prediction by changing the internal structure of LSTM. Ye et al. (2019) used the LSTM model to analyze the river water quality monitoring data in Shanghai, which proved the model's ability to extract information from long-span water quality data series.

However, on the one hand, the current standard LSTM and other methods in the prediction of water quality time series, only the water quality series forward processing, did not consider the reverse sequence selection to improve the prediction results, which will inevitably reduce the accuracy of the prediction. Bidirectional LSTM (Bi-LSTM) can consider both the positive and negative neighborhoods of time series data and effectively capture the sequence correlation of water quality data series by performing bidirectional nonlinear mapping of the sequence to obtain more accurate prediction results (Sun et al. 2018). Bi-LSTM is also widely used in the prediction of time series data, and Shahid et al. (2020) used Bi-LSTMs to predict confirmed cases, deaths, and recoveries in 10 major countries affected by COVID-19. The results show that, in most cases, the Bi-LSTM model performs better than other baseline models in terms of approval index. Le et al. (2019) proposed a power consumption prediction model that combines convolutional neural networks and Bi-LSTM, and the Bi-LSTM module with two Bi-LSTM layers uses the trend of time series in both positive and negative directions to predict. Zhang et al. (2020) proposed a deep learning model based on autoencoders and Bi-LSTM to predict PM2.5 concentrations, revealing the correlation between PM2.5 and multiple climate variables. Khullar and Singh (2022) proposed a deep learning-based Bi-LSTM model (DLBL-WQA) is introduced to forecast the water quality factors of Yamuna River, India.

On the other hand, LSTM's ability to pay attention to sub window features to varying degrees is insufficient, and using LSTM for prediction will lead to some important features being forgotten,

and the sequence correlation features of water quality time series data cannot be effectively utilized. In recent years, neural networks based on attention mechanisms (Ma et al. 2017) have been well used in the field of natural language processing (Hu 2019), such as machine translation (Tang et al. 2018), syntactic analysis (Brown et al. 2018), and speech recognition (Chorowski et al. 2015). This mechanism effectively highlights the impact of the key feature prediction model by assigning different weights to hidden layer elements of the neural network. Attention mechanisms have also been successfully applied to some time series forecasting studies; Li et al. (2018) extract valuable information from low-correlation factors through attention mechanisms and make stock price predictions through LSTM. Liu et al. (2020) proposed an air pollution forecasting based on attention-based LSTM neural network and ensemble learning that combines weather forecast information with atmospheric pollution drift characteristics for PM2.5 prediction. Lin et al. (2021) proposed an LSTM model for vehicle trajectory prediction with a spatiotemporal attention mechanism. Based on different vehicle and environmental factors such as target vehicle category, target vehicle location, and traffic density, the spatiotemporal attention weights learned in different highway scenarios are deeply analyzed.

On this basis, in order to better capture the time series characteristics of water quality data series and solve the problem that the LSTM model cannot highlight some key features, we constructed an AT-BILSTM model that can extract key features of water quality time series in watersheds. It aims to achieve the temporal correlation characteristics implied by the integrated adaptive learning of multivariate time series data to improve the accuracy of watershed water quality prediction. The contributions of this article include the following aspects:

1) Bi-LSTM is applied to water quality prediction tasks. Bi-LSTM can process data in different directions simultaneously through two interconnected hidden layers, while taking into account the information of the two neighborhoods on the time series and performing a bidirectional nonlinear mapping of the input sequence.

2) The time attention mechanism is introduced in the model, which can assign different attention weights to each hidden unit in the sequence and highlight the sequences that have an impact on the prediction moment, thereby improving the prediction ability of the model on the water quality time series data.

3) The model is tested with four actual site datasets in Lanzhou section of the Yellow River Basin of China, and the validity of the model is verified. Experimental results show that the proposed model combines the feature-weighted characteristics of the dual LSTM attention mechanism and the bidirectional nonlinear mapping capability and has the best prediction performance at four stations compared with the baseline method, which can provide a decision-making basis

for comprehensive water quality management and pollutant control in the watershed.

The rest of this article is organized as follows: the "AT-B2LSTM model based on AM and Bi-LSTM" section expounds the water quality prediction model proposed in this paper, and describes the key modules of the model. The "7" section is the introduction of datasets, pre-processing and evaluation indicators. The "10" section describes the experimental content and the result analysis, based on the four actual site datasets in Lanzhou section of the Yellow River basin of China, and analyzes the results in detail. Finally, the "15" section gives the conclusions of the study and the prospects for the future.

AT-BILSTM model based on AM and Bi-LSTM

The proposed AT-BILSTM model

Due to the complexity of the formation of water pollutants and the non-linear characteristics of concentration changes, water quality series prediction is difficult to achieve high accuracy requirements. Deep learning techniques can solve these problems by automatically training deep neural networks to better capture the characteristics of water quality sequences. When an RNN is used to process time series, some of the neuron's outputs can be passed to the neuron again as inputs, so it can make efficient use of previous information. However, the memory and storage capacity of RNNs are limited, and it is easy to produce gradient explosions and gradient disappearance problems. As a deep neural network on time series, LSTM can effectively capture the dependencies of the input time series data to a certain extent and prevent gradient disappearance and gradient explosion problems. However, the LSTM model has the problem of long-term dependence and cannot effectively capture the time correlation on each time step and the most important features in each time step.

Traditional LSTM models process time series sequentially without adequately considering forward and reverse data on time series. In order to obtain stronger feature extraction capabilities, Bi-LSTM is proposed to predict water quality sequences by making full use of the information of the two neighborhoods before and after. Bi-LSTM is able to simultaneously process sequence information in the anterior and posterior directions and then feed the information back to the current output layer, deriving correlation from the information of the two neighborhoods before and after, thereby enhancing prediction capabilities. In addition, considering that the hidden state of LSTM is usually represented by a vector of a certain length, over time, all information will be gradually compressed, and this unselected compression will weaken the

time difference between input features to a certain extent, resulting in some important information not be highlighted, affecting the prediction accuracy. The attention mechanism is introduced because historical information from water quality sequence data contributes differently to the prediction point at different times. The attention mechanism is a brain signal processing mechanism that simulates the human vision, borrowing from the human brain to obtain the focus or target area that wants to be focused on by quickly scanning information, investing more attention in the focus of attention, and ignoring some other useless information, which can solve the problem of insufficient attention in the time series of the model. Therefore, a hybrid model combining attention and Bi-LSTM can be applied to water quality sequence prediction to effectively capture the time correlation on each time step and the important features in each time step, thereby improving the model prediction accuracy.

We propose an improved AT-BILSTM water quality prediction model to predict future water quality sequences in watersheds. The main idea of the model is to change the LSTM network to Bi-LSTM and introduce attention mechanisms, which allows the model to process data on the sequence in both directions and weight it, effectively solving the problem of sequence correlation of the model. The model is shown in Fig. 1, there are mainly two parts, namely Bi-LSTM layer and temporal attention layer. The basic principles and implementation details of the function module Bi-LSTM layer and the temporal attention layer will be described in detail below.

Bi-LSTM layer

This module performs bidirectional nonlinear feature extraction on water quality sequence data to provide a basis for attention weight allocation in the next step. The complete Bi-LSTM module

has the same output to connect two LSTM networks with opposite timings to obtain bidirectional data information for the input sequences. The module contains a large number of LSTM cells; a single LSTM unit is shown in Fig. 2a.

In Fig. 2a, C is the cellular state of the LSTM memory, and h is the hidden layer state of the node. Each memory contains one or more cells and three gates, LSTM stores cell state information through memory cells, the gate structure is responsible for the protection and control of information, and the three gates include the forgotten gate, input gate, and output gate.

The forgotten gate determines what information we discard from the cell state. f_t determines the degree of the passage of cell state C_{t-1} at the last moment:

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

where C_{t-1} is the output of the $t-1$ moment cell, h_{t-1} is the state of the hidden layer at $t-1$ moments, σ is the *sigmoid* activation function, W_f is the input loop weight, x_t is the input value of the current moment node, and b_f is the bias item. The door reads h_{t-1} and x_t and outputs a value of f_t between 0 and 1 to determine the information to forget, 1 for "full retention" and 0 for "complete abandonment."

The input gate determines what information is added to the memory cells. The combination of "forgotten door" and "input door" enables cell status C_t updates. i_t determines what information needs to be updated, and C'_t represents what is used to update:

$$i_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{2}$$

$$C'_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \tag{3}$$

$$C_t = f_t * C_{t-1} + i_t * C'_t \tag{4}$$

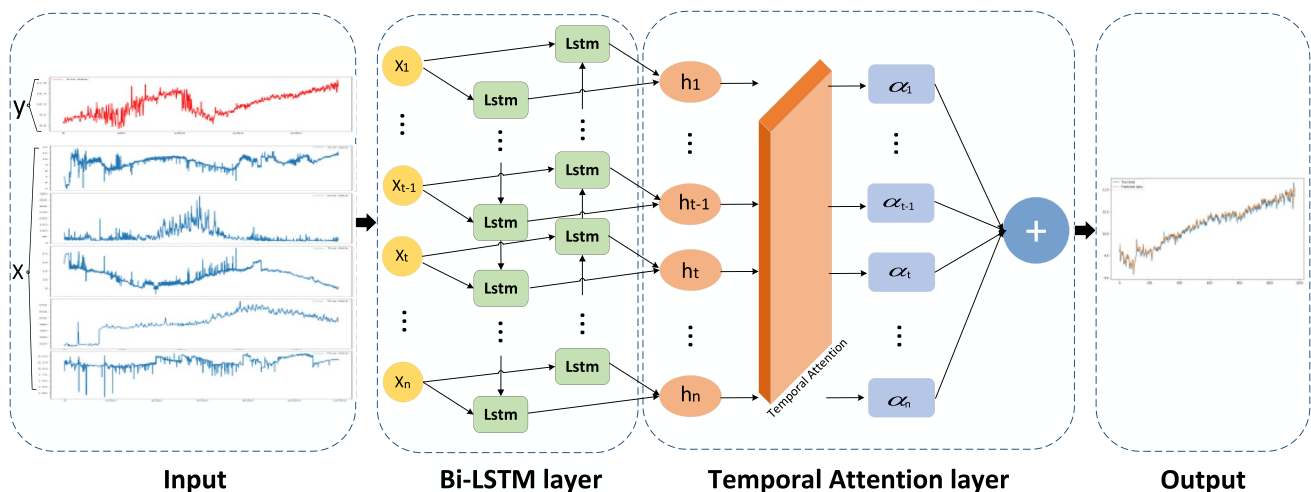where b_f and b_c are biased and W_f and W_c are the input weight.

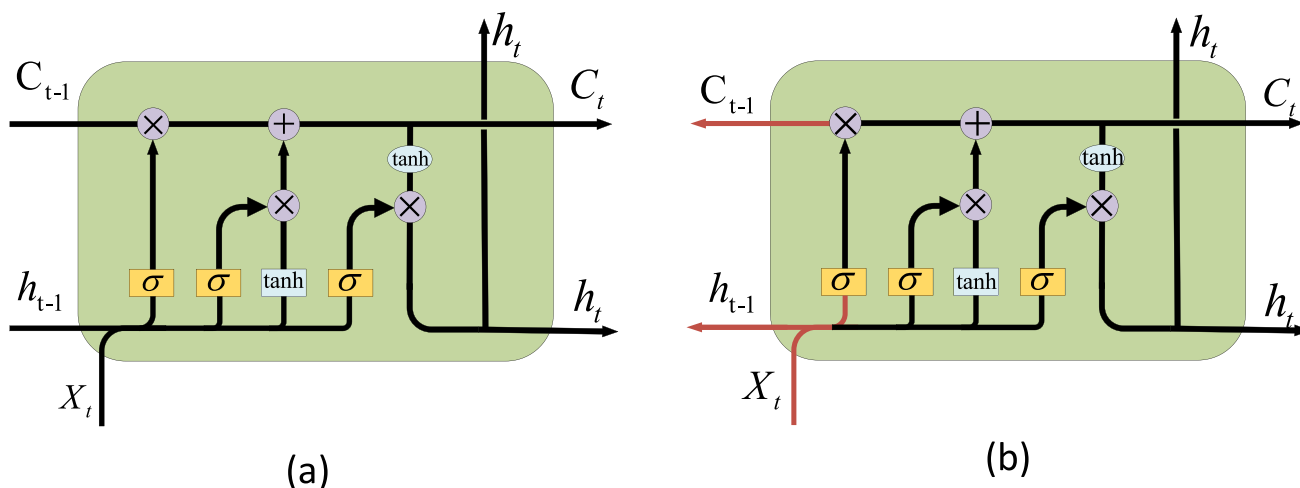**Fig. 1** The framework of the proposed AT-BILSTM model

Fig. 2 **a** Structure of LSTM. **b** Bi-LSTM single layer structure

The output gate controls the output of the cell state value, and after processing the cell state with the *tanh* activation function, the output information is multiplied by the memory unit state value. o_t is the output value; h_t is the t moment hidden layer status value:

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{5}$$

$$h_t = o_t * \tanh(C_t) \tag{6}$$

where W_o is the input weight and b_o is the biases.

These gates of LSTM effectively capture long-term and short-term dependencies on input time series data and prevent gradient disappearance and gradient explosion problems. The key to LSTM's long-term memory is that all information before each cell can be forgotten, updated, and stored in a hidden layer and exported to the next cell. The Bi-LSTM module has two LSTM networks with opposite timings that process sequence information in the front and rear directions and then feed the information back to the current output layer. Bi-LSTM's hidden state in time t includes forward and reverse, as shown below:

$$\overrightarrow{h_t} = \overrightarrow{LSTM}(h_{t-1}, x_t, c_{t-1}), t \in [1, T] \tag{7}$$

$$\overleftarrow{h_t} = \overleftarrow{LSTM}(h_{t+1}, x_t, c_{t+1}), t \in [T, 1] \tag{8}$$

$$H_t = [\overrightarrow{h_t}, \overleftarrow{h_t}] \tag{9}$$

A Bi-LSTM unit such as Fig. 2b is shown.

The unfolding structure of the Bi-LSTM layer such as Fig. 3 shown.

Because the hidden state of the LSTM unit is usually represented by a vector of a certain length, the information contained in the water quality data is gradually forgotten over time. This kind of indiscriminate forgetting will weaken the time difference between the input features to a certain extent, resulting in some important information that cannot be highlighted, affecting the accuracy of the prediction. Therefore, the attention layer is set up, and the data processed by the implicit layer is output to the time attention layer for weight calculation to make appropriate improvements to the recognition ability of Bi-LSTM.

Temporal attention layer

Attentional mechanisms originate from the simulation of the attentional features of the human brain. The attention mechanism filters useful information by assigning a higher weight to the important information in the input sequence features, thereby reasonably changing the external attention to the information, finding more important influences, and improving the accuracy of data processing.

The model sets up an attention layer that assigns the weights of the features learned by the model to the output vectors learned by the implicit layer, highlighting the impact of key features on water quality prediction. The essence of the attention mechanism is shown in Fig. 4a. We can conceive the data in the input data source as a series of $< key, value >$ key pair composition. At this point, a given target element Query, the value of the weight coefficient corresponding to each key is obtained by computing the similarity or correlation between Query and each key, and then the value is weighted to obtain the final attention value. The attention mechanism is a weighted sum of the values of the value of the elements in the source, and Query and Key are used to calculate the weight coefficients for the corresponding value. The calculation is shown below.

$$Att = \sum_{i=1}^{L_x} Similarity(Query, Key_i) * Value_i \tag{10}$$
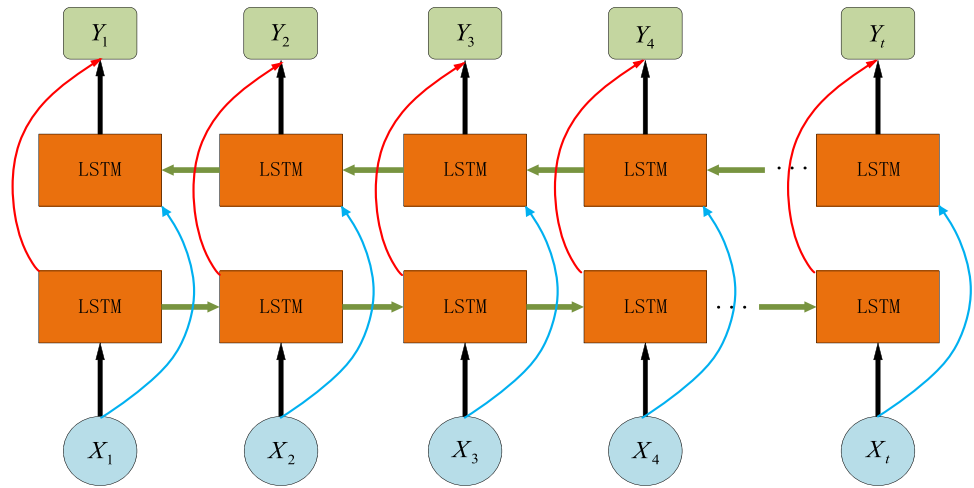
Fig. 3 Unfolded Bi-LSTM network

Figure 4b is the three stages of attention calculation, where $F(Q, K)$ is the attention value association function and s_i and $a_i(\mathrm{i} = 1, 2, \ldots, n)$ are the correlation degree and attention weight of the ith element of the input dataset. Figure 4c shows the basic attention model structure.

The traditional neural network water quality prediction model ignores that the contribution of each input feature to load prediction is different. From Fig. 4c, it can be seen that by introducing the attention mechanism in the neural network, the attention weight of the input feature X_i can be calculated, and the corresponding input features can be weighted, and the weighted features replace the original input with the input of neural networks. The implementation of the attention mechanism can be expressed as follows:
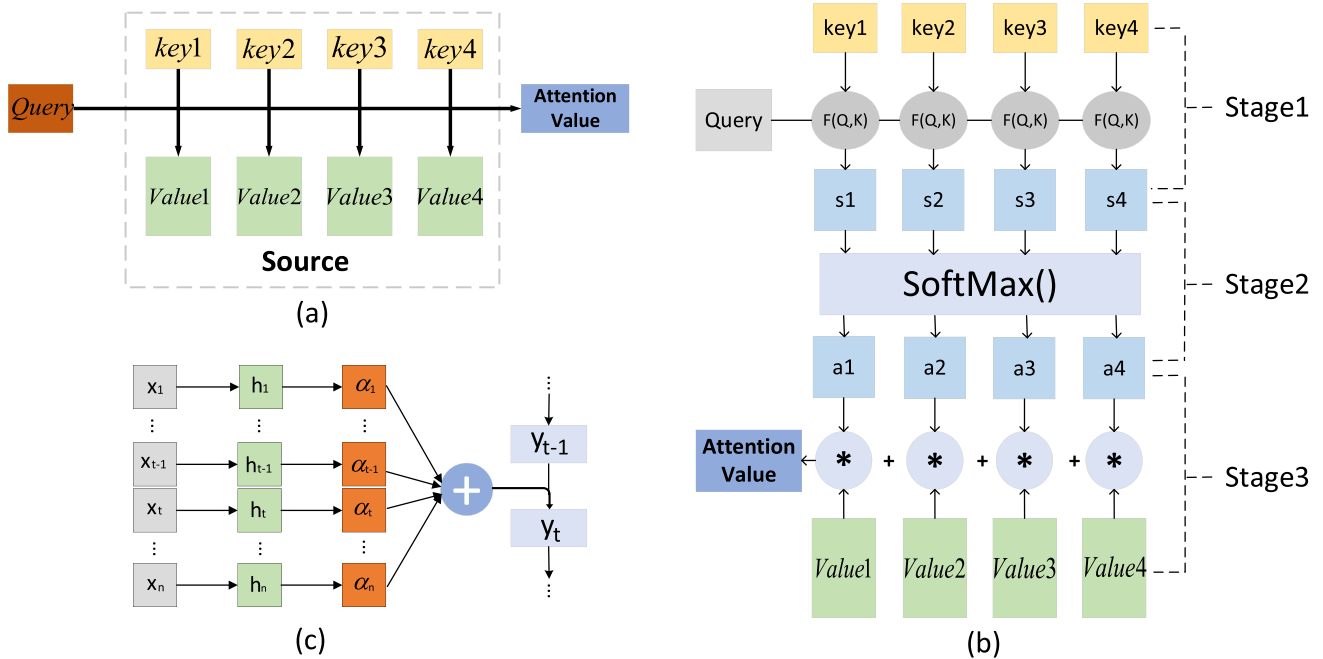
Fig. 4 Attention mechanism. **a** The core idea of attention mechanism. **b** Three stages of attention value calculation. **c** Structure of attention model

$$S_{ki} = v \tanh(Wh_k + Uh_i + b) \tag{11}$$

$$a_{ki} = \frac{\exp(S_{ki})}{\sum_{j=i}^{T_x} \exp(S_{kj})} \tag{12}$$

$$C = \sum_{i=1}^{T_x} a_{ki}h_i \tag{13}$$

The characteristic vector value of the final output is expressed as follows:

$$h_{k'} = H(C, h_k, x_k) \tag{14}$$

where $h_1, h_2, \cdots, h_k$ is the hidden layer state corresponding to the input sequence x_k, α_{ki} is the attention weight assigned to the current input by the hidden layer state feature h_k, and h_k' is the feature vector value of the final output.

This layer adds an attention mechanism to enable the model to learn to pay different attentions to the input characteristics of the time series data and effectively utilize the sequence correlation of the water quality time series data.

The training process for the model

In this study, the water quality data were divided into training sets, validation sets, and test sets. The predictive model is first trained with training and validation sets, and then the predictive model is evaluated using the test set. The key steps in the training process are described below:

Step1. Input: The correlation analysis of the predicted water environmental quality pollutants is performed, and the characteristics with the strongest correlation with the pollutants to be predicted are selected for input to improve the model accuracy.

Step2. Feature learning: The input vector enters the Bi-LSTM hidden layer from the input layer, which also includes the transfer of the same layer LSTM hidden state, mapping x_t to h_t in the LSTM hidden layer, where f is the nonlinear activation function LSTM, and the conversion formula is as follows:

$$h_t = f(h_{t-1}, x_t) \tag{15}$$

Step3. Add attention weight: The attention mechanism is designed to calculate the weight of the result at different moments according to the hidden state h_{t-1} and cell state C_{t-1} obtained in the Bi-LSTM hidden layer, and the calculation of the weight is seen (7), formula (8), and the resulting β_t^k is an attention weight, which contains the attention weight of k feature sequence.

Step4. Output: With the design of a full connection layer, we can get the prediction model output of t moment. The weighted input feature sequence is followed by z_t. The final output of the model is obtained. The conversion formula is as follows:

$$z_t = (\beta_t^1 x_t^1, \beta_t^2 x_t^2, ..., \beta_t^n x_t^n) \tag{16}$$

The corresponding algorithm for the training process of the AT-BILSTM model is described above in Algorithm 1:

Algorithm 1. The training process of AT-BILSTM

Input: Water quality sequences: $\{x_1, x_2, ..., x_T\}$
Output: Predicted sequences: $\{\hat{y}_1, \hat{y}_2, ..., \hat{y}_\tau\}$
1: Initialize window size(m), learning rate(lr), number of input steps(T)
2: **for** ξ=1 to T **do**
3:    Obtain the time series $\overline{X}_\xi$ with (1);
4: **end for**
5: **for** each epoch **do**
6:    **for** each batch **do**
7:      Bidirectional LSTM model forward pass;
8:      Bidirectional LSTM model backward pass;
9:      Generate h_t and C_t via Bi-LSTM model in (7)-(8) with the input $\overline{X}_t$ and h_{t-1};
10:   **end for**
11:   **for** t=1 to T **do**
12:     Get h_t form (5)-(11);
13:     α_t= Attention Mechanism (h_t);
14:   **end for**
15: **end for**
16: Generate $\hat{y}_t$ via fully connected with the input α_t;

Dataset description and pre-processing

In order to prove the effectiveness of the proposed AT-BILSTM model in watershed water quality prediction. In this section, the water quality sequence data of Fuhe Bridge, Xincheng Bridge, Shichuan Bridge, and Qingcheng Bridge in the Lanzhou section of the Yellow River Basin of China will be used for data preprocessing and dataset construction. The four sections cover the entire section of the Lanzhou section of the Yellow River Basin from west to east and receive a large number of farmlands, animal husbandry, rural and urban production and domestic pollution, and wastewater along the line, and their water quality characteristics can effectively represent the quality of the water environment in each section.

The water quality characteristics of the section include five characteristics of the section from June 2018 to December 2019, including hourly water temperature (T), pH, dissolved oxygen (DO), electrical conductance (EC), and turbidity (NTU) data, with 39,026 characteristics. The geographical location of the research area can be found in Fig. 5.

Data pre-processing

The experimental data is directly derived from the measured data collected by the field sensors in the four sections of the Lanzhou section of the Yellow River Basin, so it will be affected by the measurement environmental factors and measuring instruments, resulting in the existence of some vacant data and abnormal data. In order to ensure the scientific nature of the experiment and the accuracy of the model, it is necessary to preprocess the original data and then use the processed data for experimental simulation.

Missing value processing: To ensure the continuity of the dataset, enhance the reliability of the model output. Using the K nearest neighbor method, the closest K samples to the missing data sample are determined based on Euclidean distance, and the weighted average of the K sample values is used to calculate the missing sample data.

Outlier detection: In order to reduce the model prediction error and enhance the prediction accuracy, for the abnormal water quality data, the outlier value is detected by the isolated forest algorithm, the location of the abnormal score of each test data is calculated according to the formula of abnormal score calculation, and the outlier is replaced by the K nearest neighbor method.
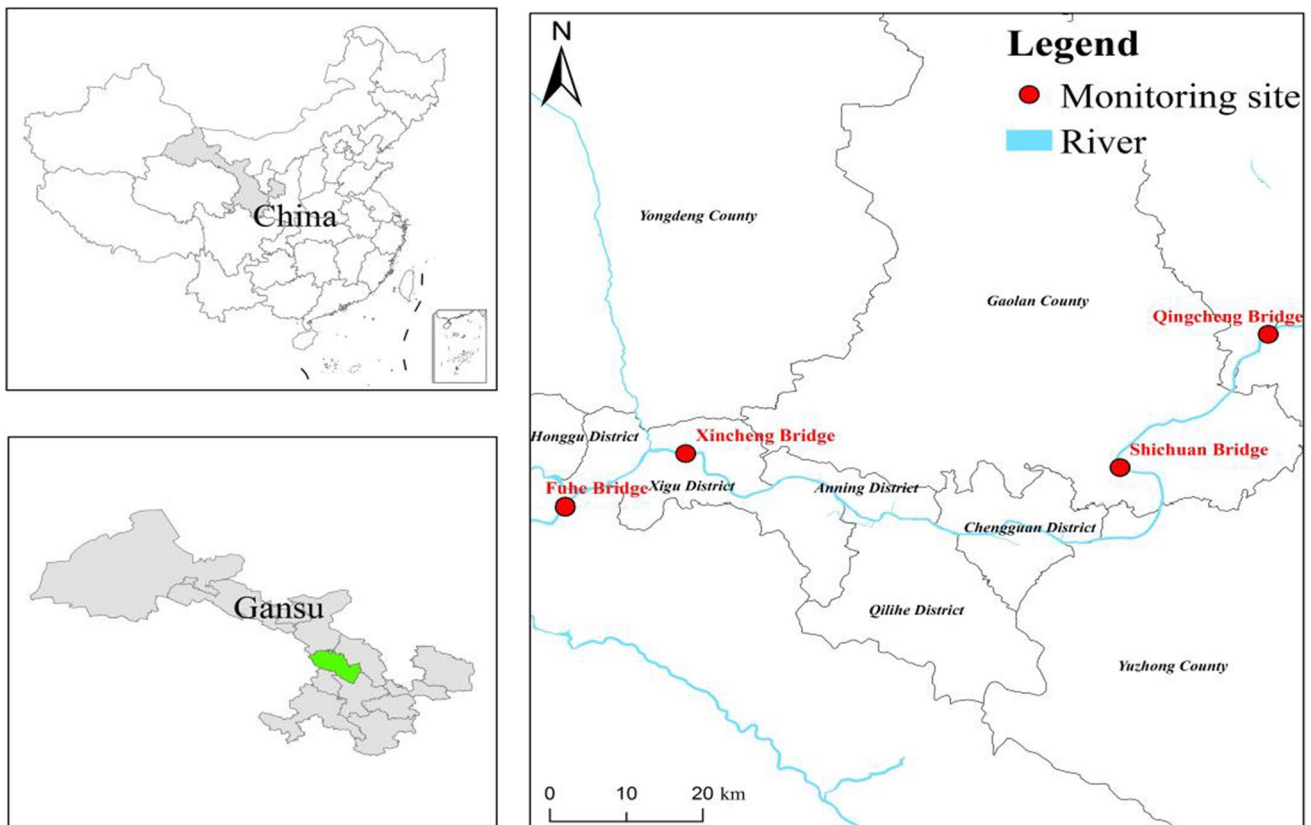

Fig. 5 Layout of water quality monitoring stations in Lanzhou City

Data standardization: In order to eliminate the effect of unit and scale differences between input features and improve convergence speed and accuracy, Max–Min is used to normalize the original water quality feature data so that the data can be mapped to the interval of [0,1].

Evaluation indicators

In order to prove the prediction performance of the AT-BIL-STM model proposed in this paper, the mean square root error (RMSE), the average absolute error (MAE), and the coefficient of determination (R^2) are selected as the evaluation index.

- RMSE: it can be used to measure the deviation between the predicted value and the true value, and the smaller the value, the more accurate the result.
- MAE: it is a loss function for regression models that are calculated as the sum of the absolute values of the difference between the target and the predicted values. Less value of MAE represents a better prediction.
- R^2: it represents fit optimization, reflecting the correlation between two random variables, and the closer to 1 indicates the better the fit.

Experiments

The water quality data of Lanzhou section of the Yellow River basin were selected as experimental data for this experiment, and the data contained five characteristics of hourly water temperature, pH, dissolved oxygen, electrical conductance, and turbidity of four of these sections from December 2017 to December 2019, with a data volume of 39,026 groups. And according to the needs of the experiment, we divide each dataset into training sets, validation sets, and test sets according to 8:1:1. In this section, empirical studies of water quality data from four sections will be conducted to demonstrate the feasibility and accuracy of the constructed water quality prediction model. Two experiments were conducted, first, the correlation detection between features was carried out by Pearson correlation analysis, the influence of different features on the model was tried, and the feature selection of the final model was determined. Then, the advantages and disadvantages of the AT-BILSTM model in each section are compared with other reference models to prove the advanced nature of the model.

Experimental environment

In this paper, the deep learning framework TensorFlow is used to build an experimental environment with the

following environmental parameters: CPU, Intel i7-6700 3.4 GHz; GPU, Nvidia GTX 1060; 8G PC memory; Win10 64-bit Operating System; python 3.6.

The effect of input characteristics on the model

Datasets contain more data features, and different features affect the predicted value of dissolved oxygen to varying degrees. To demonstrate the impact of input features on the water quality prediction model, feature extraction and identification is performed to select the most relevant information that can help with accurate predictions of the model. At this stage, we use Pearson correlation analysis to determine the extent to which features affect each other. The heat map after the calculation is such in Fig. 6 shown. The Pearson correlation coefficient refers to the measure of linear correlation between two random variables and is used to reflect the linear correlation between two continuous variables, as follows:

$$r_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2} \sqrt{\sum (y_i - \bar{y})^2}} \qquad (17)$$

where r_{xy} is valued from -1 to 1, where 1 indicates positive correlation, -1 indicates a negative correlation, and 0 indicates irrelevance.

From the above correlation analysis, we can get the correlation with dissolved oxygen (DO) to be predicted from high to low in order of temperature (T), pH, electrical conductance (EC), and turbidity (NTU). The correlation of turbidity is low, and it is discarded in feature selection. The correlation is then added from high to low as the model input. Model 1 uses only hourly dissolved oxygen information
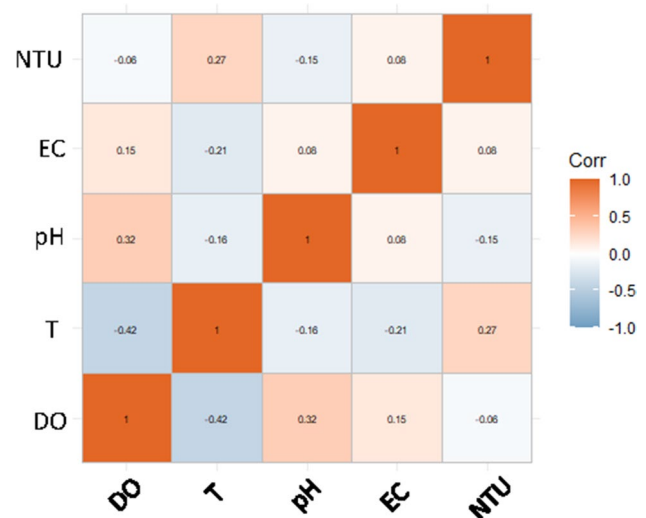

Fig. 6 Pearson correlation test results

Table 1 Pearson experimental results

Evaluation standard	MAE	RMSE
Model1	0.441	0.817
Model2	0.395	0.554
Model3	0.419	0.632
Model4	0.332	0.374

for input; model 2, model 3, and model 4 add temperature, pH, and electrical conductance characteristics as models on the basis of the previous model. Experiments compared the effect of dissolved oxygen prediction with different characteristics. Due to the different input characteristics of these four models, we did not add attention mechanism and Bi-LSTM to these four models and used LSTM neural networks to test these four models uniformly. The experimental results are shown in Table 1.

As can be seen from the table above, model 4 introduces three factors, temperature, pH, electrical conductance; there is a certain improvement, the effect is best, so we decided to use these four characteristics as the input characteristics of AT-BILSTM.

Baseline model

We compare the proposed AT-BILSTM with the following models, each of which looks like this:

- RNN: A type of recurrent neural network for processing sequence data that establishes a weighted connection between neurons between layers, and the model includes an input layer, a hidden layer, and an output layer.
- GRU: This is a typical variant of RNN that includes an update gate and a reset gate.
- LSTM: Long short-term memory network, a typical variant of RNN, consists of three gates: the forgetting gate, the input gate, and the output gate.

- Bi-LSTM: Bidirectional LSTM, which combines forward LSTM and backward LSTM for prediction.
- AT-LSTM: The model adds a time attention module to the original LSTM model, and the main structure and parameters of the model are the same as those of the LSTM model.

Experimental parameter settings

Before performing a final experimental evaluation, calibrate some important hyperparameters in the model using validation sets. We assume that the Learning Rate, Epoch, and Hidden Dimension form the core hyperparameter space of the model. Get a robust solution by searching for hyperparameter space.

Figure 7a–c shows the relationship between the RMSE of the three core models on the validation set and the learning rate, number of training, and hidden layer neurons. Based on the experimental results of the validation set, we trained all models with 60 epochs using the Adam optimizer to guarantee convergence and efficiency. Adam weight decay is set to $1e-4$, and random seed is fixed to 7. The initial learning rate is $5e-2$ with a decay rate of 0.7 after every 20 epochs, and the batch size is 200. Hidden layer neurons are set to 64. This article also uses Dropout technology to prevent overfitting and improve the performance of the model. The main parameters of the model are set as follows: Learning Rate 0.05, Epoch 60, and Hidden Dimension 64.

Results and discussions

In this section, we test the calibrated model on a test set and summarize a large number of experimental results. These results are used to analyze the predictive performance of the baseline model and the AT-BILSTM model we proposed in terms of water quality prediction. The necessity of the model attention layer and Bi-LSTM layer was confirmed by ablation experiments. And explore the benefits of combining the
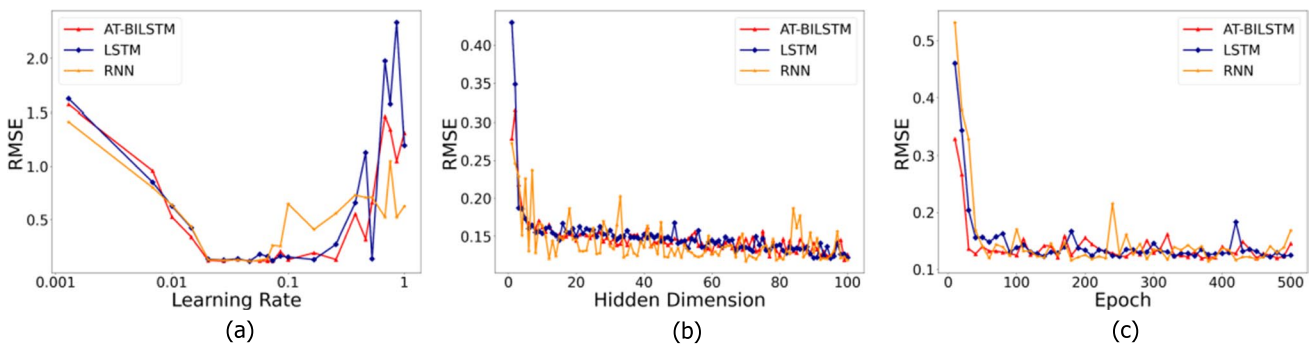

Fig. 7 Results of exploring experiments

time attention mechanism and Bi-LSTM, and finally visualize the change of time attention module weights.

Model comparison and analysis

To verify the superiority of the AT-BILSTM model proposed in this paper in terms of water quality prediction accuracy, the RNN, GRU, LSTM, Bi-LSTM, and AT-LSTM models were selected as the baseline model for comparative experiments. Compared with the AT-BILSTM model proposed in this paper, the difference between the traditional Bi-LSTM and AT-LSTM model is that the Bi-LSTM or attention layer is adopted, and the other main structures and parameters are the same as the LSTM model.

Figure 8 shows a comparison of baseline models and the AT-BILSTM model when making predictions on the Fuhe Bridge dataset. The RMSE results are displayed in the upper left of each subgraph. For all water quality prediction methods, the predicted values represented by discrete points are around a straight line representing the actual values. It can be seen that the performance of our AT-BILSTM model in

water quality prediction on the Fuhe Bridge dataset is better than that of baseline models such as RNN, GRU, and LSTM. Among the six water quality prediction methods, the RNN method was the least effective, with an RMSE of 0.1872. Whether it is the standard Bi-LSTM model or the LSTM model, the RMSE achieves a reduced effect after the introduction of attention mechanisms. This is because the attention layer in the model pays different attention to the input characteristics of different moments in the time series, highlighting the factors that have a greater impact on the prediction result and improving the prediction accuracy. The same application of the Bi-LSTM layer to the traditional model of AT-LSTM and LSTM has also improved RMSE. This is because Bi-LSTM is able to perform bidirectional learning of water quality time series data during model pretraining, greatly increasing the overall learning of the data. The model proposed in this paper introduces both Bi-LSTM and attention mechanisms into the standard LSTM model. Compared with the standard LSTM model, the water quality prediction model proposed in this paper has a 22.2%
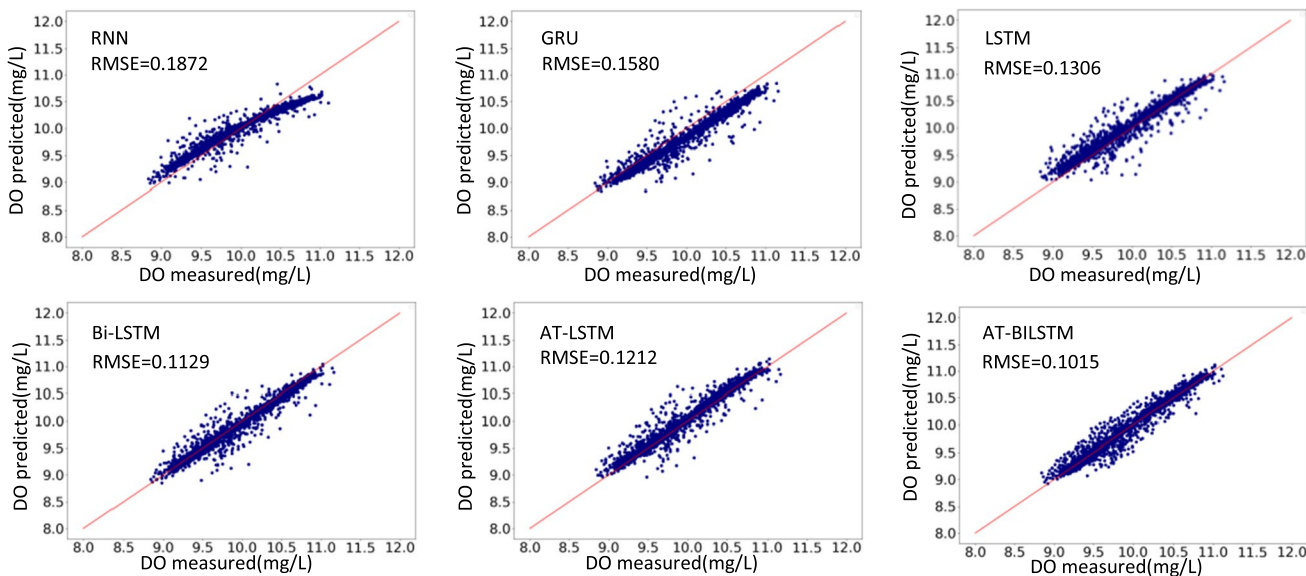

Fig. 8 Comparisons of predicted DO value with actual DO value

Table 2 Performance comparison of water quality prediction models at four stations

Models	Fuhe			Xincheng			Shichuan			Qingcheng		
	RMSE	MAE	R^2	RMSE	MAE	R^2	RMSE	MAE	R^2	RMSE	MAE	R^2
RNN	0.187	0.142	0.912	0.237	0.152	0.912	0.327	0.213	0.892	0.232	0.174	0.821
GRU	0.158	0.124	0.943	0.224	0.151	0.923	0.152	0.138	0.901	0.201	0.143	0.894
LSTM	0.130	0.133	0.951	0.193	0.160	0.921	0.145	0.161	0.893	0.173	0.143	0.903
Bi-LSTM	0.112	0.119	0.952	0.164	0.135	0.930	0.140	0.109	0.920	0.151	0.151	0.908
AT-LSTM	0.121	0.093	0.951	0.170	0.104	0.926	0.124	0.124	0.917	0.150	0.136	0.914
AT-BILSTM	0.101	0.059	0.970	0.156	0.117	0.944	0.107	0.057	0.931	0.134	0.117	0.926

reduction in RMSE on the Fuhe Bridge dataset, which is the minimum of the six models.

To further verify the feasibility of the AT-BILSTM model proposed in this paper, the AT-BILSTM model and the baseline models are applied to the Fuhe Bridge, Xincheng Bridge, Shichuan Bridge, and Qingcheng Bridge datasets for test verification. Table 2 shows how the models perform in the water quality prediction task for four sections. As can be seen from Table 2, the AT-BILSTM water quality prediction model proposed in this paper is better than the other five baseline models in terms of RMSE, MAE, and R^2. Taking the Fuhe Bridge dataset as an example, the RMSE and MAE of the AT-BILSTM model were reduced to 0.101 and 0.059, respectively, and the R^2 was increased to 0.970. The poor performance of the model on the Xincheng Bridge is due to the limitation of the size of the dataset, resulting in poor model training.

In order to show the data in Table 2 more clearly, we graphically display the predictions of the AT-BILSTM model in four sections. Figure 9 shows the prediction curve of the AT-BILSTM model in this paper over four site datasets. The red dotted lines indicate the actual collected data,

and the blue curve represents the predicted values of the model. It can be seen that the prediction values of the AT-BILSTM model proposed by us in four sections can be well in line with the actual measured values and can achieve better prediction effects on four different sections.

Ablation experiments

To analyze the effect of each component in our model, an ablation experiment was performed on the AT-BILSTM model. Table 3 lists the parameter details for each model of the ablation experiment.

The percentage of relative error of the ablation experimental model on the Fuhe bridge is calculated to evaluate the predictive performance of the model at various points. The model comparison and error distribution of the ablation experiments are shown in Fig. 10. The relative error percentage is calculated as follows:

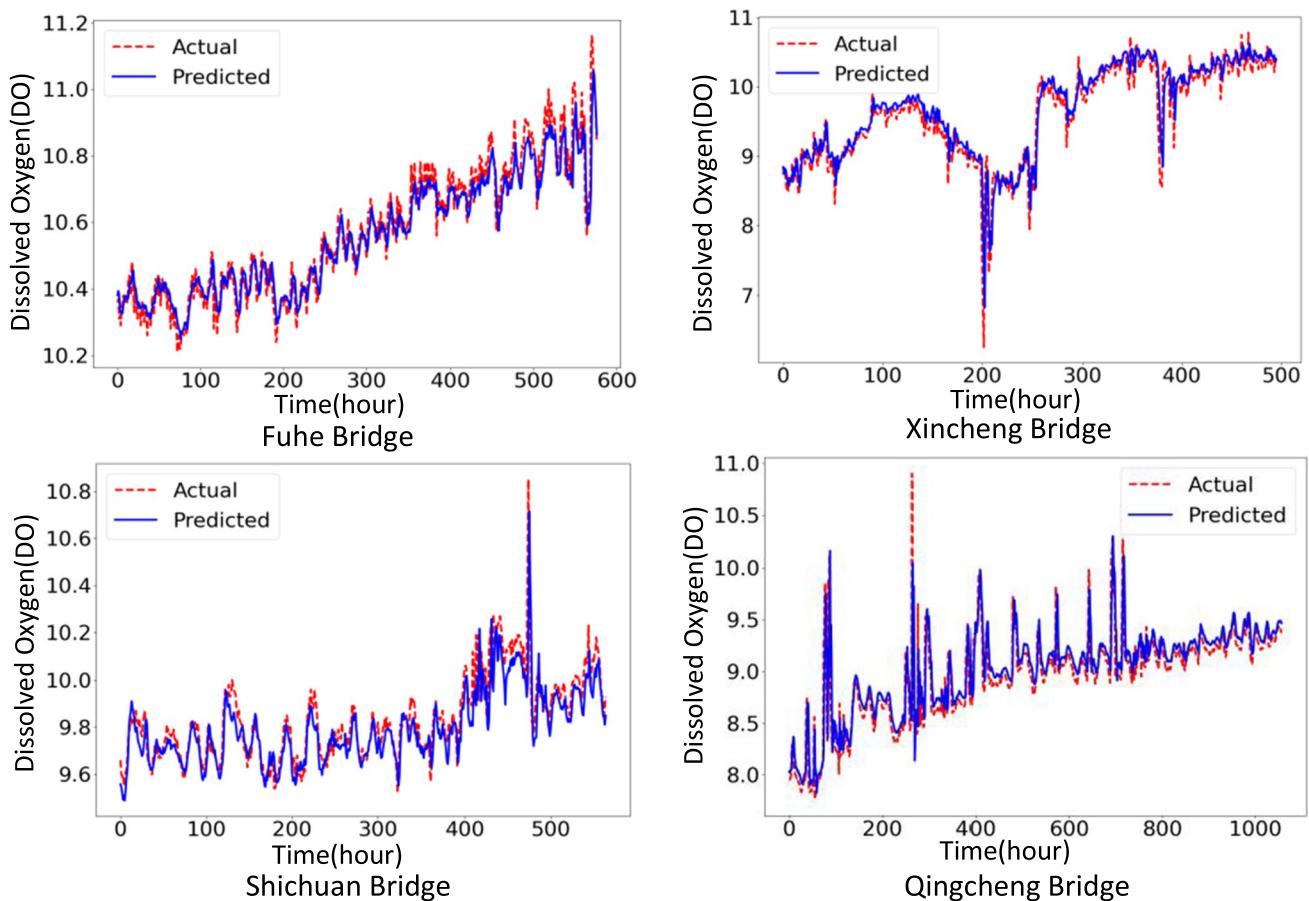$$\delta_t = \frac{|x_t - \hat{x}_t|}{x_t} \times 100\% \tag{18}$$

Fig. 9 Comparison of DO prediction and actual values on four sites

Table 3 Structures and parameters of models

Models	Layer_in	Layer_hidden	Layer_out	Parameters		
	Dimension	Dimension	Dimension	Learning rate	Epoch	Hidden dimension
LSTM	$linear : (k \times m)$	$lstm : (k \times m, m)$	$linear : (1 \times n)$	0.05	60	64
Bi-LSTM	$linear : (k \times m)$	$\overrightarrow{lstm} : (k \times m, m)$ $\overleftarrow{lstm} : (k \times m, m)$	$linear : (1 \times n)$	0.05	60	64
AT-LSTM	$linear : (k \times m)$	$lstm : (k \times m, m)$ $linear : (k \times 1, k)$	$linear : (1 \times n)$	0.05	60	64
AT-BILSTM	$linear : (k \times m)$	$\overrightarrow{lstm} : (k \times m, m)$ $\overleftarrow{lstm} : (k \times m, m)$ $linear : (k \times 1, k)$	$linear : (1 \times n)$	0.05	60	64

Fig. 10 **a** Enlarged view of each model. **b** Error distribution of each model in Fuhe Bridge dataset
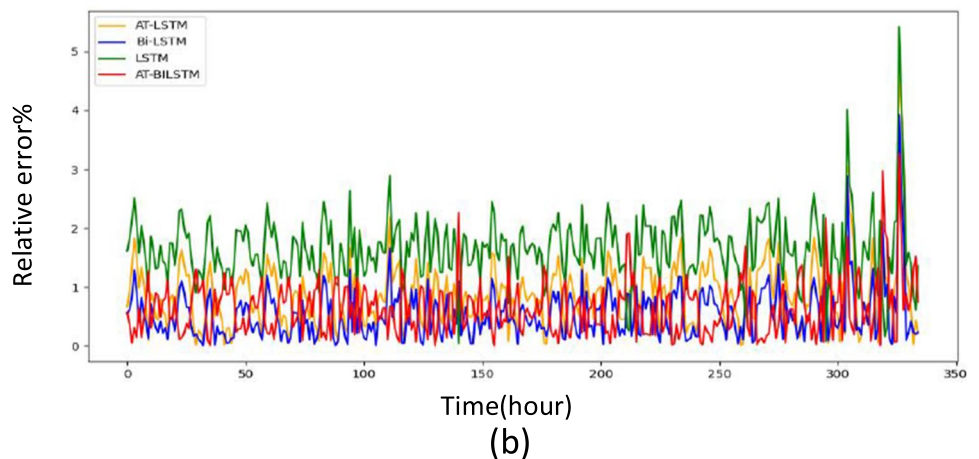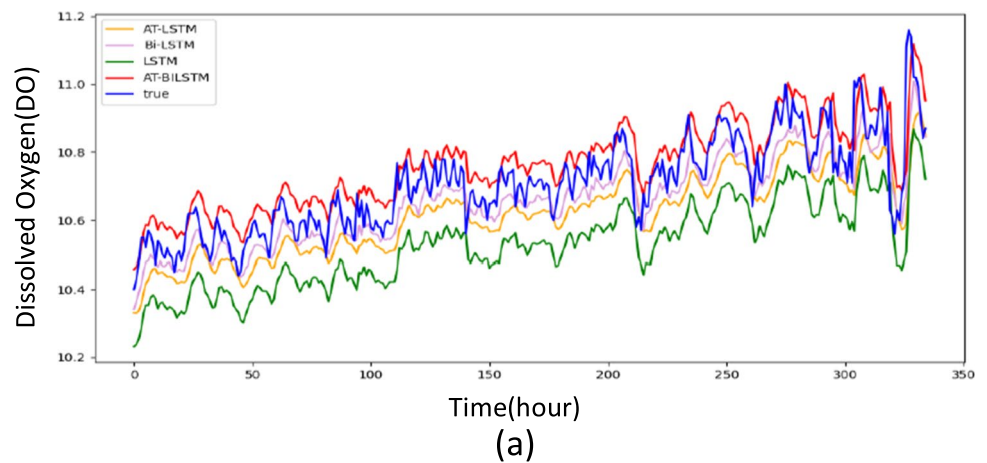
where x_t is the actual value of the moment t and $\widehat{x}_t$ is the predicted value of the moment t.

For the Fuhe Bridge dataset, the details of the various models are enlarged such as Fig. 10a shown, as can be seen from the enlarged figure, the Bi-LSTM model based on the attention mechanism proposed in this paper is superior to other models in terms of trend shape and fit the degree of upper and lower peak points. Figure 10b shows the relative error distribution for each model. It can be seen that for the Fuhe Bridge data sample, the introduction of the attention mechanism as a whole reduces the error percentage of the standard Bi-LSTM and LSTM models and improves the prediction accuracy of most sharp points.
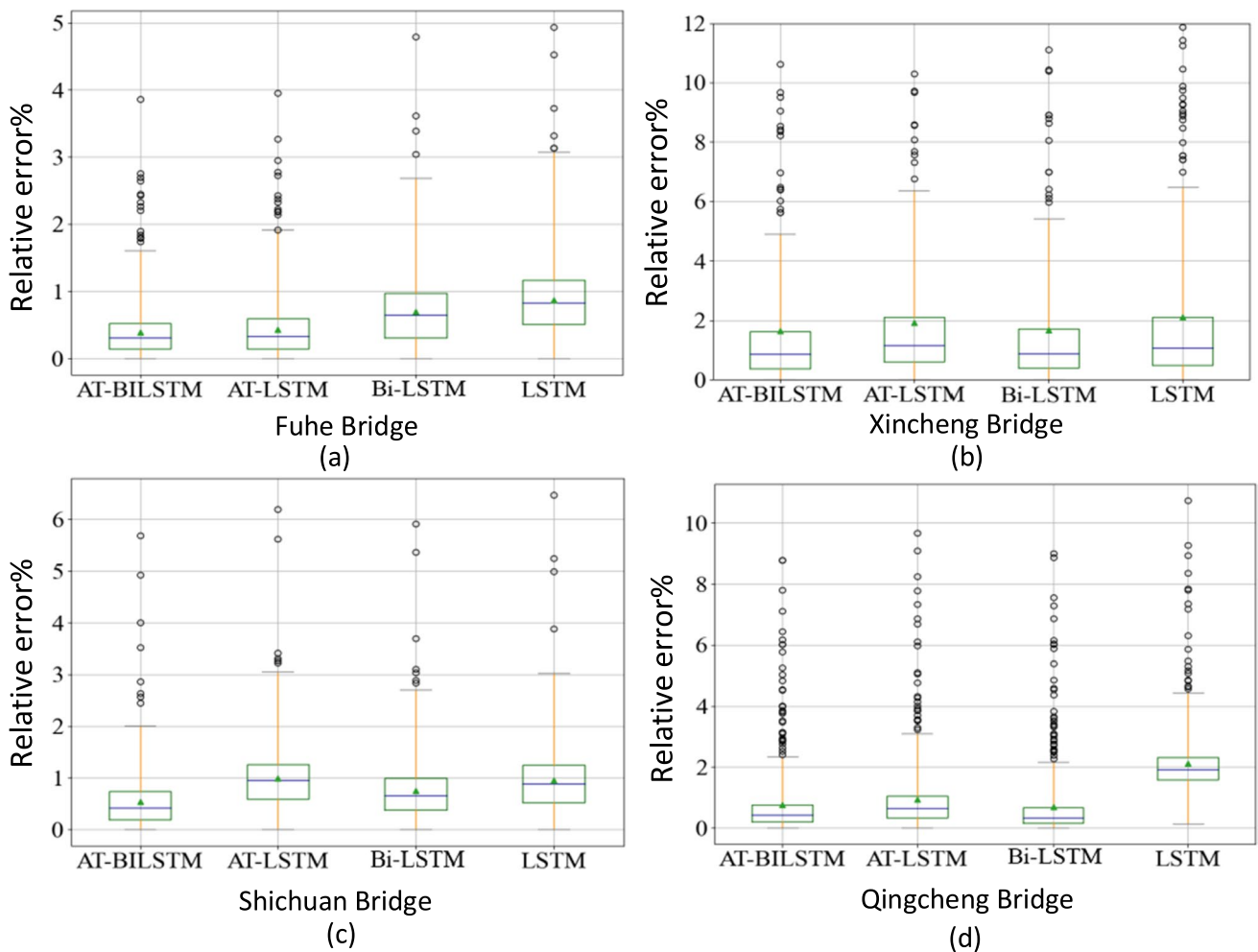
Fig. 11 Error percentage box diagram of four bridges

Figure 11 shows a box plot of the percentage relative error of the ablation model on four site datasets. It can be seen that the relative error distribution range of the AT-BILSTM model proposed in this paper is always minimal compared with other models and is better than other ablation models in most cases. But from Table 2 and Fig. 11, we can see that the models trained on the Fuhe Bridge, Shichuan Bridge, and Qingcheng Bridge data have better performance, while the models trained on the Xincheng Bridge dataset have a lower average. It follows from this that the size of the dataset has a large influence on the accuracy of the model.

The results show that the introduction of a Bi-LSTM network for bidirectional learning of data can improve the prediction accuracy of the LSTM model. On this basis, the attention layer is introduced to pay different attention to the water quality data at different times, which further improves the prediction accuracy of the model, and can fully explore the

series correlation on the water quality time series data, which is conducive to improving the accuracy of water quality prediction. It also shows once again that the neural network has good adaptability to the nonlinearity of water quality data. The AT-BILSTM model becomes more accurate and robust compared to the standard LSTM model. The temporal attention layer of the model reduces the RMSE and MAE of the LSTM model and increases R^2. This also reflects the advantages of the model we have proposed, which can improve the accuracy of water quality predictions in the watershed.

Visualization of attention mechanisms

In the previous subsection, we compared the performance of several models based on experimental details and roughly analyzed the differences between them, and it can be seen that the model based on the attention mechanism performed
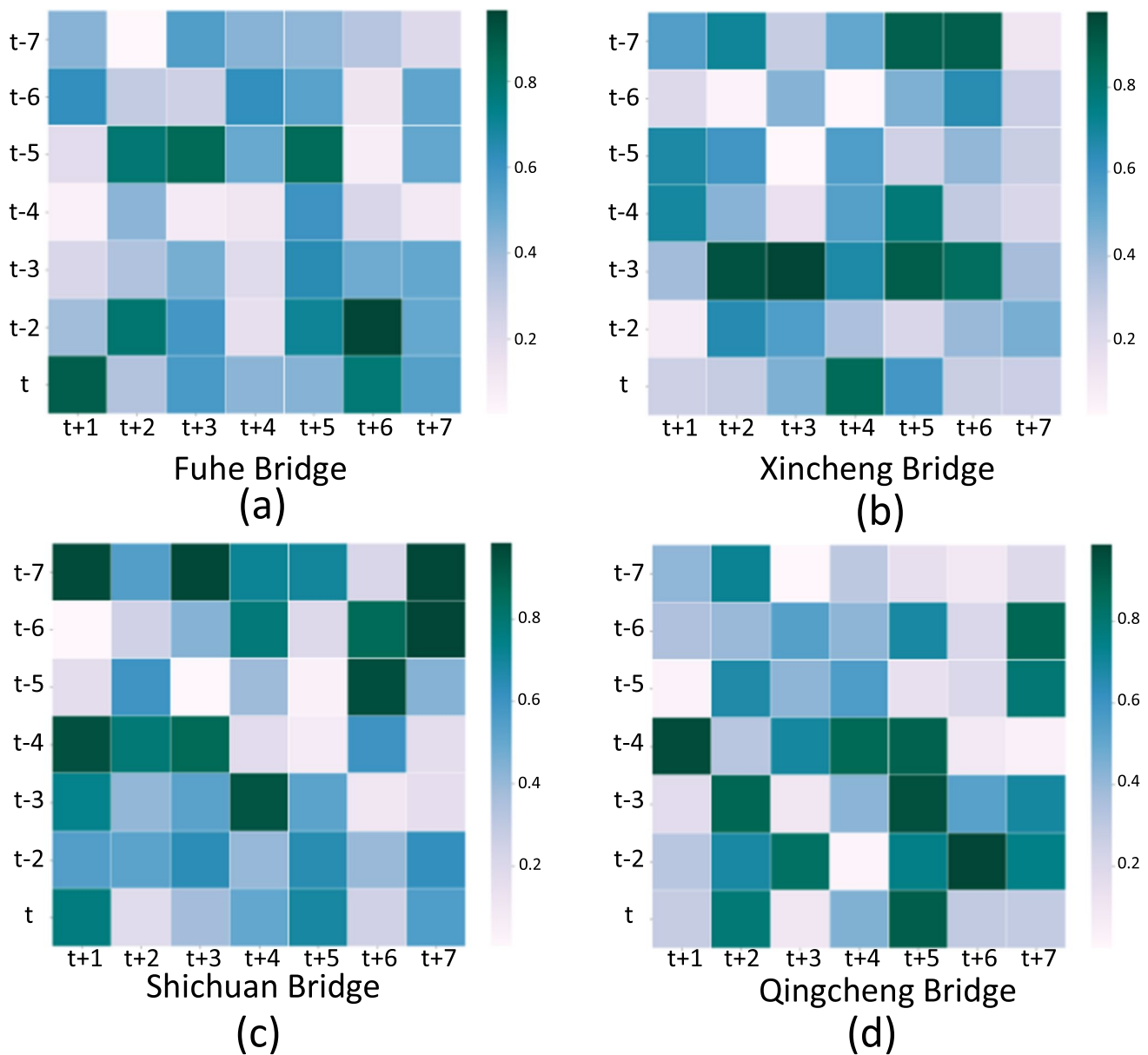
Fig. 12 This figure is the visualization result of AT-BILSTM model's time attention weights on four site datasets

better. In this section, we visualize the temporal attention weights of the AT-BILSTM model and analyze the problems of different degrees of attention of different prediction time attention mechanisms.

Figure 12 describes the change in weights of the model's temporal attention to four site datasets under the conditions of actual water quality data entry. The x coordinate is the prediction time, and the y coordinate is the historical input data. It can be seen that as the forecast time passes, the time attention weight also gradually changes, and the time attention module focuses on different past moments at different prediction moments.

Conclusion

With the promotion of fine management of water pollution prevention and control, a large number of water quality monitoring sites have been deployed in many river basins around the world. How to effectively use the large amount of data collected by these monitoring stations is an important issue with the potential to help improve the ecological environment of the watershed. In this paper, a watershed water quality prediction model (AT-BILSTM) based on the combination of attention mechanism and Bi-LSTM is

proposed and applied to the water quality prediction of multiple influencing factors in the complex environment of the Yellow River Basin in China.

Taking the four station datasets of Fuhe Bridge, Xincheng Bridge, Shichuan Bridge, and Qingcheng Bridge in Lanzhou Section of the Yellow River Basin as an example, the experimental evaluation was carried out, and the model was analyzed. The results show that:

1) For the water environment quality prediction model based on deep learning, the Bi-LSTM model is adopted so that the prediction method can learn the data series to be learned in both directions, and the characteristics of the data can be learned from the double neighborhood. In addition, after visually displaying the time attention weight of the model, it shows that the model can highlight the effective characteristics of the input dataset and obtain better accuracy of the model. Moreover, experimental studies have shown that the average prediction performance of the proposed AT-BILSTM model in four sections is better than that of the selected baseline model.

2) The quality characteristics of the water environment in different sections are different, and the trend of change is also different. Therefore, even the same model will have different performances in different sections, and the prediction performance of sections with stable data is better. The proposed AT-BILSTM model can also fit the data well when the data fluctuations are large and has better generalization ability. And because of the attention mechanism and the reference of Bi-LSTM, the model has a great advantage in feature extraction for sequence correlation.

The proposed AT-BILSTM model is a water quality prediction model based on attention mechanism and Bi-LSTM, which can also be used to predict other time series-based pollutants. With the vigorous promotion of water environment governance in river basins, the proposed water quality prediction model has great potential for application and can provide a guarantee for the safety and stability of the water quality environment. At the same time, because the deployment of water quality monitoring sites has only begun to be promoted in recent years, historical data is limited, which may have a certain impact on the training of models. In the next step, as more water quality monitoring sites are deployed, the forecasting model will be further enhanced and optimized. In addition, the performance of deep learning models is often related to different parameters and datasets. Therefore, how to improve and stabilize the forecasting ability of our models and better support online time series forecasting and early warning applications is the focus of our future research.

Declarations

Ethics approval and consent to participate Not applicable.

Consent for publication Not applicable.

Competing interests The authors declare no competing interests.

References

Brown A, Tuor A, Hutchinson B, Nichols N (2018) Recurrent neural network attention mechanisms for interpretable system log anomaly detection. In: Proceedings of the First Workshop on Machine Learning for Computing Systems. 1–8. https://doi.org/10.1145/3217871.3217872

Cho K, Van Merriënboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, Bengio Y (2014) Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv 1406–1078. https://doi.org/10.48550/arXiv.1406.1078

Chorowski JK, Bahdanau D, Serdyuk D, Cho K, Bengio Y (2015) Attention-based models for speech recognition. Advances in neural information processing systems 28. https://doi.org/10.48550/arXiv.1506.07503

Evans AE, Hanjra MA, Jiang Y, Qadir M, Drechsel P (2012) Water quality: assessment of the current situation in Asia. Int J Water Resour Dev 28(2):195–216. https://doi.org/10.1080/07900627.2012.669520

Faruk DÖ (2010) A hybrid neural network and ARIMA model for water quality time series prediction. Eng Appl Artif Intell 23(4):586–594. https://doi.org/10.1016/j.engappai.2009.09.015

Feng D, Fang K, Shen C (2020) Enhancing streamflow forecast and extracting insights using long-short term memory networks with data integration at continental scales. Water Resour Res 56(9):e2019WR026793. https://doi.org/10.1029/2019WR026793

Gholizadeh MH, Melesse AM, Reddi L (2016) Water quality assessment and apportionment of pollution sources using APCS-MLR and PMF receptor modeling techniques in three major rivers of South Florida. Sci Total Environ 566:1552–1567. https://doi.org/10.1016/j.scitotenv.2016.06.046

González FUT, Herrera-Silveira JA, Aguirre-Macedo ML (2008) Water quality variability and eutrophic trends in karstic tropical coastal lagoons of the Yucatán Peninsula. Estuar Coast Shelf Sci 76(2):418–430. https://doi.org/10.1016/j.ecss.2007.07.025

Hirsch RM, Slack JR, Smith RA (1982) Techniques of trend analysis for monthly water quality data. Water Resour Res 18(1):107–121. https://doi.org/10.1029/WR018i001p00107

Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hu D (2019) An introductory survey on attention mechanisms in NLP problems. In: Proc SAI Intell Syst Conf 432–448. https://doi.org/10.1007/978-3-030-29513-4_31

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. https://doi.org/10.3390/s19061420

Huang H, Zhang B, Lu J (2014) Quantitative identification of riverine nitrogen from point, direct runoff and base flow sources. Water Sci Technol 70(5):865–870. https://doi.org/10.2166/wst.2014.303

Huang H, Ji X, Xia F, Huang S, Shang X, Chen H, Zhang M, Dahlgren RA, Mei K (2020) Multivariate adaptive regression splines for estimating riverine constituent concentrations. Hydrol Process 34(5):1213–1227. https://doi.org/10.1002/hyp.13669

Jaynes ET (1982) On the rationale of maximum-entropy methods. Proc IEEE 70(9):939–952. https://doi.org/10.1109/PROC.1982.12425

Jouanneau S, Recoules L, Durand M, Boukabache A, Picot V, Primault Y, Lakel A, Sengelin M, Barillon B, Thouand G (2014) Methods for assessing biochemical oxygen demand (BOD): A review. Water Res 49:62–82. https://doi.org/10.1016/j.watres.2013.10.066

Khullar S, Singh N (2022) Water quality assessment of a river using deep learning Bi-LSTM methodology: forecasting and validation. Environ Sci Pollut Res 29(9):12875–12889. https://doi.org/10.1007/s11356-021-13875-w

Kisi O, Parmar KS (2016) Application of least square support vector machine and multivariate adaptive regression spline models in long term prediction of river water pollution. J Hydrol 534:104–112. https://doi.org/10.1016/j.jhydrol.2015.12.014

Le T, Vo MT, Vo B, Hwang E, Rho S, Baik SW (2019) Improving electric energy consumption prediction using CNN and Bi-LSTM. Appl Sci 9(20):4237. https://doi.org/10.3390/app9204237

Li H, Shen Y, Zhu Y (2018) Stock price prediction using attention-based multi-input LSTM. Proc Mach Learn Res 95:454–469

Lin L, Li W, Bi H, Qin L (2021) Vehicle trajectory prediction using LSTMs with spatial-temporal attention mechanisms. IEEE Intelligent Transportation Systems Magazine. https://doi.org/10.1109/MITS.2021.3049404

Liu DR, Lee SJ, Huang Y, Chiu CJ (2020) Air pollution forecasting based on attention-based LSTM neural network and ensemble learning. Expert Syst 37(3):e12511. https://doi.org/10.1111/exsy.12511

Liu S, Tai H, Ding Q, Li D, Xu L, Wei Y (2013) A hybrid approach of support vector regression with genetic algorithm optimization for aquaculture water quality prediction. Math Comput Model 58(3–4):458–465. https://doi.org/10.1016/j.mcm.2011.11.021

Ma F, Chitta R, Zhou J, You Q, Sun T, Gao J (2017) Dipole: diagnosis prediction in healthcare via attention-based bidirectional recurrent neural networks. In: Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining. 1903–1911. https://doi.org/10.1145/3097983.3098088

Noori N, Kalin L, Isik S (2020) Water quality prediction using SWAT-ANN coupled approach. J Hydrol 590:125220. https://doi.org/10.1016/j.jhydrol.2020.125220

Pulver A, Lyu S (2017) LSTM with working memory. In: 2017 International Joint Conference on Neural Networks (IJCNN). 845–851. https://doi.org/10.1109/IJCNN.2017.7965940

Shahid F, Zameer A, Muneeb M (2020) Predictions for COVID-19 with deep learning models of LSTM, GRU and Bi-LSTM. Chaos, Solitons Fractals 140:110212. https://doi.org/10.1016/j.chaos.2020.110212

Sun Q, Jankovic MV, Bally L, Mougiakakou SG (2018) Predicting blood glucose with an lstm and bi-lstm based deep neural network. In: 2018 14th Symposium on Neural Networks and Applications (NEUREL). 1–5. https://doi.org/10.1109/NEUREL.2018.8586990

Tang G, Sennrich R, Nivre J (2018) An Analysis of Attention Mechanisms: The Case of Word Sense Disambiguation in Neural Machine Translation. In Proceedings of the Third Conference on Machine Translation: Research Papers, pages 26–35, Brussels, Belgium. Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-6304

Wang L, Zou H, Su J, Li L, Chaudhry S (2013) An ARIMA-ANN hybrid model for time series forecasting. Syst Res Behav Sci 30(3):244–259. https://doi.org/10.1002/sres.2179

Ye Q, Yang X, Chen C, Wang J (2019) River water quality parameters prediction method based on LSTM-RNN model. In: 2019 Chinese Control And Decision Conference (CCDC). 3024–3028. https://doi.org/10.1109/CCDC.2019.8832885

Zhang B, Zhang H, Zhao G, Lian J (2020) Constructing a PM2.5 concentration prediction model by combining auto-encoder with Bi-LSTM neural networks. Environ Model Software 124:104600. https://doi.org/10.1016/j.envsoft.2019.104600

Zhu X, Li D, He D, Wang J, Ma D, Li F (2010) A remote wireless system for water quality online monitoring in intensive fish culture. Comput Electron Agric 71:S3–S9. https://doi.org/10.1016/j.compag.2009.10.004