Water Science & Technology

IWA PUBLISHING

Water Science & Technology Vol 88 No 3, 530 doi: 10.2166/wst.2023.156

Research on water quality spatiotemporal forecasting model based on ST-BIGRU-SVR neural network

Rongli Gai and Jiahui Yang *

School of Information Engineering, Dalian University, Dalian 116622, China
*Corresponding author. E-mail: 2670281026@qq.com

JY, 0000-0003-1021-2894

ABSTRACT

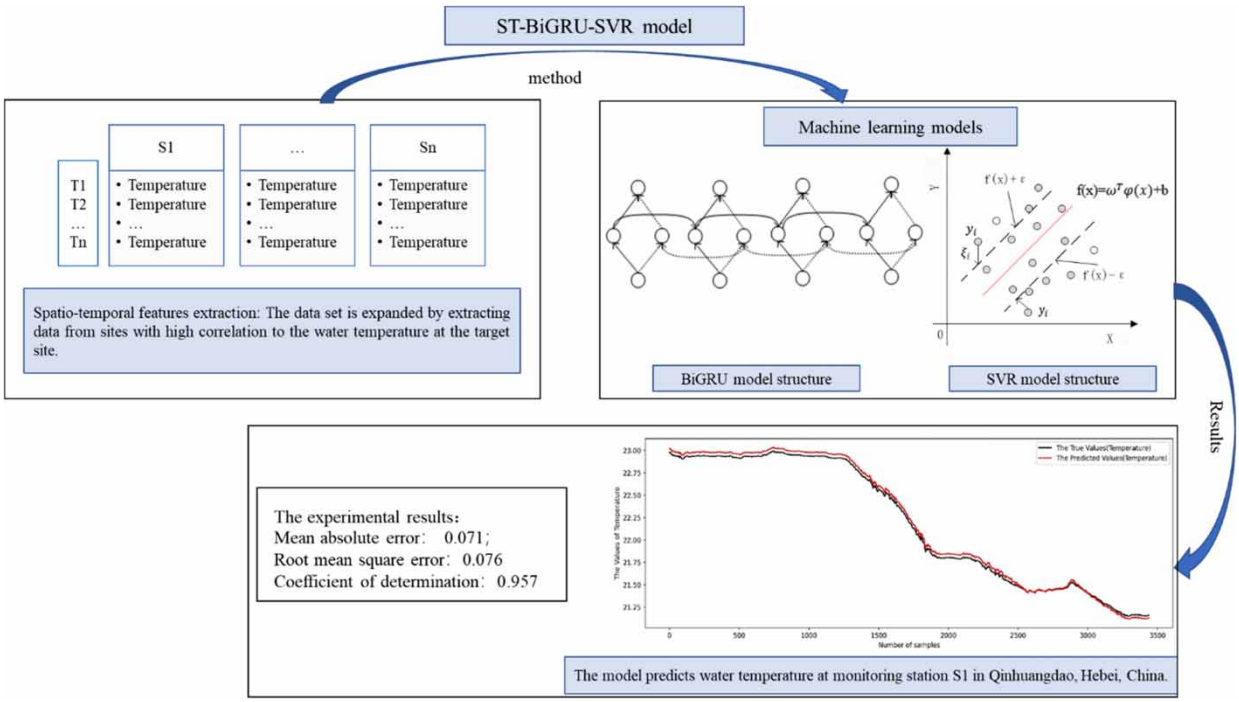
With the serious deterioration of the water environment, accurate prediction of water quality changes has become a topic of increasing concern. To further improve the accuracy of water quality prediction and the stability and generalization ability of the model, we propose a new water quality spatiotemporal forecast model to predict future water quality. To capture the spatiotemporal characteristics of water quality pollution data, the three sites (station S1, station S2, station S4) with the highest temperature time series concentration correlation at the experimental sites were first extracted to predict the water temperature at station S1, and 17,380 records were collected at each monitoring station, and the spatiotemporal characteristics were extracted by BiGRU-SVR network model. This paper's prediction test is based on the actual water quality data of the Qinhuangdao sea area in Hebei province from 2 September to 26 September 2013 and compared with other baseline models. The experimental results show that the proposed model is better than other baseline models and effectively improves the accuracy of water quality prediction, and the mean absolute error (MAE), root mean square error (RMSE), and coefficient of determination (R^2) are 0.071, 0.076, and 0.957, respectively, which have good robustness.

Key words: hybrid model, machine learning, water quality data, water quality prediction, spatiotemporal correlation

HIGHLIGHTS

- In order to capture the spatiotemporal characteristics of water pollution data, a new bidirectional gated recurrent unit networks and support vector regression model hybrid neural network model proposed in this paper focuses on water quality data trends and contextual temporal attributes to capture the spatiotemporal characteristics of water pollution data.
- Multiparameter water quality prediction is realized, with comprehensive consideration of the correlations between data.

GRAPHICAL ABSTRACT



1. INTRODUCTION

With the rapid development of urbanization and industrialization in recent years, water quality has deteriorated and water safety has suffered from great threats and challenges. Water quality safety is a guarantee for human health and aquaculture, and timely monitoring and scientific management of water quality are of great importance to ensure water safety. The Ministry of Ecology and Environment of China released the 'Ecological Environment Monitoring Planning Outline (2020–2035)', which emphasizes the need to speed up the implementation of the Ministry of Water Resources to build an automatic water quality monitoring network to ensure that the monitoring data are 'accurate and complete'. Water quality prediction is an important part of the automatic water quality monitoring network, which is a prerequisite for all subsequent work and an important tool for water resources protection. As an important factor affecting the stability of water ecosystems, water temperature not only restricts the reproduction and growth of aquatic plants and animals, the living environment, and the population distribution of each aquatic species but also affects the dissolved oxygen and chemical reactions of toxic substances in water (Fu *et al.* 2021). Therefore, water temperature plays a key role in water resource management decisions. Advance prediction of water quality parameter concentrations is the basis for water quality pollution prevention and control, achieving integrated environmental management and is important for public health and governmental decision-making.

However, the current water temperature prototype observation has great limitations due to the influence of factors, such as observation point location and instrument accuracy (Quan *et al.* 2020). Therefore, it is necessary to explore the high-precision water temperature simulation method, which has important reference value for the management and protection of the reservoir water environment and aquaculture. In water quality prediction, the existing methods include the water quality simulation model, gray theory method (Wu *et al.* 2021), regression analysis method (Bauwe *et al.* 2019), time series method (Li *et al.* 2021), and neural network method (Deng *et al.* 2021). Water quality simulation model is an early and widely used method for water quality research, through a series of water quality indicators, to establish a mathematical model to predict future water quality changes, such as the Soil and Water Assessment Tool (SWAT) model (Akoko *et al.* 2021). This type of model can accurately simulate the basic water quality laws but is often only applicable to a small range of waters, such as specific lakes and rivers, the generality is poor. Gray theory method, through a small amount of incomplete information, the establishment of a gray differential prediction model, the development of the law of things to make a fuzzy long-term description. However, its drawback is the large average error in prediction and the inability to predict

interval number time series. Regression analysis method, using regression equations to fit the relationship between the independent and dependent variables, in the prediction of water quality needs to first analyze the correlation coefficient between water quality indicators, commonly used linear regression (LR) models, multiple regression analysis, and so on. For the water environment affected by the interaction of multiple factors, regression analysis considering multiple variables has the unique advantage of predicting better when the situation is simple, but it is difficult to express the highly complex data well. The time series method, using the principles of mathematical statistics, analysis, and collation of the historical series of water quality data itself, the study of water quality data parameters change trends and thus achieve the purpose of predicting the future, the method is often used for short and medium-term water quality forecasting, such as autoregressive integrated moving average model (ARIMA) (Chen *et al.* 2021), autoregressive conditional heteroskedastic (ARCH) (Wu *et al.* 2012), and so on. Its advantage is that it does not need to rely on other variables; however, it has the following limitations: it can only use its data for prediction and requires a long enough historical series; the data must be auto-correlated; it can only capture linear relationships, not nonlinear relationships between attributes.

The use of machine learning neural network models for water resources domain research is a current research hotspot. Neural networks are extensive and interconnected complex network structures composed of a large number of simple units with adaptability, which can simulate the process of human brain information processing through effective learning mechanisms (Haghiabi *et al.* 2018). Able to learn by themselves the nonlinear mapping relationship between the historical parameters of water quality and the external factor variables, through this mapping relationship applied to the future forecast. To efficiently integrate relevant information in the context of time series data, Liu *et al.* (2020) and other scholars successfully applied bidirectional stacked simple recursive units (Bi-S-SRU) to mariculture water quality prediction and demonstrated the feasibility of bidirectional neural networks to predict water quality parameters. Yan *et al.* (2021) used one-deep residual convolutional neural network (CNN) (1-DRCNN) and bidirectional gating recurrent unit (BIGRU) hybrid neural networks to predict the water quality of the Luan River, fully extracting the potential local features among water quality parameters and integrating the before and after time series information. Hu *et al.* (2019) for smart mariculture proposed a water quality prediction method based on deep long short term memory (LSTM) learning network to predict PH and water temperature. Wang *et al.* (2019) established a water pollution prediction model with eutrophication indicators total phosphorus and total nitrogen as parameters based on a time series ARIMA model with the introduction of the Holt–Winters seasonal model for optimization. Valadkhan *et al.* (2022) proposed a groundwater quality parameter prediction method with new effective parameters based on LSTM and recurrent neural networks (RNN) to prevent water pollution. Zhou *et al.* (2022) proposed the integrated wavelet decomposition, autoregressive integrated moving average, and gated recurrent unit (W-ARIMA-GRU) model for water quality prediction for decomposing the original water quality index data series into two series of trends and fluctuations, and the characteristics of the decomposed series data. Haq & Harigovindan (2022)) used CNN to obtain aquaculture water quality characteristics, LSTM and GRU to learn long-term dependencies in time series data and proposed a hybrid deep learning model for water quality prediction. Yan *et al.* (2020) predicted water quality based on a deep belief network and a least squares support vector regression model (PSO-DBN-LSSVR). In addition to error back-propagation (BP) networks (Huang *et al.* 2022) support vector regression (SVR) (Su *et al.* 2022), temporal convolutional networks (TCN) models (Li *et al.* 2022), and other machine learning methods are also used in water quality prediction. These different structures of machine learning models provide an important reference for the study in this paper.

In this paper, spatially, the spatial dependence is captured based on the influence of water quality parameters by other monitoring sites; in the temporal dimension, the water quality temperature at the experimental site is also influenced by the past water temperature at the site, so the temporal dependence is captured. We propose a new hybrid neural network model combining SVR and BiGRU to focus on the trend and contextual temporal attributes of water quality data and then predict the water temperature in Qinhuangdao water, Hebei, China. Finally, the experimental structure is compared with a total of seven models, both non-deep and deep models, and the accuracy and stability of each method are evaluated based on real values.

2. DATA AND METHOD

2.1. Research area and data

The study area of this paper is located between 119°34′–119°54′ E and 39°25′–39°42′ N in Qinhuangdao Sea, Bohai Sea, China, and the data are obtained from the water quality temperature and salinity observation data of Qinhuangdao sea

monitoring station. The temperature and salinity water quality dataset use data records from 2 September to 26 September 2013, with monitoring station sensors collecting experimental data every 10 min, and 17,380 records were collected at each monitoring station. The name, data volume, and location information of each monitoring station are listed in Table 1. In this paper, one of the four monitoring stations was selected as the evaluation point (S1) for water temperature prediction.

Water temperature data were collected from four monitoring stations via water temperature sensors to provide data support for subsequent data analysis modeling. The monitoring station sensors collected data at a frequency of one data item every 10 min, with data collected from 2 September to 26 September 2013, with 17,380 data items collected at each monitoring station.

2.2. Extraction for spatial factors

According to Tobler's first law of geography, everything is related to everything else, and similar things are more closely connected, neighboring sites have a greater influence on the experimental site than distant sites (Li *et al.* 2021a, 2021b). To illustrate the spatial characteristics of the water temperature data, the authors calculated the distance between two monitoring points and the Pearson correlation coefficient of the water temperature between each monitoring point.

The distance between two stations is calculated based on latitude and longitude according to the Haversine formula (Mundu *et al.* 2022). The formula uses a sine function to keep enough valid numbers, even if the distance is small. The formula is as follows:

$$\text{Haversin}\left(\frac{d}{R}\right) = \text{haversin}(\varphi_2 - \varphi_1) + \cos(\varphi_1)\cos(\varphi_2)\text{haversin}(\lambda_2 - \lambda_1) \tag{1}$$

Haversin is an abbreviation for semi-positive vector function:

$$\text{Haversin}(\theta) = \sin^2\left(\frac{\theta}{2}\right) = \frac{1 - \cos(\theta)}{2} \tag{2}$$

where d is the distance between the two points and R is the radius of the Earth, taking the average value of 6,371 km, φ_1, φ_2 indicates the latitude of the two stations, and λ_1, λ_2 indicates the longitude of the two stations.

The Pearson correlation coefficient method (Wang *et al.* 2020) is commonly used to qualitatively measure and analyze the degree of linear correlation between variables. The formula is as follows:

$$r_{xy} = \frac{S_{xy}}{S_x S_y} \quad \epsilon[-1, 1] \tag{3}$$

$$S_{xy} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n - 1} \tag{4}$$

$$S_x = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n - 1}} \tag{5}$$

Table 1 | Monitoring station name and location

Monitoring station name	Number of data	Location	
		Longitude	Latitude
Station 1 (S1)	17,380	119°37′ 52.3812″E	39°39′ 55.3212″N
Station 2 (S2)	17,380	119°54′ 57.5994″E	39°42′ 21.3588″N
Station 3 (S3)	17,380	119°45′ 55.4394″E	39°37′ 1.4406″N
Station 4 (S4)	17,380	119°34′ 56.46″E	39°25′ 55.9812″N

$$S_y = \sqrt{\frac{\sum (y_i - \bar{y})^2}{n-1}} \tag{6}$$

where r_{xy} represents the sample correlation coefficient, S_{xy} represents the sample covariance, S_x represents the standard deviation of sample X, and S_y represents the standard deviation of sample Y. The closer the value of $|r_{xy}|$ to 1, which indicates that the stronger the data correlation, $r_{xy} > 0$ indicates that the two are positively correlated; $r_{xy} < 0$ indicates a negative correlation.

Knowing the target site S1(φ_1, λ_1) and the other sites Sn(φ_2, λ_2), $n \in \{2, 3, 4\}$ brings the coordinate values of the surrounding sites Sn into the Haversin formula above to calculate the distance between the other sites Sn and the target site S1. The Pearson correlation coefficient is used to calculate the correlation between the water temperature of the surrounding sites and the target site, and the water temperature data from the sites with a high correlation with the target site are selected and fused with the target site's water temperature for later model input.

Based on the water temperature correlation analysis of the monitoring stations, water temperature data from stations with correlation coefficient values greater than 0.8 were selected to improve the prediction accuracy of the target stations by using adjacent monitoring stations. The process of spatial factor extraction is shown in Figure 1. The initial dataset used in this paper is time series data collected from four stations taken every 10 min. Using station S1 as the target station, the correlation coefficient between the stations was calculated according to the above formula, and the correlation between station S2 and station S4 was calculated to meet 0.8. The water temperature data of station S2 and station S4 were extracted, and finally, the time series data of the three monitoring stations were obtained. The data are in the form of three features per data record (own water temperature, neighboring station S2_water temperature, neighboring station S4_water temperature), which are used as initial data for the next stage. The water temperature data collected as a special time series is time series in the sense that the changes in the water temperature data are time-dependent. To further reflect the spatiotemporal correlation between stations, the station time data obtained above was fed into the model and then its spatiotemporal correlation was extracted using a neural network.

2.3. Data cleaning

The water quality data used in this paper come from different collection devices of automatic monitoring stations. Due to equipment failure or human error records and other factors, water quality data will inevitably have abnormal values and vacant values. These non-conforming data can lead the algorithm to have a poor grasp of the direction of the predicted values, leading to a decrease in the accuracy of the model. Therefore, the dataset must be cleaned before constructing the prediction model. In this paper, the Pauta criterion and the generative adversarial networks approach are used to correct water quality data.

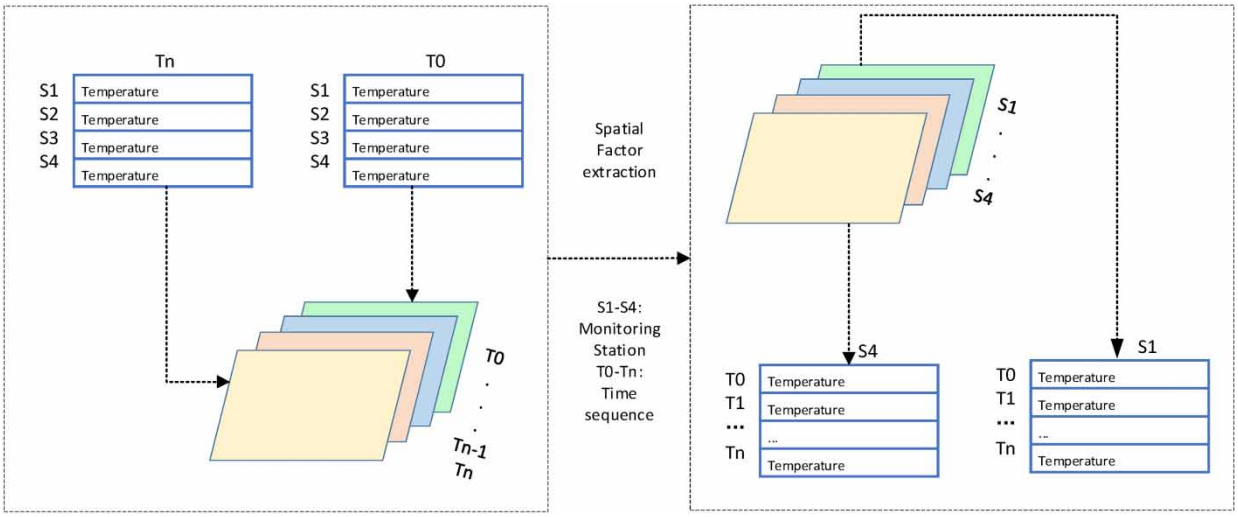


Figure 1 | Spatial factor extraction.

2.3.1. Pauta criterion

The literature (Shen *et al.* 2020) used the Pauta criterion (also known as the 3σ principle) to effectively detect anomalous values in highway traffic flow, showing that the algorithm is able to correctly identify anomalous data. Pauta criterion is to assume that a set of test data contains only random errors, which are calculated and processed to obtain the standard deviation, and determine an interval with a certain probability, and consider that any error exceeding this interval is an outlier (the general threshold value is $\mu + 3\sigma$). The Pauta criterion formula is:

$$|V_i| = |x_i - \bar{x}| > 3\sigma \tag{7}$$

where $\bar{x}$ is the sample mean, $V_i = x_i - \bar{x}$, and σ is the standard deviation. If a value of the sample satisfies Equation (7), x_i is considered to be excluded.

2.3.2. Generative adversarial network

The basic structure of a generative adversarial network (GAN) consists of two network structures: a generative model (G), and a discriminator model (D). The G network generates false samples satisfying the positive sample distribution as much as possible, while the D network discriminates the true and false samples as much as possible, and the performance of both networks becomes better and better in this game process. GAN has two main features compared to other generative models: it does not rely on any a priori assumption. Many traditional methods assume that the data obey a certain distribution and then use a great likelihood to estimate the data distribution. GAN is a generative model for learning data distribution by adversarial means, and its aim is to obtain a generative network with good results through this adversarial game. The structure of the GAN model is shown in Figure 2.

2.4. ST-BiGRU-SVR

The prediction framework diagram of the model proposed in this paper is shown in Figure 3. The input to the model is a two-part data fusion that includes inter-site auto-correlated water temperature and adjacent site water temperature data. The output is the predicted value of the water temperature at the experimental site at $t+1, t+2, \ldots, t+N$. The model is divided into three parts: site self-correlation water temperature and adjacent site correlation water temperature data extraction, auxiliary data fusion and spatiotemporal feature extraction, and future water temperature prediction.

The first part is the extraction of spatial factors. That is, the extraction of station self-correlation water temperature and water temperature data from neighboring stations. The details have been described in detail in Section 2.2.

The second part is the fusion of auxiliary data. Auxiliary data are incorporated to extract more spatiotemporal features during model training. All data are cleaned and processed for missing values before use, and records with outliers are removed. For the water temperature values at each site, the authors used the Pauta criterion and GAN model padding to process the water quality data. The merged data are normalized and used as input for the next stage.

The last part is the extraction of spatiotemporal features and the prediction of future water temperature. After cleaning the data and normalizing the data, the enhanced dataset is obtained. The spatiotemporal features of the normalized time series

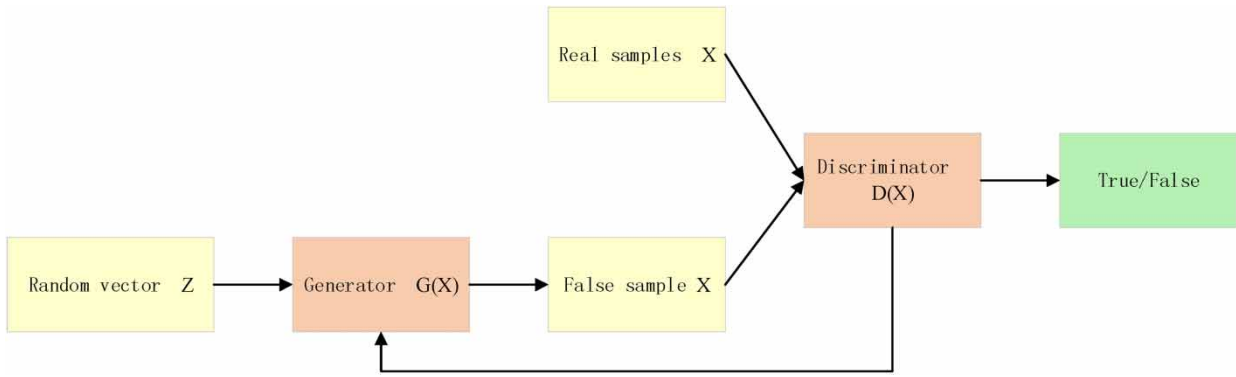


Figure 2 | GAN model.

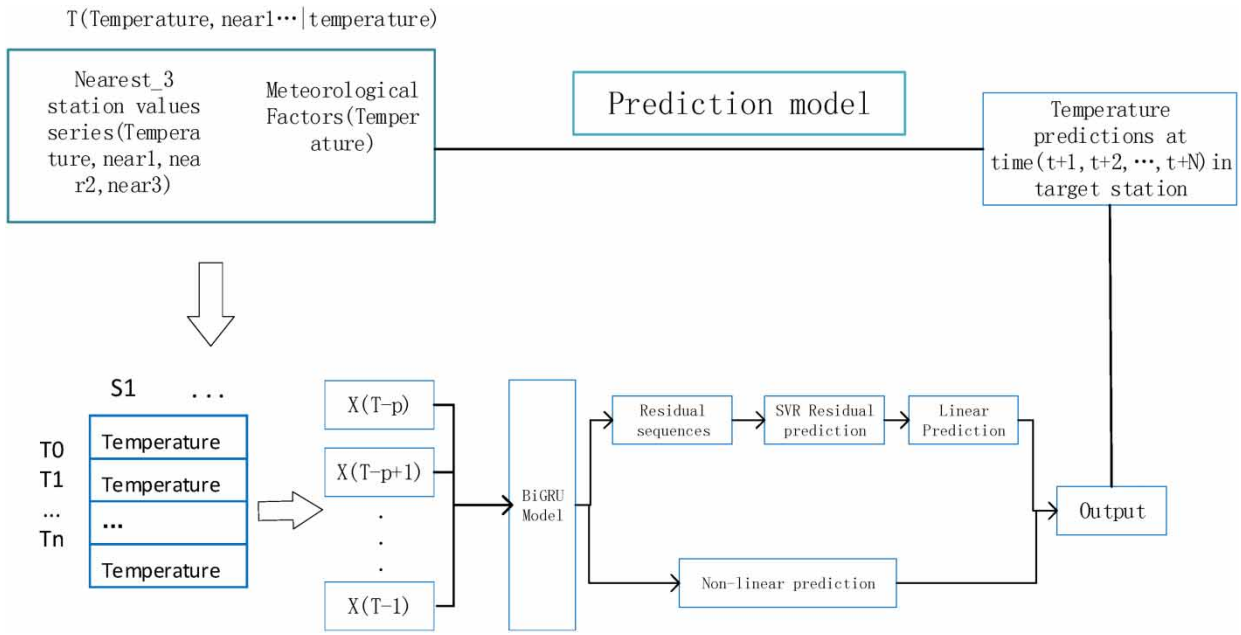


Figure 3 | Overall model prediction framework.

data are extracted by the BiGRU model and the SVR model. The predicted series values at $(t+1, t+2, \ldots, t+N)$ time the predicted series values are predicted using lagged data from past time t.

The structure of the combined water quality prediction model is shown in Figure 4.

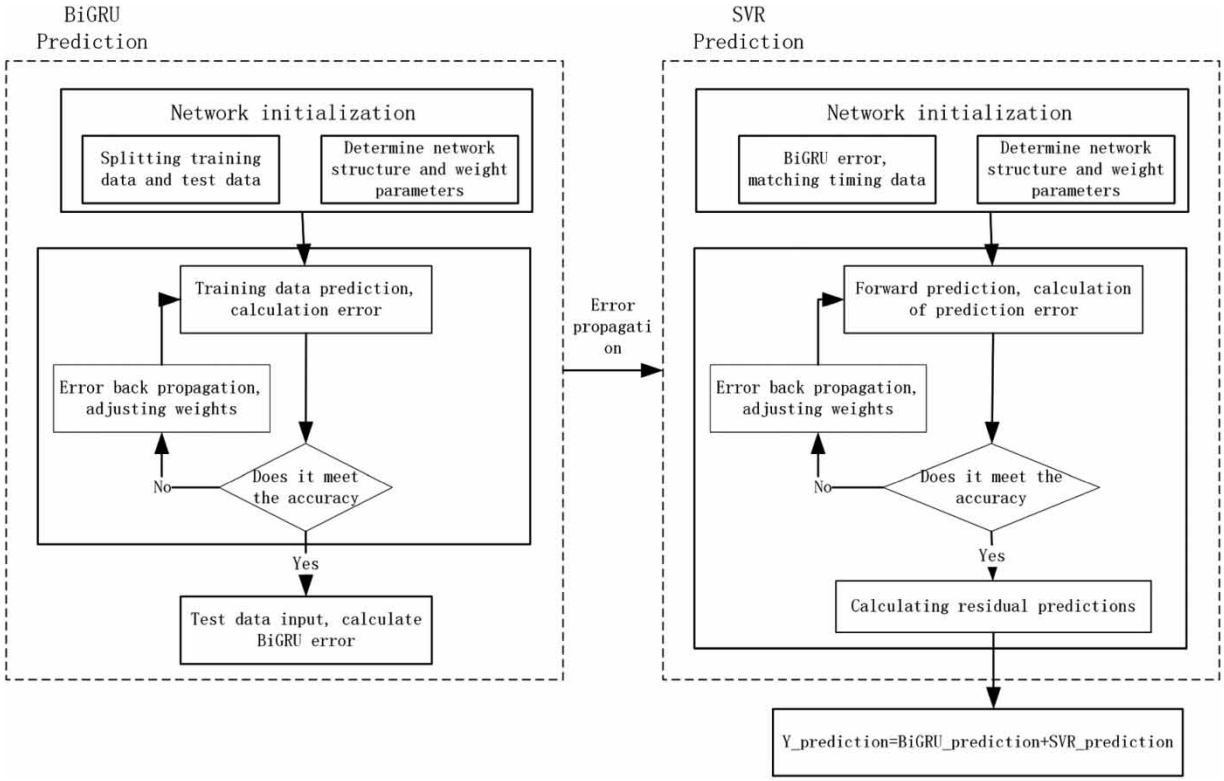


Figure 4 | Combined model prediction process.

In this paper, the BIGRU model is fused with the SVR model to obtain the ST-BIGRU-SVR water temperature prediction model. First split the input data to form the training data and test data for the ST-BIGRU model, determine the ST-BIGRU network structure and weights, queue parameters, train the network, make predictions on the test data, and calculate the prediction error. Then, based on the ST-BIGRU error, the timing data are matched again and normalized as the input data of the ST-SVR network, the ST-SVR network structure and the weight and queue parameters are determined, training and prediction are performed, and the ST-SVR residual predictions are obtained. Finally, the ST-BIGRU and ST-SVR residual prediction results are summed to obtain the final predicted values.

3. EXPERIMENTS

3.1. Evaluation

To evaluate the performance of the model proposed in this paper, the authors used three evaluation metrics, namely, mean absolute error (MAE), root mean square error (RMSE), and decision coefficient (R^2). Due to the limitations of RMSE and MAE, the same algorithmic model solving different problems does not reflect the merits of this model for different problems. It is impossible to determine which model is more suitable for predicting which problem because the data differ in different practical applications and it is not possible to directly compare the predicted values. Therefore, the prediction results are converted to accuracy, and the results are all between [0,1], and the prediction accuracy for different problems can be compared to determine which model is more suitable for predicting which problem. R^2 is the best indicator of LR. The formula for calculating the three indicators is as follows:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \widehat{y_i}| \tag{8}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y_i})^2} \quad \epsilon[0, +\infty) \tag{9}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \widehat{y_i})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad \epsilon[-\infty, 1) \tag{10}$$

In the formula, y_i is the actual value, $\widehat{y_i}$ is the predicted value, and n is the total number of samples. In this paper, three evaluation metrics, MAE, RMSE, and coefficient of determination (R^2), are used to evaluate the performance of the model. R^2 in statistics characterizes how well the regression equation explains the variation in the dependent variable and can be used to determine the degree of fit between the true and interpolated values, with closer to 1 indicating a better fit between the variables, that the model explains 100% of the variation in the target values.

3.2. Baseline model

To test the overall performance of the current model, the authors conducted a series of comparative experiments on two baseline models, the non-deep learning model, and the deep learning model.

(1) Non-deep learning baseline models. These include LR models, random forest (RF), ARIMA model, and SVR model.
(2) Deep learning baseline models. These include the temporal convolutional network (TCN) model, the gated recurrent unit network (GRU) model, the bidirectional gated recurrent unit network (BIGRU) model, and ST_ BIGRU_SVR model (the model proposed in this paper).

4. RESULTS AND DISCUSSION

4.1. Prediction performance

After determining the optimal network architecture for the current prediction task, the current ST-BiGRU-SVR model is trained using the training set until convergence and then evaluated on the test set. In this paper, the water temperature values of monitoring station S1 in Qinhuangdao, Hebei are predicted and the model predictions are compared with the actual values. The predicted next-moment water temperature values, and the observed water temperature values at site 1

in Qinhuangdao, Hebei, drawn using the ST-BiGRU-SVR model proposed in this paper are shown in Figure 5. As can be seen from the figure, the predicted values are generally consistent with the observed values. The R^2 values between observed and predicted data indicate that the model can capture 96% of the explained variance. The feasibility and accuracy of the proposed model were verified.

Also, the authors plotted the curves of the true and predicted values on the test set. The relationship between the true and predicted values of monitoring station S1 in Qinhuangdao, Hebei, is shown in Figure 6 (All values are at the same moment corresponding to the conditions). The authors observe from the graph that the two curves have approximately the same trend and a good fit. It is shown that the model proposed in this paper can accurately capture the spatial and temporal variation of water temperature and achieve relatively accurate prediction of future water quality (predicted future water temperature data).

4.2. Comparison of experiments

A comparison of the performance of the model proposed in this paper with the other seven baseline predictions is shown in Table 2 regarding the three evaluation metrics MAE, RMSE and R^2. It can be found that the deep learning baseline model performs better than the non-deep learning baseline model, where the ARIMA non-deep learning baseline model performs

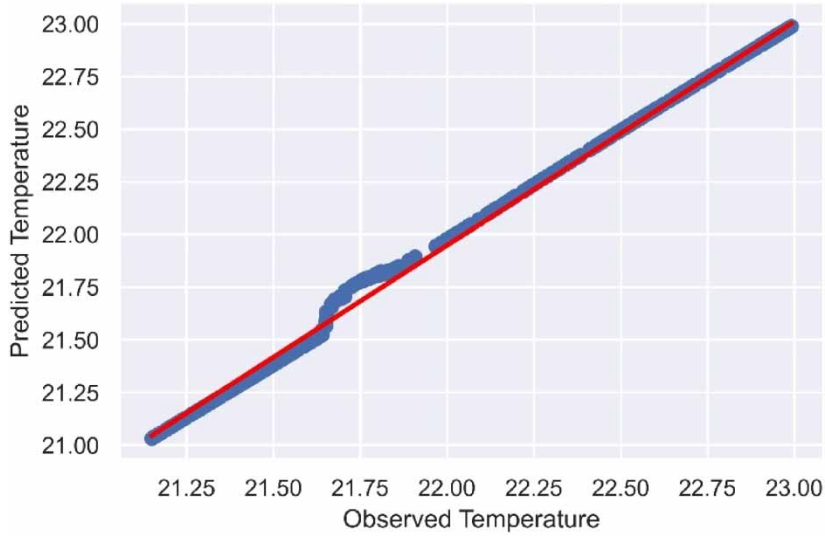


Figure 5 | Scatter plot of observed and true values of water temperature.

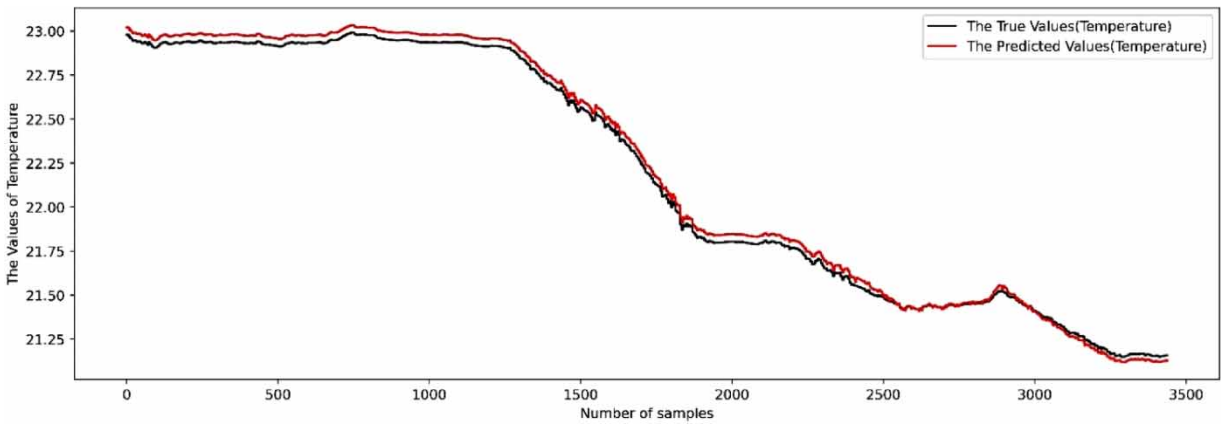


Figure 6 | Fitting curve of actual and predicted values of water temperature.

Table 2 | Comparison of model metrics for each baseline model with ST-BiGRU-SVR

Baseline model	MAE	RMSE	R^2
LR	0.202	0.260	0.831
RF	0.112	0.171	0.868
ARIMA	1.63	2.620	0.581
SVR	0.124	0.186	0.845
TCN	0.081	0.094	0.903
GRU	0.212	0.084	0.916
BIGRU	0.078	0.071	0.922
ST-BIGRU-SVR	0.071	0.076	0.957

the worst. Comparing the three evaluation metrics of all the deep learning baseline models, it is found that the proposed model performs best in terms of prediction performance and has the lowest model error. The reason for the high prediction accuracy of the ST-BIGRU-SVR proposed in this paper may be that the water temperature of the surrounding monitoring station influences the water temperature of the current station S1 to be predicted, while other models only consider the water temperature data of the current station to be predicted. Because the water temperature data are affected by the water temperature of the surrounding stations and the time attributes of the water temperature context, all the influencing factors are used as the input of the ST-BIGRU-SVR model, and the predicted water temperature data of the S1 station is used as the model output. ST-BIGRU-SVR model, in essence, the water temperature data of neighboring stations with higher correlation with S1 of the predicted site is determined as the input of the model by calculating the distance between the monitoring stations and the water temperature correlation coefficient between the monitoring stations. According to the temporal characteristics of water temperature data, BiGRU is used to process the water quality data at different times, mine the time series information of water quality data, and make full use of the time series characteristics of water quality data. Aiming at the nonlinear characteristics of water quality datasets, the SVR model is used to transform the nonlinear regression problem into a LR problem in high-dimensional space, because the SVR model has the advantages of global optimality, simple structure, strong generalization ability, etc., which is very suitable for the prediction of nonlinear data, and the core idea of SVR is to find a curve in high-dimensional space to represent the relationship between input data and output data and obtain the desired predicted value through the curve function. Therefore, the ST-BiGRU-SVR model predicts better than the general non-deep learning baseline model (LR, RF, ARIMA, SVR) and deep model (TCN, GRU, BiGRU).

5. CONCLUSIONS AND OUTLOOK

Currently, most reservoirs and intelligent aquaculture systems mainly use sensor-based IoT systems to monitor water quality in real-time, and this method of real-time monitoring has a lag. If water quality can be accurately predicted for a long period time in the future, it will allow water quality monitors to propose measures to prevent water pollution in advance so that farmers can take measures in advance to effectively counteract farming risks and improve production efficiency. However, the current trend of water quality is mainly based on regular manual surveys and monitoring by inspectors, farmers' long-term accumulated experience in speculation, monitoring methods have a strong subjectivity, poor reliability, and poor timeliness.

For the long-term prediction of water quality parameters, this paper proposes a hybrid neural network-based future water quality spatiotemporal prediction model. The model achieves more accurate and stable future prediction by integrating historical time water quality data, and nearest neighbor water quality data into the model. After pre-processing the data with GAN interpolation, we also use the Pearson correlation coefficient to analyze the correlation of water quality parameters. Finally, we input the prior information and preprocessed data into the constructed model for training. The experimental results show that the model proposed in this paper has a higher prediction accuracy compared to the non-depth model and the depth model. Specifically, this paper proposes a long-term prediction method for water quality parameters with a prediction accuracy of 95.7%, which makes the effectiveness of model in the paper. The proposed model still has some limitations: (1) water quality parameters by a variety of complex factors, however, due to equipment funding issues, failure to obtain more parameters of the site, the future can be integrated with other parameters of water quality, meteorological factors,

etc., to improve predictive performance. (2) In addition, to make the water quality prediction model more robust and further reduce the impact of water quality changes on the model prediction, a variety of factors such as seasonal changes and water quality changes can be incorporated into the deep neural network as a priori information, so that the prediction model can obtain longer-term prediction results. (3) The model proposed in the paper was only evaluated on the dataset of the Qinhuangdao watershed in Hebei, which has some limitations. In the future, it is hoped that more monitoring data (such as dissolved oxygen, chemical oxygen demand, ammonia nitrogen, etc.) from other watershed water quality monitoring sites can be collected to further validate the performance of the model.

ACKNOWLEDGEMENTS

This work was financially supported by the Dalian Science and Technology Innovation Fund Project: Liupanshui Citizens' Drinking Water Source Water Quality Monitoring and Early Warning System Construction(2020JJ27SN106).

DATA AVAILABILITY STATEMENT

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

CONFLICT OF INTEREST

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Rongli Gai reports financial support was provided by Department of Science and Technology of Liaoning Province.

REFERENCES

Akoko, G., Le, T. H., Gomi, T. & Kato, T. 2021 A review of SWAT model application in Africa. *Water* **13** (9), 1313.

Bauwe, A., Eckhardt, K.-U. & Lennartz, B. 2019 Predicting dissolved reactive phosphorus in tile-drained catchments using a modified SWAT model. *Ecohydrology & Hydrobiology* **19** (2), 198–209.

Chen, W., Xu, H., Chen, Z. & Jiang, M. 2021 A novel method for time series prediction based on error decomposition and nonlinear combination of forecasters. *Neurocomputing* **426**, 85–103.

Deng, T., Chau, K.-W. & Duan, H.-F. 2021 Machine learning based marine water quality prediction for coastal hydro-environment management. *Journal of Environmental Management* **284**, 112051.

Fu, Y., Hu, Z., Zhao, Y. & Huang, M. 2021 A long-Term water quality prediction method based on the temporal convolutional network in smart mariculture. *Water* **13** (20), 2907.

Haghiabi, A. H., Nasrolahi, A. H. & Parsaie, A. 2018 Water quality prediction using machine learning methods. *Water Quality Research Journal* **53** (1), 3–13.

Haq, K. R. A. & Harigovindan, V. P. 2022 Water quality prediction for smart aquaculture using hybrid deep learning models. *IEEE Access*, **10**, 60078-60098

Haq, KP Rasheed Abdul, and V. P. Harigovindan. "Water Quality Prediction for Smart Aquaculture using Hybrid Deep Learning Models." IEEE Access (2022).

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.

Huang, H., Zhang, Z., Lin, Z. & Liu, S. 2022 Hourly water demand forecasting using a hybrid model based on mind evolutionary algorithm. *Water Supply* **22** (1), 917–927.

Li, W., Hsu, C.-Y. & Hu, M. 2021 Tobler's First Law in GeoAI: a spatially explicit deep learning model for terrain feature detection under weak supervision. *Annals of the American Association of Geographers* **111** (7), 1887–1905.

Li, W., Wei, Y., An, D., Jiao, Y. & Wei, Q. 2022 LSTM-TCN: Dissolved oxygen prediction in aquaculture, based on combined model of long short-term memory network and temporal convolutional network. *Environmental Science and Pollution Research* **29** (26), 39545–39556.

Liu, J., Yu, C., Hu, Z., Zhao, Y., Bai, Y., Xie, M. & Luo, J. 2020 Accurate prediction scheme of water quality in smart mariculture with deep Bi-S-SRU learning network. *IEEE Access* **8**, 24784–24798.

Mundu, M. M., Nnamchi, S. N., Ukagwu, K. J., Peter, B. A., Nnamchi, O. A. & Ssempewo, J. I. 2022 Numerical modelling of wind flow for solar power generation in a case study of the tropical zones. *Modeling Earth Systems and Environment* **8** (3), 4123–4134.

Quan, Q., Hao, Z., Xifeng, H. & Jingchun, L. 2020 Research on water temperature prediction based on improved support vector regression. *Neural Computing and Applications* **34** (11), 8501–8510.

Shen, L., Lu, J., Geng, D. & Deng, L. 2020 Peak traffic flow predictions: exploiting toll data from large expressway networks. *Sustainability* **13** (1), 260.

Su, X., He, X., Zhang, G., Chen, Y. & Li, K. 2022 Research on SVR water quality prediction model based on improved sparrow search algorithm. *Computational Intelligence and Neuroscience* **2022**.

Valadkhan, D., Moghaddasi, R. & Mohammadinejad, A. 2022 Groundwater quality prediction based on LSTM RNN: an Iranian experience. *International Journal of Environmental Science and Technology*, 1–12.

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, M., Cheng, Q., Huang, J. & Cheng, G. 2020 Analysis of the European stock market's advance response time to COVID-19 based on Pearson correlation Coefficient. In *2020 3rd International Conference on Algorithms, Computing and Artificial Intelligence*.

Wu, E. M.-Y., Kuo, S.-L. & Liu, W.-C. 2012 Generalized autoregressive conditional heteroskedastic model for water quality analyses and time series investigation in reservoir watersheds. *Environmental Engineering Science* **29** (4), 227–237.

Wu, X., Zhou, J., Yu, H., Liu, D., Xie, K., Chen, Y., Hu, J., Sun, H. & Xing, F. 2021 The development of a hybrid wavelet-ARIMA-LSTM model for precipitation amounts and drought analysis. *Atmosphere* **12** (1), 74.

Yan, J., Gao, Y., Yu, Y., Xu, H. & Xu, Z. 2020 A prediction model based on deep belief network and least squares SVR applied to cross-section water quality. *Water* **12** (7), 1929.

Yan, J., Liu, J., Yu, Y. & Xu, H. 2021 Water quality prediction in the luan river based on 1-DRCNN and bigru hybrid neural network model. *Water* **13** (9), 1273.

Zhou, S., Song, C., Zhang, J., Chang, W., Hou, W. & Yang, L. 2022 A hybrid prediction framework for water quality with integrated W-ARIMA-GRU and LightGBM methods. *Water* **14** (9), 1322.

First received 1 December 2022; accepted in revised form 26 April 2023. Available online 7 June 2023