A novel operational water quality mobile prediction system with LSTM-Seq2Seq model

Xie, Lizi; Zhao, Yanxin; Fang, Pan; Cheng, Meiling; Chen, Zhuo; Wang, Yonggui

DOI
10.1016/j.envsoft.2024.106290

Publication date
2024

Document Version
Final published version

Published in
Environmental Modelling and Software

Citation (APA)
Xie, L., Zhao, Y., Fang, P., Cheng, M., Chen, Z., & Wang, Y. (2024). A novel operational water quality mobile prediction system with LSTM-Seq2Seq model. *Environmental Modelling and Software*, *185*, Article 106290. https://doi.org/10.1016/j.envsoft.2024.106290

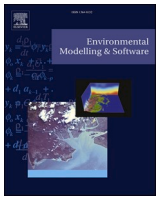
Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Environmental Modelling and Software 185 (2025) 106290

Contents lists available at ScienceDirect

Environmental Modelling and Software

journal homepage: www.elsevier.com/locate/envsoft



A novel operational water quality mobile prediction system with LSTM-Seq2Seq model

Lizi Xie [a,b], Yanxin Zhao [c], Pan Fang [d], Meiling Cheng [e], Zhuo Chen [a,b], Yonggui Wang [a,b,*]

[a] Hubei Key Laboratory of Regional Ecology and Environmental Change, School of Geography and Information Engineering, China University of Geosciences, Wuhan, China
[b] State Key Laboratory of Hydraulic Engineering Intelligent Construction and Operation, Tianjin University, China
[c] United Center for Eco-Environment in Yangtze River Economic Belt, Chinese Academy for Environmental Planning, Beijing, 100012, China
[d] Institute of Advanced Studies, China University of Geosciences, Wuhan, China
[e] Department of Geoscience and Remote Sensing, Delft University of Technology, Delft, Netherlands

ARTICLE INFO

Keywords:
Water quality
LSTM-seq2seq
Time step
Transfer application

ABSTRACT

An adequate water quality prediction mobile system is crucial for real-time, proactive, and convenient water environment monitoring through mobile devices to reduce or prevent water environmental threats. After exploring the feasibility and superiority of the LSTM-seq2seq model for predicting various water quality indicators, the optimal time step range for different length predictions was proposed. To verify the generalizability and reusability of the model, the performance differences of migrating models was investigated. Based on the entire process, we have developed a cost-effective, widely applicable, and sustainable operational prediction system framework. It was successfully applied in the Huangshui River Basin for two years. Results indicated that the model can achieve an *NSE* of above 0.5 for indicators with high coefficient of variation and above 0.75 for more stable indicators. When carrying out transfer applications, the model can achieve an *NSE* performance of above 0.5 for most sites in short to medium-term forecasting.

1. Introduction

Water resources are fundamental for achieving sustainable social and economic development. However, over the past 100 years, rapid urbanization and industrialization have resulted in significant water pollution and deterioration of water quality. Research indicates that nearly 80% (4.8 billion) of the global population faces a high level of water security threat in the early 21st century (Vörösmarty et al., 2010). It also reveals that water pollution seriously endangers human health and ecological environmental protection, becoming a global challenge (Gad et al., 2021). For a prolonged period, we have been employing passive measures for these water pollution issues. That is, addressing pollution only where it occurs. A large number of applications have demonstrated that preventing and controlling pollution before it occurs can effectively safeguard the water environment and lower treatment costs (L. Li et al., 2024; Liu et al., 2020; Wang et al., 2018; Zhi et al., 2024). Thus, it is of great significance to take more proactive measures to monitor the water environment and implement early pollution control, such as water quality prediction and early warning.

From a policy perspective, since the 1980s, developed countries have deliberately shifted the predominant approach to water environmental governance to watershed water quality target management (He et al., 2020) and have advocated for improving the capacity to prevent water environment crisis events by advancing the development of risk early warning systems (Alfieri et al., 2012). Additionally, increasing the availability of and giving access to early warning systems is one of the main targets given by the Sendai Framework for Disaster Risk Reduction 2015–2030 (Mosimann et al., 2024). In response to policy-driven initiatives, regional and national-level water quality monitoring and prediction systems have made significant advancements. Early water quality prediction systems were initially based on laboratory analysis, such as the online monitoring system of the River Trent in the UK (Drage et al., 1998), which could not provide real-time prediction and warning information, resulting in slow response times for formulating countermeasures. Afterward, advancements in digital technologies have made it possible to use online water quality sensors for real-time monitoring of

* Corresponding author. Hubei Key Laboratory of Regional Ecology and Environmental Change, School of Geography and Information Engineering, China University of Geosciences, Wuhan, China.
E-mail address: wangyg@cug.edu.cn (Y. Wang).

https://doi.org/10.1016/j.envsoft.2024.106290
Received 24 April 2024; Received in revised form 15 November 2024; Accepted 3 December 2024
Available online 9 December 2024

water quality, and these water quality sensors provide continuous data streams that enable timely detection and response to water quality issues (Scarbrough et al., 2023). However, deploying a sufficient number of sensors to monitor the water quality of the entire network poses significant challenges in cost and practical application (Z. Li et al., 2024; Storey et al., 2011). Process-based hydrodynamic and water quality models can serve as soft sensors, predicting water quality through expert estimation and mathematical description of physical, biogeochemical, and hydrochemical processes, thereby providing water systems monitoring and playing a crucial role in operational prediction systems (Najah et al., 2012; Tiyasha et al., 2020; Li et al., 2024). For instance, the regional water quality warning system developed by the Ohio River Valley Water Sanitation Commission (ORSANCO) (GULLICK et al., 2004), the Qiantang River water pollution emergency dynamic warning system (Song et al., 2013), the urban drinking water quality warning and control system framework (Hou et al., 2013), and the mobile environmental decision support system for ungauged basins (MEWSUB) (Wang et al., 2015), to name a few. Ranging from 1D, 2D to 3D formulations, several types of software and tools exist with varying complexities and capabilities (Mohammed et al., 2022). It is worth noting that the applicability and generality of these models are limited. According to (Herath et al., 2021), for large watersheds characterized by significant spatial heterogeneity, while lumped models may yield accurate results, the inferences (predictions) derived from these models may be unreasonable or unrealistic. Furthermore, distributed models exhibit limited generalizability and encounter issues related to over-parameterization. For example, EFDC can be used for very large watersheds or waterbodies, but there is a trade-off in ability to model or predict a small region within the large system accurately, while some models (AQUATOX and SWMM) are suitable for use in small-scale applications, extrapolating to larger areas may result in significant inaccuracies (Keller et al., 2023). In other words, different watersheds require suitable models for modeling and predicting, which may also involve optimization and calibration of some unobserved model parameters (Beven, 2000), all of which increase the complexity and reduce the generality of the operational prediction system using process models.

Fortunately, with the increasing computational performance and the continuous expansion of related datasets, deep learning methods have shown broad application prospects in water environmental management systems. This is primarily attributed to its easy-to-use programming languages, efficient algorithms, and ability to rapidly and accurately solve nonlinear and highly stochastic prediction problems through dynamic and adaptive model adjustments (Wan et al., 2022; Liu et al., 2022). Additionally, some studies have leveraged this advantage and integrated physical mechanisms to enhance the interpretability and adaptability of water environment predictions (Cai et al., 2024; Chadalawada et al., 2020; Zhan et al., 2024). However, the approach of physically guided deep learning is still in the exploratory and developmental phase, lacking the maturity required for widespread application and posing significant challenges to scientific innovation and interdisciplinary collaboration. Consequently, the immediate priority remains to improve the efficient application of deep learning models within operational systems, further optimizing their computational performance and stability to ensure their feasibility and reliability in real-world applications. Researchers have already utilized specific types of neural networks in early warning systems to forecast water quality a few hours in advance and demonstrated excellent prediction accuracy and reliability (Jin et al., 2019). For instance, Long Short Term Memory (LSTM) which can not only capture long-term dependencies but also address the issues of gradient vanishing and exploding encountered by regular Recurrent Neural Networks (RNNs) (Hochreiter, 1998), has been widely used and confirmed the exceptional performance in predicting the dynamic changes and seasonal variations of water quality time series (Azrour et al., 2022; Docheshmeh Gorgij et al., 2023; Hu et al., 2019; Kouadri et al., 2022; Liu et al., 2019; Wang et al., 2017; Zamani et al., 2023; Zhou et al., 2018). Furthermore, some researchers have applied LSTM deep learning networks for multi-step ahead prediction in water quality operational systems, achieving higher prediction accuracy and lower time cost compared to standard time series prediction models such as ARIMA and traditional artificial neural networks (Mohammed et al., 2021; Rasheed Abdul Haq and Harigovindan, 2022; Zhao et al., 2020). In terms of multi-step ahead prediction for the model, there are four output strategies: direct strategy (DS), recursive strategy (RS), combined direct and recursive strategy (DRS), and multi-input and multi-output strategy (MIMO) (Zhou et al., 2023). In comparison to the first three models, which have high model complexities or error accumulation, the MIMO type consists of the sequence-to-sequence (seq2seq) models can better meet the business requirements of different output lengths due to its great flexibility in setting input and output lengths as well as the continuity of the output sequence (Jia et al., 2021). Therefore, LSTM-seq2seq was selected to develop an operational prediction system for water quality early warning and management.

However, some issues need to be taken into consideration during the application of LSTM-seq2seq. Firstly, currently, for the input of deep learning models, most researchers tend to choose multi-variate collaborative inputs to achieve higher accuracy (Hu et al., 2019; Li et al., 2023; Mohammed et al., 2021; Saeed et al., 2024; Wang et al., 2019). For the output, most research focuses on short-term prediction, with relatively limited research on long-term prediction (Chen et al., 2020). While in operational systems, the key objective is to achieve high-precision predictions in the most efficient, cost-effective, and streamlined manner. This involves exploring the intrinsic characteristics of univariate predictive indicators, optimizing the use of single-variable data resources, and addressing the varying prediction lengths requirements of decision-makers. Secondly, recent studies have begun exploring the application of deep learning models in the environmental domain through transfer learning by utilizing model transferability assessments or data post-processing techniques, based on multi-site data (Sangiorgio and Guariso, 2024; Zhou, 2020). Particularly from a practical perspective, we should also consider how to achieve the economic value of reusing deep learning models and how to implement model transfer applications conveniently and effectively after spending some time training a deep learning model. Lastly, with the rise of the 5G wave, it is worth considering how to make it more convenient for managers and users to access water quality data and enjoy water quality prediction and early warning services anytime and anywhere, just like checking the weather forecast.

In response to these issues, the objectives of this research are to 1) build univariate LSTM-seq2seq models under different input-output conditions and explore the standard range of input lengths for short-term, medium-term, and long-term predictions; 2) investigate the feasibility and suitability of model transfer application; 3) construct a complete operational water quality prediction system based on WeChat mini-program. This study can promote the flexibility and convenience of input-output length selection for water quality prediction, achieve real-time, automatic, and reusable water quality prediction and warning in a lightweight, all-encompassing service approach, and provide a universal and effective system framework for water pollution prediction and early warning.

2. Materials and methods

2.1. Study area

The method and framework were tested in the Tuojiang River basin, and the transfer application and validation of results were conducted in the Huangshui River basin (the locations of these two basins are shown in Fig. 1).

The Tuojiang River is a primary tributary of the upper Yangtze River, passing through cities such as Chengdu, Deyang, Meishan, Neijiang, Yibin, Ziyang, and Zigong in Sichuan Province, joining the Yangtze River in Luzhou City, with a total length of 627.4 km and a drainage area

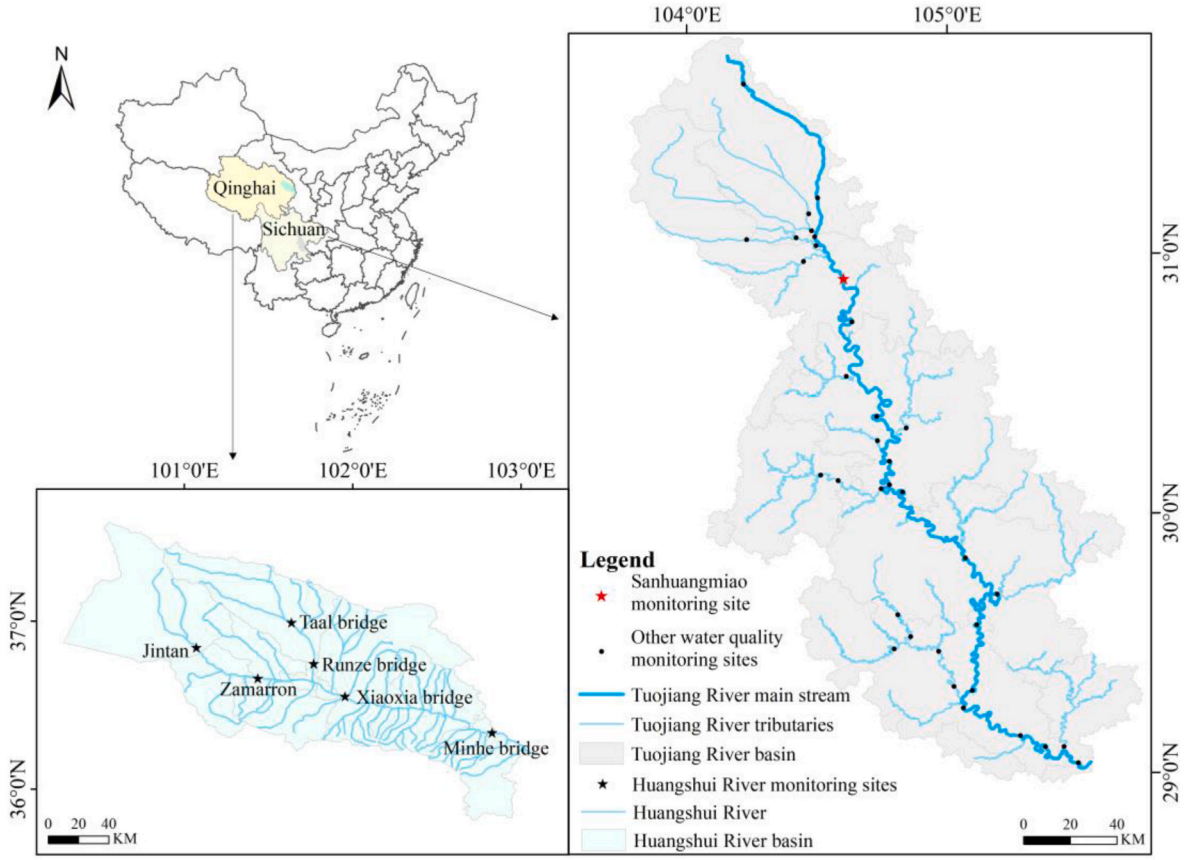


Fig. 1. Location of the tuojiang and huangshui river basin.

of 25576 km^2. The Tuojiang River basin supports 30% of the water supply and 26.2% of the population in Sichuan Province, with only 3.5% of the water resources, which has significant impacts on the ecological environment of Sichuan Province and even the water quality of the Yangtze River basin (Bai et al., 2021). The Huangshui River is a significant tributary of the upper Yellow River, often referred to as the "mother river" of Qinghai Province, with a total length of 374 km and a basin area of 17733 km^2. In recent years, with the rapid development of the Xining urban agglomeration, the water environment and aquatic ecology of the Huangshui River basin have been threatened (Wang et al., 2023). There is a significant discontinuity and a missing rate of up to 65% in the data from certain automatic water quality monitoring stations in the Huangshui River basin, which has led to challenges in effectively managing water resources and predicting and issuing early warnings for water quality. Therefore, if the model's migration application is successful in the Huangshui River basin, it will not only improve the adaptability and practicality of deep learning models but also offer a viable and effective approach for water quality prediction, early warning, and high-quality water resource management in data-scarce basins.

Considering that the Sanhuangmiao station is a crucial node where the upstream tributaries of the Tuojiang River converge with the mainstream, this study chose water quality sampling data from the Sanhuangmiao station for modeling. The water quality indicators considered in this study include water temperature (WT), pH value, dissolved oxygen (DO), turbidity (Tur), electrical conductivity (EC), permanganate index (COD_{Mn}), and ammonia nitrogen (NH_3-N). The modeling used the seven water quality indicator data with a frequency of every 4 h from December 2015 to June 2023 at this station. The statistical indicators for each water quality data are shown in Table 1. Several preprocessing steps need to be applied to the original water quality data. Firstly, due to the inherent randomness and uncertainty in the process of acquiring water quality monitoring data, there may be a certain degree of false reports. The 3σ rule is used to identify abnormal values in the data, and any data beyond the (μ-3σ, μ+3σ) interval are transformed into NaN values. Secondly, for certain periods, the water quality data is sampled at a frequency of 1 h, and it needs to be down-sampled using the mean to create a complete time series. Finally, the nearest neighbor method is used to handle missing values, completing the preprocessing of the water quality monitoring data. After preprocessing, each indicator has 16409 data points, which are then divided into a training set (60%), validation set (20%), and test set (20%).

Table 1
Descriptive statistics of water quality indicators at Sanhuangmiao Station.

Indicator	Unit	Min	Max	Mean	Std	CV
WT	°C	2.000	34.200	18.563	5.636	30.36%
DO	mg/L	1.370	13.610	7.722	1.701	22.03%
EC	μS/cm	153.000	881.000	543.000	109.339	20.13%
pH	/	6.058	8.950	7.859	0.279	3.55%
Tur	NTU	0.600	732.200	49.211	85.652	174.05%
COD_{Mn}	mg/L	0.030	13.030	2.570	1.090	42.42%
NH_3-N	mg/L	0.002	2.114	0.352	0.308	87.57%

2.2. Methods

2.2.1. LSTM-seq2seq model

Long Short-Term Memory (LSTM) is a specialized recurrent neural network architecture that is capable of learning both long-term (static) and short-term (dynamic) dependencies in time series data (Zhou, 2020). It replaces the hidden layers of traditional neural networks with memory blocks containing hidden states and cell states. Each memory

block consists of four components: a forget gate, an input gate, an output gate, and a memory cell, as shown in Fig. S1. The constant error back-propagation within the LSTM memory cell can compensate for the long time delay in predicting time series data (Hochreiter and Schmidhuber, 1997). Additionally, the linear interactions among LSTM units enable the cell state to store invariant information for a long time, overcoming the gradient vanishing and exploding problems of traditional neural networks (Kratzert et al., 2018). The three gate structures regulate the flow of information in the memory cell. Among these, the forget gate controls the information that needs to be forgotten to prevent the unbounded growth of internal values (Gers et al., 1999). The input gate controls which input is updated to protect the contents in the memory unit from irrelevant inputs, while the output gate controls which elements of the memory cell are used to update the hidden state, preventing disruption of short-term memories when storing long-term memories (Hochreiter and Schmidhuber, 1997). Therefore, for predicting water quality data with characteristics such as periodicity, long-term dependency, and slow time-varying trends (Zou et al., 2020), LSTM demonstrates unlimited potential and extensive application prospects.

Using LSTM to predict the water quality sequence at time t, such as $X = \{x_1, x_2, x_3, \ldots, x_T\}$, as shown in Fig. S2. When the previous hidden state h_{t-1} and the input x_t at this step pass through the forget gate, a vector f_t between 0 and 1 is output using the *Sigmoid* activation function as described in Eq. (3). According to f_t, the selective deletion or retention of the previous time step's cell state C_{t-1} is performed, where 0 represents complete deletion and 1 represents complete retention. When passing h_{t-1} and x_t through the input gate, on the one hand, they go through the *Sigmoid* layer to determine whether the current input value should be retained in the state or updated, resulting in a vector i_t between 0 and 1, as shown in Eq. (4). On the other hand, a candidate vector $\widetilde{C}_t$ is obtained through a tanh layer, as shown in Eq. (5). The new information is selectively recorded in the cell state through the combined action of i_t and $\widetilde{C}_t$. The current memory cell C_t is updated by combining the cell state after forgetting the old cell information through the forgetting gate with the candidate cell state after passing through the input gate, as shown in Eq. (6), which finalizes the update of the cell state. When passing through the output gate, C_t is directly sent to the next unit. The vectors h_{t-1} and x_t are activated using the Sigmoid function to obtain the vector o_t, determining which parts of the memory cell C_t to output at this time, as shown in Eq. (7). After adjustment through the tanh function, the output h_t for this time step is obtained, as shown in Eq. (8). Eq. 1 and 2 represent the expressions for the *Sigmoid* function and the tanh function, respectively.

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

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

$$f_t = \sigma\left(W_f \bullet x_t + U_f \bullet h_{t-1} + b_f\right) \tag{3}$$

$$i_t = \sigma(W_i \bullet x_t + U_i \bullet h_{t-1} + b_i) \tag{4}$$

$$\widetilde{C}_t = tanh\left(W_g \bullet x_t + U_g \bullet h_{t-1} + b_g\right) \tag{5}$$

$$C_t = f_t \times C_{t-1} + i_t \times \widetilde{C}_t \tag{6}$$

$$o_t = \sigma(W_o \bullet x_t + U_o \bullet h_{t-1} + b_o) \tag{7}$$

$$h_t = o_t \times tanh\,(C_t) \tag{8}$$

Where $\times$ denotes element-wise multiplication; $\bullet$ denotes matrix multiplication; W and U are the weight matrices of the neural network; b is the bias vector of the neural network; σ and tanh are activation functions; f_t、 i_t、 $\widetilde{C}_t$、 o_t、 C_t、 h_t represent the forget gate, input gate, candidate cell, output gate, cell state, and hidden state at time step t, respectively; x_t represents the input sequence data.

The traditional LSTM, commonly employed for addressing N-step prediction problems, incorporates a fully connected layer into the output h_t (Wunsch et al., 2021). The number of units in the fully connected layer is set to n, indicating that the hidden state is mapped to an n-dimensional vector through a simple fully connected layer to align with the data format of the problem. At this point, the n-dimensional vector outputs are completely independent and do not influence each other. Therefore, when using LSTM for prediction, the sequence correlation between output labels cannot be considered (Mu et al., 2023). The Seq2seq model is a common framework for time series prediction. The encoder maps variable-length source sequences to a fixed-length vector to extract important features from the sequence, while the decoder maps these features or vector representations to variable-length target sequences (Fang et al., 2021; Kow et al., 2024). It considers the dependencies and correlations between sequences, solves the problems existing in the traditional LSTM mentioned above, and can process input and output at different time steps (Zhang et al., 2022), eliminates the limitation on the lengths of input and output sequences. Therefore, this paper selects the LSTM-seq2seq model as the deep learning model for predicting water quality.

The LSTM-seq2seq model integrates LSTM into the encoder and decoder of the seq2seq model. Its objective is to maximize the conditional probability of the target sequence based on the given input sequence, as shown in Eq. (9). The process, as illustrated in Fig. S2, involves the LSTM encoder transforming the state of the last time step of the input sequence into a fixed-length intermediate semantic vector v (which theoretically considering the input information of each previous time step), and then passing v to the LSTM decoder. The decoder takes the vector v as the initial state and decodes it step by step, ultimately mapping it to the target sequence (Zheng et al., 2024). This approach provides greater flexibility in handling input and output sequences of varying lengths, while also takes into account the correlations between output sequences, learns long-term dependency information, eliminates long-term time lags, and enhances the performance of water quality prediction models.

$$P(y_1, \ldots, y_m | x_1, \ldots, x_n) = \prod_{t=1}^{m} P(y_m | y_1, \ldots, y_{m-1}, v) \tag{9}$$

Where $(x_1, \ldots, x_n)$ is the input sequence; $(y_1, \ldots, y_m)$ is the corresponding output sequence; n is the input time steps, and m is the output time steps.

2.2.2. Autocorrelation analysis

Autocorrelation refers to the correlation of a sequence of data records with the records from a previous time period (referred to as the lag period), used to measure the impact of historical data on the current moment's data. Given a dataset X_N, where X_1、 X_2、 … 、 X_N correspond to data at times t_1、 t_2、 … 、 t_N, the autocorrelation function (ACF) with lag k is calculated as shown in Eq. (10) and illustrated in Fig. S3. ACF within the confidence interval indicates that the autocorrelation is not significant within the corresponding lag period and lacks statistical significance. The confidence interval (CI) widens as the lag increases, calculated as shown in Eq. (11).

$$ACF(k) = \frac{\sum_{i=1}^{N-k}(X_i - \overline{X})(X_{i+k} - \overline{X})}{\sum_{i=1}^{N}(X_i - \overline{X})^2} \tag{10}$$

$$CI = \pm z_{1-\alpha/2}\sqrt{\frac{1}{N}\left(1 + 2\sum_{i=1}^{k} ACF(k)^2\right)} \tag{11}$$

Where X_t is the original data set; X_{t+k} is the same data set that has undergone k-lagged displacements; $\overline{X}$ is the average value of the complete sequence of the original data set; N is the sample size; z is the cumulative

distribution function of the standard normal distribution; α is the significance level.

On the one hand, a strong autocorrelation indicates that the data sequence has a similar pattern with its future sequences, which may result in higher predictive accuracy. This paper will explore whether the features of the data itself directly or indirectly determine the training accuracy through the comparison of experimental results. On the other hand, the window size and autocorrelation function characterize the relationship between current data and historical or future data, so determining the window size based on the autocorrelation function is theoretically feasible. This study attempts to compare the experimental results of different input-output lengths with autocorrelation values to investigate whether the autocorrelation of the sequence can be used for input step selection and to determine the range of multiples of the input steps relative to the output steps.

2.2.3. Logical framework of the system

The general architecture of the Operational Water Quality Prediction Mobile System (WQPMS) developed in this study is shown in Fig. 2. This system framework is used for the automatic prediction and warning of water quality indicators, achieves the cyclical reuse of deep learning models, ensures the maximization of data resource benefits, provides the most economical and practical decision support, and emergency response for water quality management, and has been successfully applied in the management system of the Huangshui River basin.

In this system, the water quality data of the automatic monitoring station stored in the database will be preprocessed first. Then, the autocorrelation of the data sequence will be examined to determine if the data sequence is suitable for simulating and prediction using the LSTM-seq2seq model. If it passes, the system will retrieve the corresponding model from the model repository, conduct accuracy verification, and directly output the prediction results if the verification is

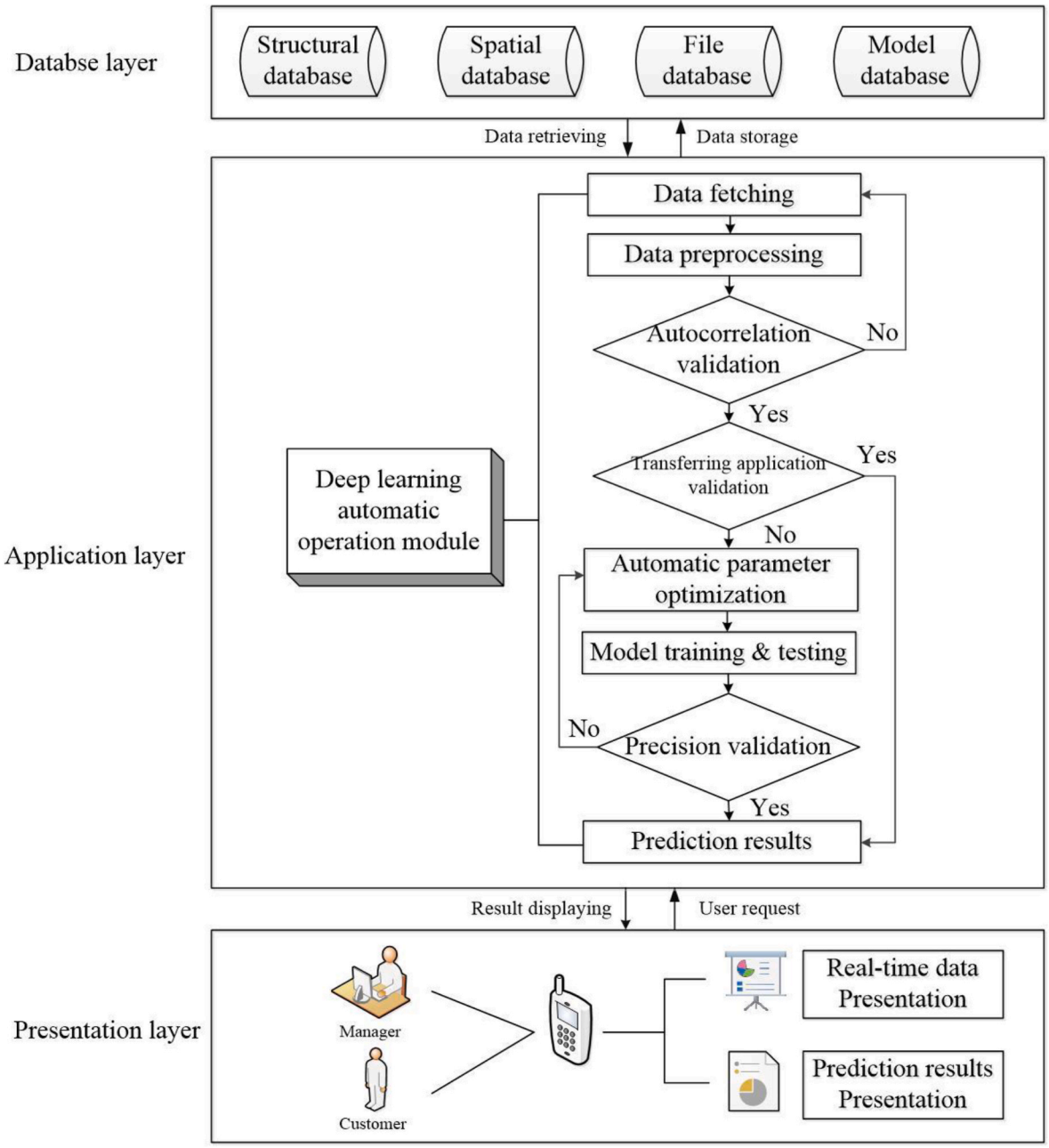


Fig. 2. Logical framework of the WQPMS.

successful; if not, the system will automatically perform parameter tuning, training, testing, and accuracy validation. It should be noted that automatic parameter adjustment can include input parameters such as input time step and hyperparameters such as learning rate and hidden layer size, where the parameter range of the input time step is determined based on standard range intervals obtained from the experiments in this article. The logical architecture of the system includes the presentation, application, and database layer, as shown in Fig. 2.

a) The presentation layer. The presentation layer receives user input, calls encapsulated methods through the application layer, and then visually displays the results processed by the application layer. Users are divided into administrators and ordinary users, and administrators can modify the model code through configuration files. It presents real-time monitoring information on water quality for various sections in tables and maps, and visualizes the statistical trend results of water quality predictions for various sections. Also, it displays the classification results of water quality monitoring data and prediction data based on the Surface Water Environmental Quality Standards (GB3838-2002), highlighting the data of water quality exceeding standards, allowing user interaction, and supporting section-based and time-based queries.
b) The application layer. It is the core component of the system. After the model has been identified through autocorrelation, transferring application or training as described above, the deep learning model is automatically invoked daily to predict water quality indicators at the scheduled time. It can also receive execution instructions from the presentation layer, read corresponding data from the database, simulate the prediction of water quality indicators, and store the simulation results in the corresponding database table.
c) The database layer. Centralized storage and management of structural database (such as water quality data, meteorological data, hydrological data, etc.), spatial database (such as water system data, administrative divisions, water functional zones, etc.), file database (such as statistical yearbooks, monitoring reports, environmental protection manuals, etc.), and model database (in H5 or tf format). Specifically, the structural database comprises raw water quality monitoring data collected from automatic monitoring stations, and preprocessed water quality data that has undergone operations such as data cleaning and handling missing values. Additionally, it includes simulated future water quality index data predicted by deep learning models at the application layer.

2.3. Experiment setups

This study aims to develop a straightforward, economical, and practical operational water quality prediction system framework that is applicable to different river basins. This involves quickly determining the optimal input time steps corresponding to the output time steps required for business needs, exploring the hidden relationship between the autocorrelation features of the data and the input time steps or model accuracy, and maximizing the utilization of the deep learning model. Therefore, this study does not involve modifying the structure of the LSTM network or achieving the highest accuracy of the LSTM model within the study area.

2.3.1. Model parameterization

The size of the hidden layer in the LSTM-seq2seq network is determined through random search within the search space of the number and size of hidden layers, searching 25 times to determine. The optimizer used in the model to minimize error and loss is Adam, which combines the advantages of the AdaGrad and RMSProp algorithms, can dynamically adjust the learning rate, requires fewer memory resources, converges faster, has high computational efficiency, and is easy to implement (Khan et al., 2020). The activation function of the Dense layer in the encoder and decoder uses *tanh*. The initial learning rate is set to 0.001, and if the loss value on the validation dataset does not decrease for several epochs, the learning rate will be automatically reduced. The loss function is a key component of the neural network, measuring the error between the model's predicted value and the true value, determining the update and optimization of the model parameters. It was demonstrated that the overall model performance is higher when using *NSE* as the loss function than when using *MSE* as the loss function (Kratzert et al., 2019), and the model selects a portion of *NSE* as the loss function (*NSE* will be detailed in 2.3.3), as shown in Eq. (12). If the batch size is too small, the loss function will be difficult to converge, and if it is too large, the model may exhibit partial bias, leading to a significant decrease in performance (Xiang et al., 2020). After conducting multiple experiments, the batch size has been set to 256, and the number of iterations (epochs) has been set to 600. Since the process of deep learning modeling involves randomness and uncertainty, each model is run 15 times. The proposed LSTM-seq2seq model is implemented in the TensorFlow framework using Python, and preprocessing and result analysis are conducted in Jupyter Lab.

$$Loss\ Function = \frac{\sum_{i=1}^{n}\left(y_i^o - y_i^s\right)^2}{\sum_{i=1}^{n}\left(y_i^o - \bar{y}\right)^2} \tag{12}$$

2.3.2. Input and output time steps setting

The time steps of the input and output sequence (window sizes) determine how far back the LSTM-seq2seq model will connect the data with its past values and will heavily affect the correctness of the model constraints. For example, a smaller window will result in the absence of dependency relationships within subsequences, while a larger window will significantly increase the model's computational complexity (Li et al., 2022). Therefore, choosing the appropriate window size is crucial. In operating systems, the output sequence length is generally determined based on managerial requirements, so selecting the corresponding optimal input time length becomes an important research problem that needs to be addressed urgently.

This study employs five different input time steps: 6, 18, 42, 90 180, corresponding to 24 h, 72 h, 168 h, 360 h, 720 h, respectively. It also uses four different output time lengths: 2, 6, 18, 42, corresponding to 8 h, 24 h, 72 h, 168 h in the future. The selection of progressive time steps aims to demonstrate the cascade of long-term dependency in time series, facilitating the exploration of the advantages of daily, weekly, or monthly-scale inputs and meeting the predicted needs of different short-, medium-, and long-term lengths.

2.3.3. Evaluation metrics

This study uses the Nash efficiency coefficient (*NSE*), the root mean square error (*RMSE*), and the normalized root mean square error (*NRMSE*) to evaluate the performance of the model. *NSE* is a normalized statistic that assesses the goodness of fit of the model by comparing the ratio of the simulated value sequence to the observed value sequence with the 1:1 line, and it is the best objective function to reflect the overall fit of the hydrological process (Servat and Dezetter, 1991). *NSE* ranges from (-∞, 1], and when *NSE* is less than 0, the model performance is considered worse than a model that only produces the mean value, indicating that the model performance is unacceptable. When *NSE* is between 0 and 1, the model performance is generally considered acceptable, and the closer it is to 1, the more accurate the model can predict water quality indicators, as shown in Eq. (13). However, *NSE* can lead to inaccurate evaluation results when the variance of the observed data is slight or close to zero, so *RMSE* and *NRMSE* are used for additional assessment. *RMSE* is an evaluation index that measures the proximity between observed and simulated values and is sensitive to outliers. Its calculation formula is shown in Eq. (14), with a range of [0, +∞), and the closer it is to 0, the more accurate the model prediction. *NRMSE* normalizes *RMSE*, allowing for comparison between different

datasets, as shown in Eq. (15).

$$NSE = 1 - \frac{\sum_{i=1}^{n}\left(y_i^o - y_i^s\right)^2}{\sum_{i=1}^{n}\left(y_i^o - \overline{y^o}\right)^2} \tag{13}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}\left(y_i^o - y_i^s\right)^2}{n}} \tag{14}$$

$$NRMSE = \frac{\sqrt{\sum_{i=1}^{n}\left(y_i^o - y_i^s\right)^2 \Big/ n}}{\overline{y^o}} \tag{15}$$

Where y_i^o is the i-th observed value of the evaluated indicator; y_i^s is the i-th simulated value of the evaluated indicator; $\overline{y^o}$ is the average value of the observed data of the evaluated indicator; n is the total number of observed data.

3. Results

3.1. Predictive performance analysis of LSTM-seq2seq models

3.1.1. The predictive performance on different water quality indicators

Fig. 3 and Fig. S4 show the distribution of *NSE* and *NRMSE* results of LSTM-seq2seq on the test set for each input-output time steps combination. Overall, there is an obvious pattern: with the increase of time steps, the model performance of all indicators shows a significant downward trend (*NSE* gradually decreases, *NRMSE* gradually increases). The degree of this decrease in accuracy varies with different indicators. Fig. 3 shows that all indicators except Tur can achieve a performance status with *NSE*>0.75 in short-term forecasts. However, across 15 repeated random experiments, the maximum precision achievable by each indicator varies, as does the probability of achieving excellent performance. Specifically, for WT, DO, EC, pH, COD_{Mn} and NH_3-N, with relatively small coefficients of variation, all *NSE* are above 0.8, with the majority even surpassing 0.9. For WT and DO, some random experimental results show discrete low values. For Tur indicator, with a difference of up to 731.6 between the maximum and minimum concentration values (as shown in Table 1), the precision decreases below the *NSE* = 0.75 line with increasing input step length. In mid to long-term forecasts, the performance gap between each indicator

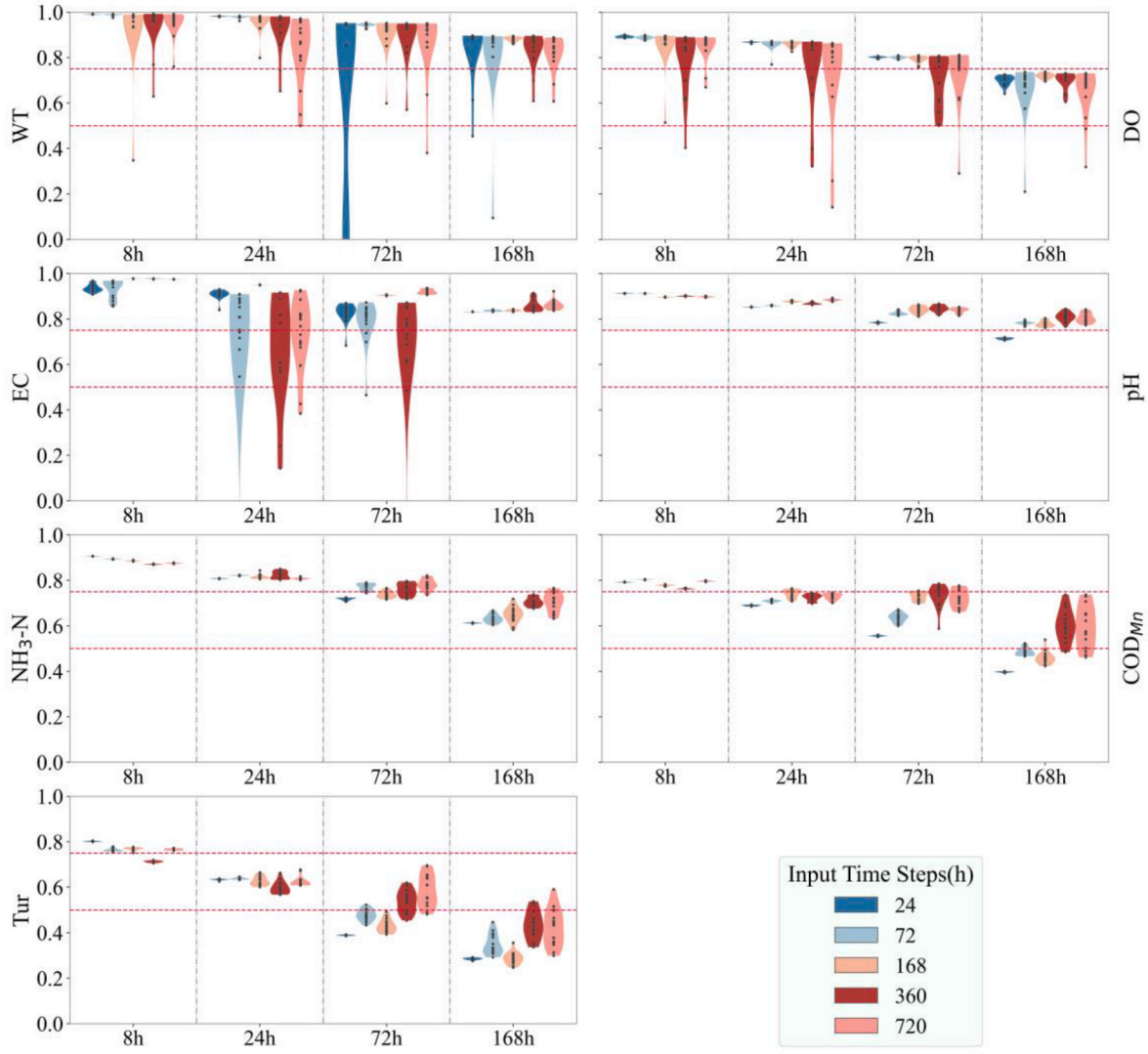


Fig. 3. Violin plots of *NSE* distribution for each indicator on the test set. The red lines represent the baselines of *NSE* = 0.75 and *NSE* = 0.5 where previous studies considered the model satisfactory when *NSE*≥0.5 and excellent when *NSE*≥0.75 (Xiang et al., 2020).

becomes increasingly significant. With an output time step of 168 h, *NSE* for WT, EC, and pH can still reach 0.8, or even around 0.9 under various input conditions, while the precision of DO and NH_3-N indicators decreases to below the *NSE* = 0.75 line, and for COD_{Mn} and Tur, when the input time step is less than the output time step, indicator precision can even drop below the 0.5 level. As we observe Table 1, we can see that the coefficients of variation for NH_3-N, COD_{Mn}, and Tur are the top three highest.

The above indicates that the distribution characteristics of the data itself are related to the performance of the model to a certain extent. In order to further explore this specific relationship and the differences in the performance of various indicators, this study statistically analyzed the relationship between model performance and the autocorrelation values of the data. As shown in Fig. 4, the Pearson correlation coefficient (r = 0.703) revealed a significant positive correlation between autocorrelation values and model accuracy, which means the higher the autocorrelation value, the higher the accuracy the indicator can achieve. It can be observed from Fig. 4 that each water quality indicator exhibits a specific clustering trend in its distribution on the graph, indicating that the autocorrelation values for each indicator at different window sizes are relatively concentrated, and the corresponding accuracies also fall within a range of relative clustering. Due to the decrease in autocorrelation values with increasing lag in the early stages and the variable speed and degree of reduction for different indicators, the disparity becomes more significant with longer lags, especially in long-term predictions. As depicted in Fig. 4, in long-term forecasting, the differences in autocorrelation values and distribution precision between different indicators are more pronounced and dispersed, while those between the same indicators are more concentrated. In particular, the accuracy of WT, EC, DO and pH in long-term predictions is far higher than the other three indicators, with their autocorrelation values also being relatively high. In contrast, Tur exhibits lower autocorrelation values across all window sizes in long-term predictions, and its predictive accuracy is less than ideal. It is evident from this that the inherent characteristics of the indicator, to some extent, determine the maximum accuracy achievable when the LSTM-seq2seq model predicts water quality indicators. In practice, the suitability of the model for predicting a certain indicator can be determined through this relationship before applying the LSTM-seq2seq model.

3.1.2. Comparison between LSTM-seq2seq and LSTM, lasso models

To evaluate model performance, we used Lasso regression model and the LSTM model without the seq2seq structure as the benchmark models. We trained both models under the same input-output step conditions, and the *NSE* results on the test set are shown in Table S1. For the Lasso model, as the input time step increases, the model can capture more collinear information from the data, leading to higher accuracy. However, when the input time step is shorter, the window size becomes smaller, resulting in more data blocks and higher data complexity. The model cannot identify collinear trends from the complex dataset, leading to lower accuracy. The most evident observation pertains to the WT indicator at an output time step of 168 h. When the input time step is set to 720 h, the model achieves a *NSE* of 0.89. However, as the input time step is reduced to 24 h, the model's NSE plummets to 0.28, and for the NH_3-N and Tur indicators, the model's NSE even falls below zero. In contrast, when the input time step changes, especially when the input time step is smaller than the output time step, the accuracy of LSTM and LSTM-seq2seq shows a more stable state. This indicates that the two models can better capture the randomness, nonlinearity, and multi-timescale variations of water quality data, extracting more helpful information from nonlinear data. For the LSTM model, the accuracy in short to medium-term prediction is not as good as LSTM-seq2seq. While with an increase in output time step, the accuracy of LSTM-seq2seq decreases, and LSTM shows a similar accuracy to LSTM-seq2seq in predicting the DO, EC, and pH indicators in long-term prediction. In certain cases, the *NSE* of LSTM is even slightly higher than that of LSTM-

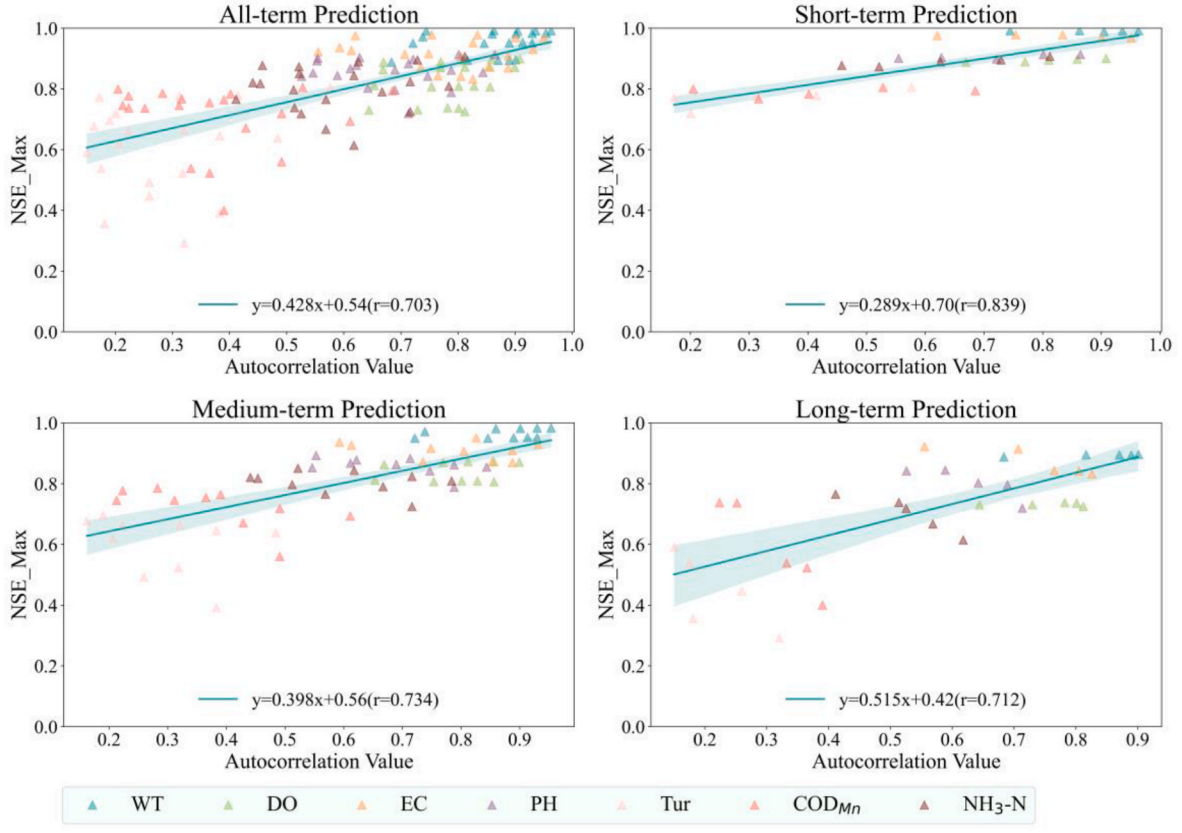


Fig. 4. The relationship between autocorrelation of different indicators and model performance under different forecast time steps.

seq2seq but does not exceed 0.1. For instance, When both the input and output time steps of the DO indicator are set to 168 h, the *NSE* of the LSTM model surpasses that of the LSTM-seq2seq model by approximately 0.057.

In order to more comprehensively compare the predictive performance of the three models, we will gradually display the prediction accuracy of each indicator at each output time step under 168h conditions for input and output, as shown in Fig. 5 and Fig. S5. It can be observed that both the Lasso model and the LSTM-seq2seq model exhibit a trend of decreasing accuracy with increasing step length, which aligns with the pattern discussed above. Furthermore, the LSTM-seq2seq model demonstrates better accuracy in predicting each indicator and the accuracy decreases more gradually with the time step than the Lasso linear regression model. For the LSTM model, in terms of the accuracy at each step, it is less stable and exhibits greater fluctuation than LSTM-seq2seq. For example, the difference between the highest and lowest *NSE* of the Tur indicator output sequence can reach 0.65, and the *NSE* of the last step of NH_3-N is as low as −4.407. Comparatively, the variations of LSTM-seq2seq are much more stable.

3.2. *Effect analysis of input time steps*

The window size is the sum of the input and output time steps. In operational systems, managers usually specify the desired output time step. Therefore, determining the most appropriate input time step range for this specified output length becomes crucial. The above indicates that the characteristics of the data itself, to some extent, determine the training accuracy. To explore whether the autocorrelation of the sequence can be used for the selection of input time steps, we compare the *NSE* and autocorrelation values of different input-output combinations of different indicators. Through Fig. S6, we can clearly observe that the autocorrelation values of all indicators exhibit a monotonically decreasing trend with increasing window size. For the same output time step, the model accuracy shows complex fluctuations with changes in the input time step rather than a monotonic increasing or decreasing trend. Therefore, selecting input time steps based on autocorrelation poses particular challenges.

To investigate the complex fluctuation phenomenon, we fitted the *NSE* means across 15 experiments with different input time steps of different indicators, as shown in Fig. 6. It can be observed that, under the same output steps, the accuracy change trend of different indicators at

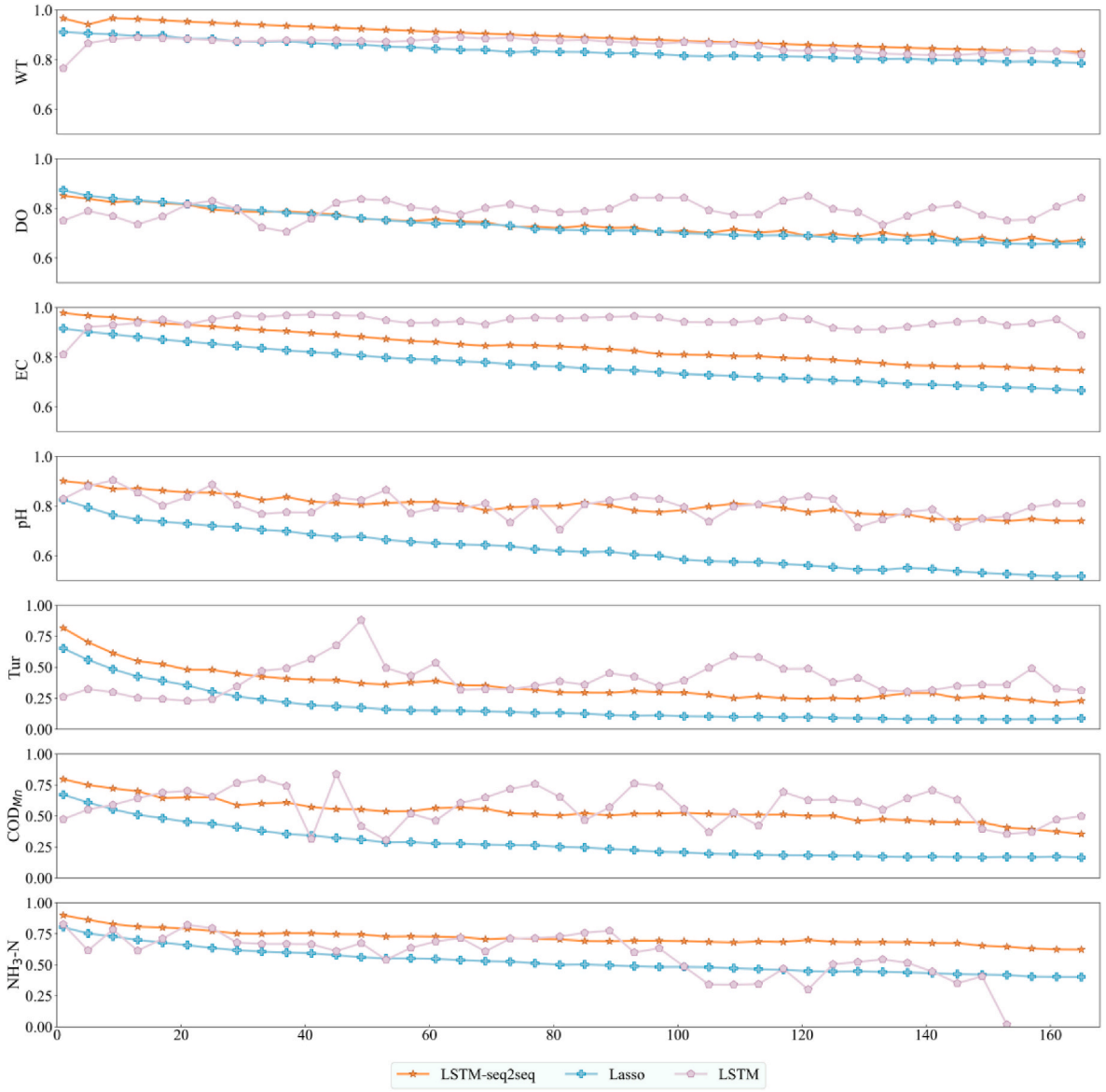


Fig. 5. The *NSE* accuracy at each output time step under 168h conditions for input and output.

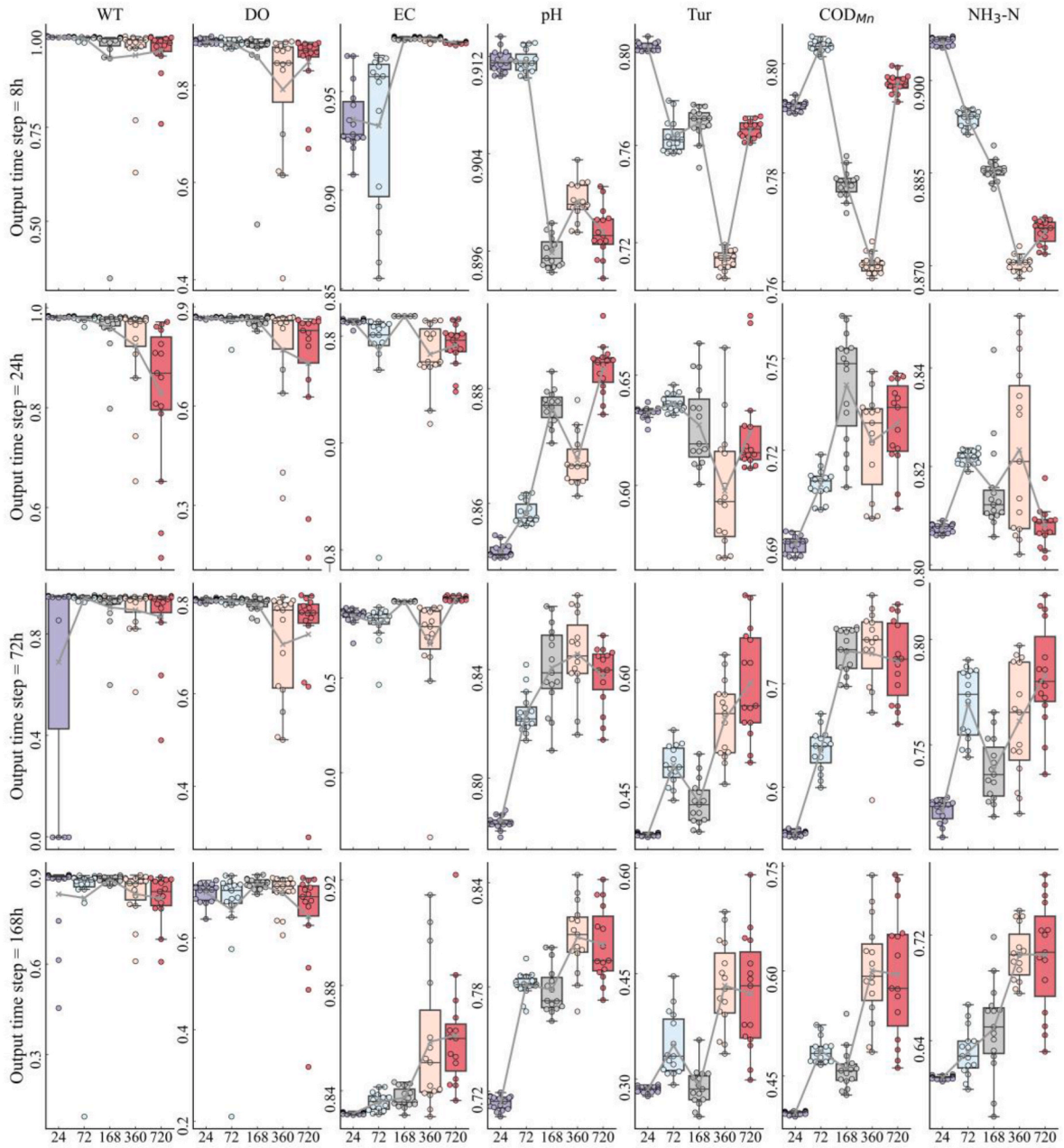


Fig. 6. The *NSE* change fitting plots at different input time step for various each indicator.

different input step sizes is basically consistent, which means that the input step has a specific common effect on the performance of different indicator models. For short-term prediction, the model accuracy initially decreases with the increase in input steps and rebounds after a particular time step. For example, when the input time step is 168 h, the *NSE* distribution decreases to the lowest for WT and pH, while for DO, Tur, NH_3-N, and COD_{Mn} with an input time step of 360 h. After the rebound, the accuracy distribution is more dispersed, and most of the highest values after the rebound are lower than the lowest values under the conditions of the previous few input steps, indicating a substantial reduction in model efficiency. For the EC indicator, the model's accuracy reaches the lowest point at an input time step of 72 h, and under subsequent input conditions, the accuracy becomes more concentrated and higher. Considering the high accuracy of the EC indicator under all input time steps, and the trends of most indicators, we believe that the optimal input time step for short-term forecasting is the time step at a daily scale within 72 h. For mid-term prediction, when the input time steps are 24 h and 720 h, the *NSE* distribution of WT is more scattered, leading to a decrease in model efficiency, and DO, EC show a decrease in efficiency when the input time step is 360 h. Indicators pH, Tur, NH_3-N, and COD_{Mn} exhibit lower accuracy at the input time step of 24 h, achieving a maximum accuracy lower than the minimum under neighboring step conditions. However, due to these four indicators' concentrated overall accuracy distribution, the difference in *NSE* is insignificant. For the input step of 360 h, the *NSE* distribution of these four indicators starts to become relatively scattered. Therefore, input time steps between 72 h and 168 h have a competitive advantage for mid-term forecasting. In long-term forecasting, the majority precision of the WT and DO in most random experiments is concentrated, with numerous shallow values at input time steps of 24 h and 720 h, affecting the distribution of *NSE*. The distributions of the other four indicators are more concentrated, with no shallow values. In 15 random experiments at each input time step, the precision at an input time step of 24 h is lower, reaching its mean or upper edge peak at 360 h and more scattered

at 720 h, with the median and lower quartile relatively low. Therefore, we consider the optimal time step for long-term forecasting to be between 72 h and 360 h, or approximately 0.5–2 times the output time step.

3.3. Assessment of model transfer application

We applied the model trained in the Tuojiang River basin to the Huangshui River basin, with the *NSE* and *NRMSE* results of each site and each index under different input-output time step combinations shown in Fig. 7 and Fig. S7, respectively. It is evident that the differences in accuracy under different conditions are significant. When the output time step increases under the same conditions as the station and input time step, the accuracy of the same indicators will decrease, and this is an inevitable result. For short-term forecasting, the *NSE* values of WT, DO, pH, and COD_{Mn} are all above 0.5. The accuracy at the Taal, Runze, and Zamarron sites is relatively low, but the accuracy can reach 0.75 or higher, with relatively low *NRMSE* values at the remaining three sites. The differences between EC, Tur, and NH_3-N across different sites are more pronounced. For instance, when the input time step is 72h, the *NSE* value for EC at the Minhe site is as high as 0.859, while at the Taal site, it is only 0.095. When predicting in the mid-term, the *NSE* of WT, DO, and pH can generally reach above 0.5; the accuracy of the EC indicator is relatively low at the Taal and Xiaoxia stations, with *NSE* around 0 and *NRMSE* greater than 1. Tur, COD_{Mn}, and NH_3-N have *NRMSE* greater than 1 at all stations, indicating that the error of the prediction model exceeds the range of the original data series, resulting in low accuracy. When making long-term predictions, the accuracy of various indicators notably decreases. While WT, Tur, COD_{Mn}, NH_3-N, and EC at some stations exhibit *NSE* higher than 0.5, the *NRMSE* at almost all stations exceeds 1.5. This suggests that the model effectively captures water quality variations, explaining a significant portion of the data's variance, but the predictive values are not sufficiently accurate. This discrepancy may be due to the presence of outliers or anomalous values in the water quality index sequence of the Huangshui River Basin compared to the data range of the Tuojiang River Basin.

In our study, the autocorrelation values significantly correlate with model accuracy mentioned in section 3.1.1. To validate this result and explain the differences in accuracy between different sites in transfer applications, we also conducted a statistical analysis on the relationship between autocorrelation of each water quality indicator series at various sites and *NSE*, as shown in Fig. 8. For overall term forecasts, the Pearson correlation coefficient between the two is 0.452, indicating a moderate correlation. It can be observed from the graph that the lower the autocorrelation coefficient of the same indicator at different sites, the relatively lower the accuracy at different sites. However, the accuracy of different indicators and autocorrelation values show varying rates of change with the time window, potentially leading to lower correlation coefficients. For instance, in short-term forecasts, as mentioned above, there is a clear difference between indicators such as WT and NH_3-N. The autocorrelation value and accuracy of the WT indicator are both around 0.8, whereas the accuracy of NH_3-N at the Jintan site is relatively concentrated around 0.4, with its autocorrelation value also relatively low and showing a more dispersed distribution with changes in the window, dropping to around 0. In mid-term forecasting, it can be observed that the autocorrelation values of WT and DO are generally higher than 0.4, with accuracies around 0.5. The pH accuracy is higher, but the autocorrelation values are more dispersed, ranging as low as 0. In long-term forecasting, the *NSE* of the NH_3-N indicator at the Jintan, Taal, and Zamarron sites is relatively low, around 0. The autocorrelation values at these sites are also relatively low, mostly less than 0.5, while the distribution of WT remains relatively concentrated.

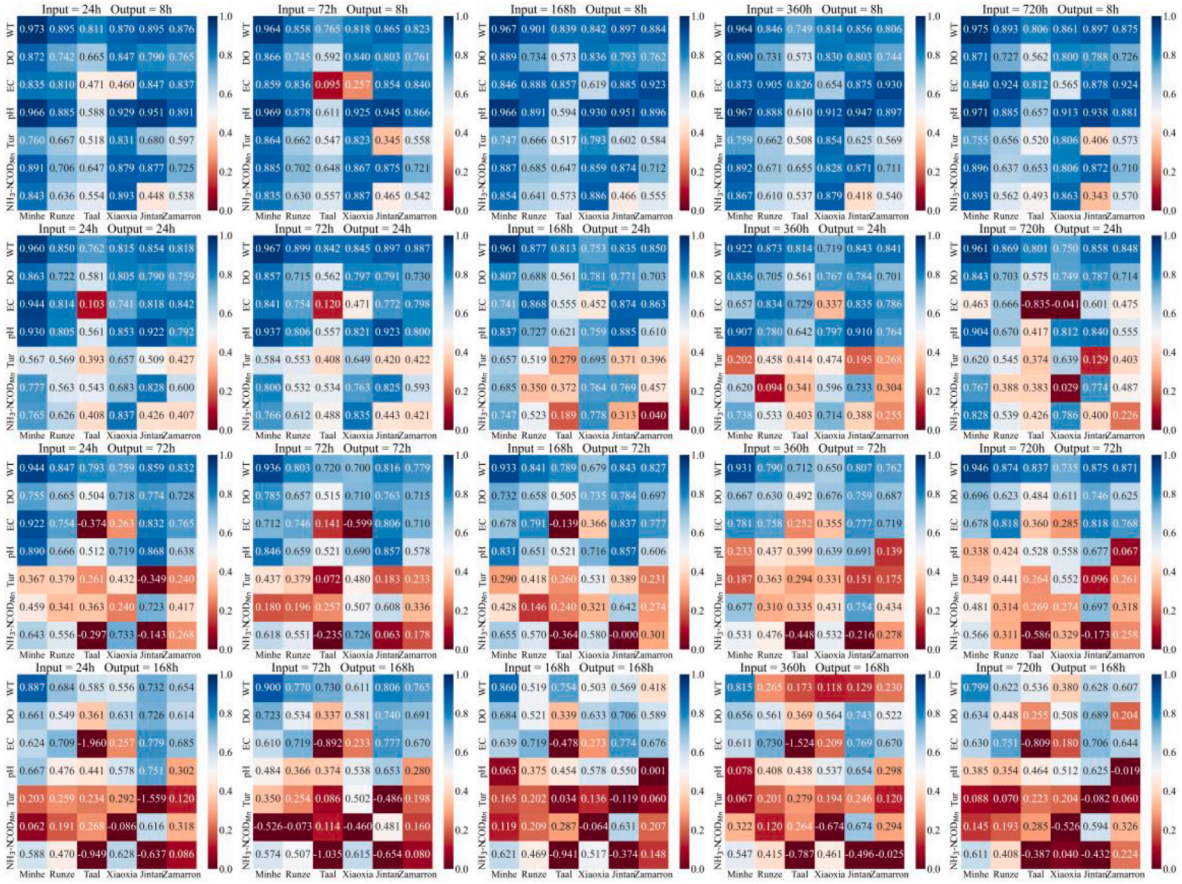


Fig. 7. *NSE* for migration applications for each indicator at each site in the Huangshui River.

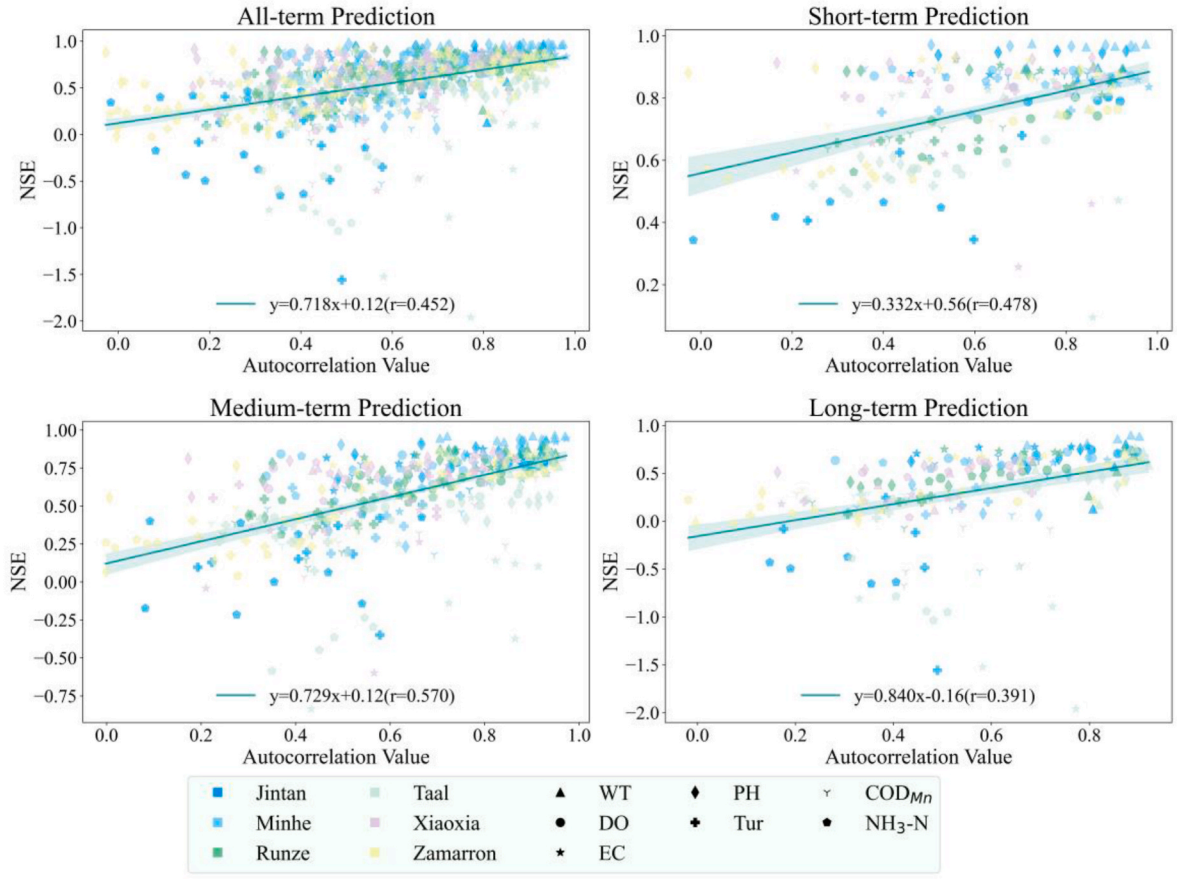


Fig. 8. The relationship between autocorrelation and model performance of migration applications under different forecast time steps.

Overall, although the correlation coefficient is not perfect, it can still indicate a specific relationship between model accuracy and the characteristics of the data itself and can also be used to assess the reasonability of the migration application using autocorrelation preliminarily.

3.4. Implementation of the system

3.4.1. System development

We applied the core concept of WQPMS to the water quality management process of the Huangshui River Basin, forming the Mobile Intelligent Integrated Management System for the Ecological Environment of the Huangshui Basin, as shown in Figs. 9–13. It has realized real-time, efficient monitoring, prediction, and water quality management in the Huangshui River Basin, providing higher-level technical support for management personnel decision-making. The system includes five primary functional pages.

a) Home page: The water quality warning information is the core of the prediction and warning system. The top of the system's homepage displays early warning message notification bar, as shown in Fig. 9. If the system predicts that the water quality exceeds the standard range in the Chinese Surface Water Quality Criteria (GB3838-2002) in the next seven days, it will display the information of the sites and the concentration of the indicators in the warning information bar. In order to allow users to conveniently observe the change trends of each indicator at each station in a certain time period, the system provides options for different sites, periods, and indicators, and users can visualize the water quality classification change trend, the proportion of different classifications, and the classification attainment status under user-selected conditions.

b) Monitor page: The monitor page of the system displays the water quality monitoring status of all sections of the entire Huangshui River basin in the form of a map, as shown in Fig. 10. The water quality status categorizes indicator concentrations into categories according to the Chinese Surface Water Quality Criteria (GB3838-2002). Additionally, the system provides various interactive options for user selection. Clicking "Filter" allows the selection of water quality indicators, time range, and section/station. Clicking "Map" can change the base map, including the satellite vector maps. Clicking "Basin" can show the range and boundary of the Huangshui River basin. Clicking "Stream" can display the water system distribution of the Huangshui River basin. Clicking "City" can show the cities involved in the Huangshui River basin. Clicking "Table" can display detailed water quality data for each section, including section number, section name, water quality category, and detection time. Clicking "Surrounding" can display livestock and poultry breeding, industrial enterprises, and sewage treatment plants for the entire basin, screen range, and current location. Clicking "Location" can locate the current position. Clicking the water quality category on the map can display detailed information, change trends, and forecast information for the station and can also set a manual buffer zone to query livestock and poultry breeding, industrial enterprises, and sewage treatment plants within the surrounding buffer zone.

c) Forecast page: For the convenience of users or administrators to view the forecast results, the system's prediction page displays specific concentration information and trend of changes for the site's future seven-day water quality forecast, as shown in Fig. 11. Users can choose different monitoring sites in the Huangshui River basin, and the page first displays the comprehensive water quality category, the water quality target category specified by the manager, and the water quality compliance status of the selected site. Next, the

Fig. 9. Home page of the WQPMS.

predicted concentrations of each indicator for the next seven days are displayed in text form. Finally, the trend of each indicator for the next seven days is displayed, and users can choose between viewing in the form of charts or tables. This type of statistics is conducted both on a daily basis and in various indicators, providing users with more diverse and comprehensive information.

d) Monitoring problems page: The Monitoring problems page summarizes all the detected exceedance information for display, as shown in Fig. 12. Users can search by time period and section, and the page will display refined exceedance information based on the selected criteria. This information includes section names, water quality classification, exceedance indicators and specific concentrations, exceedance time. Users can also navigate to the map to view spatial information on exceedance situations.

e) Messages page: The message subscription page can subscribe to different categories of information, including future seven-day water quality forecasts, automatic station water quality warnings, automatic station water quality monthly reports, and daily air quality warning reports. Subscription information will be pushed on fixed dates or at fixed times and can be viewed and managed on the message management page, as shown in Fig. 13.

3.4.2. *Water quality prediction performance*

Since the official launch in January 2022, the system has been automatically conducting water quality forecasts for the next 7 day at 00:00 every day. The forecast results are updated in the system at 12:00 every noon, displaying the status of each water quality indicator under the Environmental Quality Standards for Surface Water of China (GB3838-2002). Once the water quality at a particular section exceeds the standard, a warning will be issued on the homepage. The system operates stably overall, and some of the predictive effectiveness of the model are shown in Fig. 14 and Fig. S8. Notably, according to the WQPMS framework, the Tur indicator failed the autocorrelation test and lies outside the management requirements. Therefore, we do not present the model performance for the Tur indicator here. Additionally, based on the model transfer application effectiveness assessment, the WT, DO, EC, and pH indicators from the Jintan station, along with the WT, DO, EC, and NH_3-N indicators from the Minhe station, meet the precision requirements. We use existing models as predictive models for these parameters, while other indicators will undergo retraining.

For the Minhe site, all indicators, except for DO, demonstrate exceptional performance. It not only accurately depicts the changing trends and directions but also predicts extreme values, demonstrating the outstanding generalization and accuracy of the LSTM-seq2seq model to a certain extent. However, for the DO indicator, the model's prediction performance is not stable. The model predicts more accurately during periods of smooth change, whereas for time periods with more abrupt values, the model is able to predict the trend of change but not the extremes accurately. This issue may arise from uncertainties associated with applying transfer models. For the Jintan site, only the NH_3-N index exhibits unsatisfactory performance. The model could not accurately predict the extreme values of the NH_3-N index, even after we retrained the model using data from this site.

In conclusion, despite some dissatisfaction in predicting extreme values, this result is acceptable for long-term forecasting of migration applications, and this system framework provides a feasible approach to

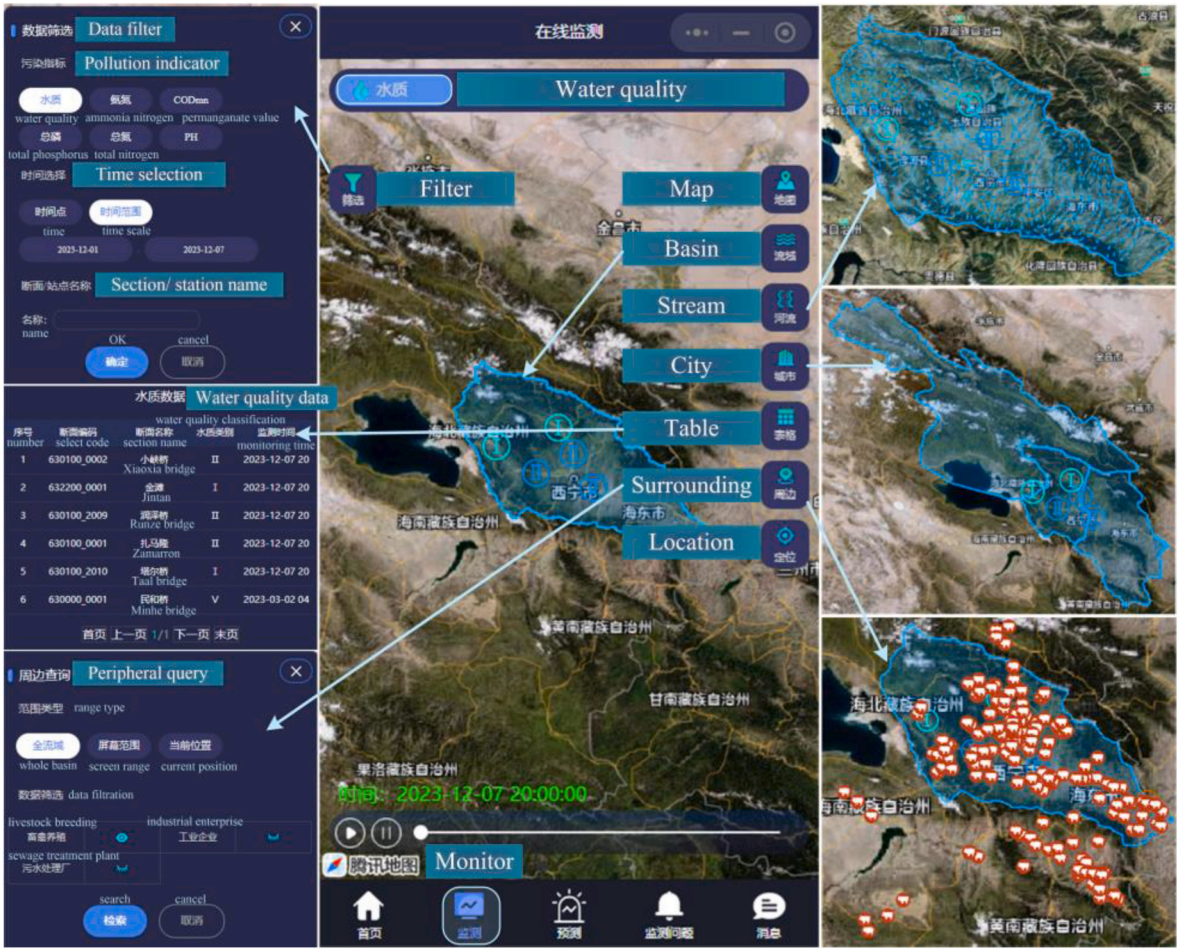


Fig. 10. Monitoring page of the WQPMS.

conveniently and efficiently utilize the advantages of artificial intelligence for effective water quality prediction and dramatically promotes the practical application of artificial neural networks in real life.

4. Discussion

4.1. The advantage of LSTM-seq2seq model in water quality prediction

To address the complexities of input variable requirements and the insufficient diversity in output lengths when applying existing deep learning models to operational systems, we developed univariate water quality prediction models based on the LSTM-seq2seq and explored the model's performance under various combinations of progressive input and output time step. Comparative analyses with linear Lasso model and LSTM model without the seq2seq structure demonstrated the superiority of the LSTM-seq2seq model in predicting water quality.

Compared to the Lasso model, LSTM-seq2seq possesses a relatively complex neural network and encoder-decoder structure, enabling it to capture complex non-linearity and non-stationary fluctuations in water quality sequences and maintaining relatively stable predictive accuracy even with shorter input time step. Relative to LSTM, on one hand, as described in section 3.1.2, LSTM-seq2seq and LSTM demonstrated similar results in predicting DO, EC, and pH indicators. We observed in Table 1 that DO, EC, and pH indicators have the lowest coefficient of variation. It indicates that even without using the seq2seq structure, LSTM can effectively forecast the trends and changes in these three indicators. In other words, LSTM excels in predicting stable water quality indicators, yet struggles with highly variable metrics. It has been indicated that the higher the degree of data dispersion and the presence of small values (close to 0) in the data, the higher the error caused by small changes, making it more difficult to predict (Wu et al., 2023). But after incorporating the seq2seq structure, the predictive accuracy of indicators with significant variations has been enhanced. This suggests the superiority of LSTM-seq2seq in handling more drastic changes in indicators. On the other hand, LSTM-seq2seq shows greater stability in predictions at each step, even though model performance may gradually decline as the output sequence length increases. This may be due to the seq2seq architecture, where the encoder and decoder process sequences rather than producing direct or recursive outputs. Consequently, this takes into account the continuity of the output sequence, resulting in more stable performance for predictions at each step (Zhou et al., 2023). Regarding the issue of diminishing accuracy, similar results have also been obtained in previous research and explained that this may be because the longer the sequence, the more uncertain noise in the data, which disturbs the model's ability to learn complex relationships from the sequence data and leads to reduced accuracy in predictions (Xiang et al., 2020; Chen et al., 2020; Sabzipour et al., 2023). In long-term forecasting, uncertainty factors will increase with the prediction length, leading to error accumulation and reduced prediction accuracy.

In summary, LSTM-seq2seq performs better in predicting indicators with more drastic changes and considers the continuity of the output sequence, ensuring more stability in predictions at each step. During our exploration of the aforementioned issues, we were surprised to discover the significant role that the intrinsic characteristics of data play in model predictions. The data exhibits a high coefficient of variation. Longitudinally, the model struggles to accurately predict extreme values due to an insufficient number of extreme sample data (Man et al., 2023). Horizontally, as discussed in section 3.1.1, the low interdependence

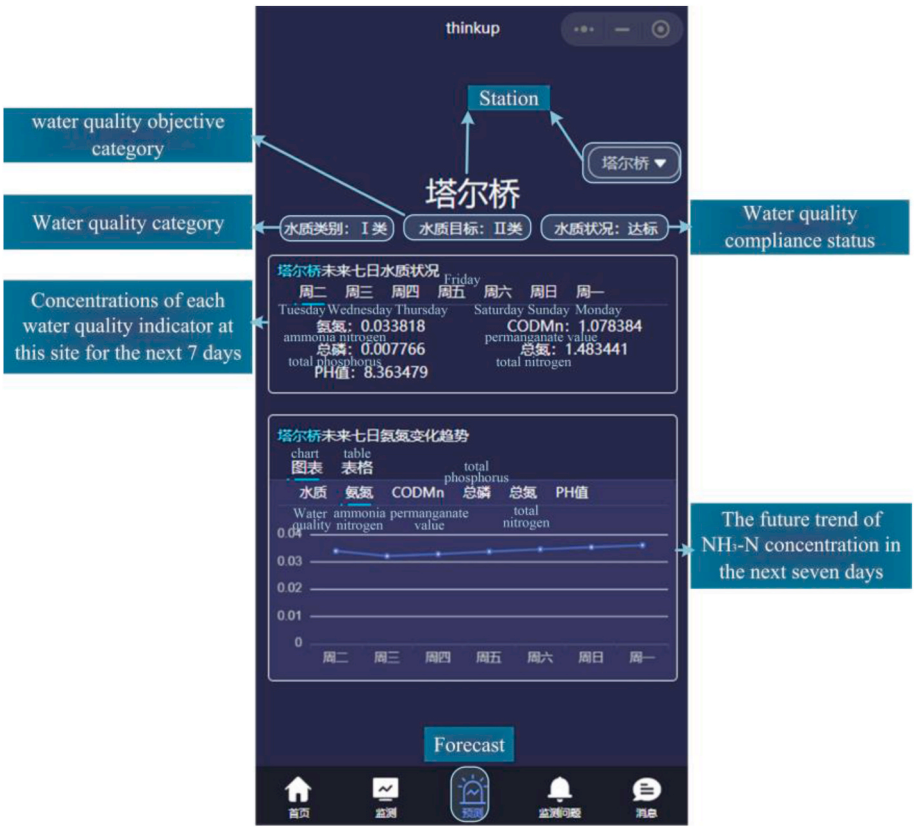


Fig. 11. Forecasting page of the WQPMS.

among data points hampers the model's ability to infer future trends from prior information. Prior research often involved preprocessing data using methods like smoothing or detrending (Zou et al., 2020), aiming to eliminate the impact of the data's inherent features. While this approach has its merits, it risks losing crucial information and complicating inverse transformations. Perhaps for modelers, it is crucial to leverage systematic principles and inherent rules of geospatial data to enhance models' extrapolation and adaptability, thereby transcending the limitations posed by the data itself. This represents a key focus for our future efforts.

4.2. Application of the WQPMS framework

Considering the practicality of deep learning models in operational systems, we propose a relatively simple and feasible framework for the cyclical reuse of models, as shown in Fig. 2. First, for selecting input steps, we analyzed the common patterns across 15 experiments for each of 20 different combinations using seven indicators. To be honest, the range of multiples we provide is approximate and does not need to be completely precise to several decimal places. We aim to provide researchers with a relatively small range and standard to facilitate quicker selection, whether using manual settings or automated optimization algorithms. Secondly, we assess the suitability of indicators for prediction by utilizing the autocorrelation features of the data. For time series data, the most prominent characteristic is temporal dependence, which indicates that the data from past states is correlated with the data from current or future states. The autocorrelation value serves as a quantitative measure of this relationship, determining the correlation of the series at various lags (Jónsdóttir and Milano, 2019). A high autocorrelation value indicates a strong correlation between data sequences, suggesting that the model can effectively leverage previously learned patterns to make accurate predictions about future states, thereby achieving a high level of predictive accuracy, as illustrated in Fig. S3 and Figs. 3 and 4. This bears a certain resemblance to the research conducted by (Recuero et al., 2019) on enhancing the classification accuracy of machine learning models through the use of autocorrelation values. Indeed, the accuracy of the model is influenced by various factors. This paper offers a relatively convenient criterion for assessing model suitability based on the common characteristics of the relationship between multi-indicator autocorrelation and accuracy. This enables managers to leverage the intrinsic characteristics of the data to estimate the potential accuracy of the model, thereby saving significant computational resources and time costs. Furthermore, for the target water quality sequence, we invoke the appropriate models from the model database under the corresponding conditions, conduct predictions, and evaluate prediction performance to achieve model transfer application. Transferring the trained model only takes a few seconds to predict the results, saving a lot of time and data resources and providing a feasible and effective method for water quality prediction and early warning in watersheds with insufficient data. However, as stated in section 3.3, the model of the migrating application generally exhibits excellent performance in short-term predictions, but as the output time step increases, the predictive performance of transfer learning models on indicators with low autocorrelation is suboptimal. This may explain the instability or inadequacy of transfer application models when predicting long-term outcomes with highly variable or weakly autocorrelated indicators. This is also why we have evaluated model suitability based on data autocorrelation from the outset. Ultimately, we established the WQPMS framework and successfully facilitated the development and operational functionality of the water quality prediction component within the Huangshui River management system. In light of the relatively poor predictive performance of the NH_3-N index at the Jintan site, it is noteworthy that, firstly, the data for the Huangshui River basin is scarce. The retraining of the NH_3-N model relied on just 6030 data points, which may have led to underfitting due to insufficient samples. Secondly, as mentioned in section 4.1, the NH_3-N concentration values

Fig. 12. Monitoring problems page of the WQPMS.

Fig. 13. Messages page of the WQPMS.

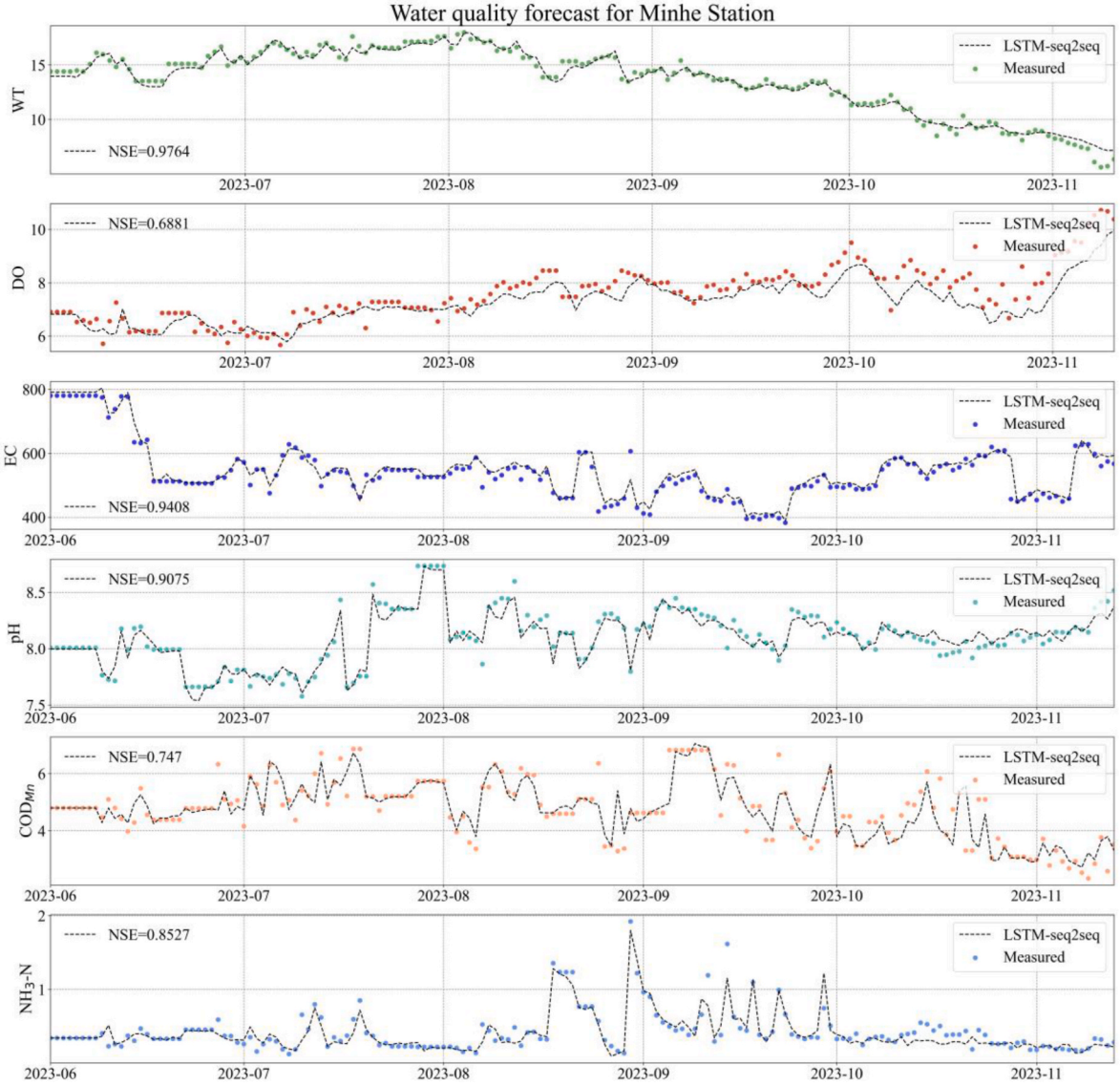


Fig. 14. Long-term prediction effectiveness of six indicators at Minhe Station from June to November 2023.

approach zero. Consequently, slight variations during data preprocessing and model training could yield significant changes in outcomes, such as outlier treatment, interpolation, normalization, and denormalization operations. Regarding the indexes that require retraining for other sites in the Huangshui River Basin, they have demonstrated excellent performance with less data. This further highlights the superiority of the LSTM-seq2seq model. As discussed in section 4.1, we anticipate further improvements in future models, resulting in reduced data requirements and more stable performance. We also look forward to the broader application of the WQPMS framework in other operational systems.

5. Conclusion

The accuracy and efficiency of water quality prediction and early warning system are crucially dependent on a simple, convenient, and reliable system processes and framework. In this study, we constructed the high-precision univariate water quality prediction model based on LSTM-seq2seq, proposed a reference standard range of input steps for short, medium, and long-term prediction by exploring the relationship between the input and output step lengths, and maximized the utilization of single-variable data resources. Also, we achieved the transfer and recurrent application of deep learning models. Finally, we integrated the entire method framework to form a more cost-effective, adaptable, and sustainable water environment prediction systems.

The comparison between experimental and transfer application results has demonstrated the effectiveness of the proposed system framework. Compared to Lasso (a linear regression model) and LSTM (a model without seq2seq structure), the LSTM-seq2seq model is capable of capturing the nonlinear features of complex time series and exhibits superior performance considering the continuity of output sequences. Applying the model transfer to the Huangshui River Basin also shows satisfactory performance in short-term forecasting *NSE*, which can result in significant savings in resources and time costs. However, the method framework still has some limitations and shortcomings. Fistly, the long-term predictive performance still needs improvement for indicators with poor stability, such as COD_{Mn} and Tur. Subsequent efforts can be made to improve the predictive effects by increasing the complexity of the model structure, by enhancing parameterization (Wang et al., 2016) or incorporating physical mechanisms (Maciąg et al., 2023; Tan et al., 2023; Zhan et al., 2024). Secondly, the universality of training models for some indicators still needs further improvement, and subsequent efforts can be made to enhance the interpretability of the model to improve its generalization ability further (Cai et al., 2024).

In conclusion, the system framework introduced in this study can be

used for water quality prediction and early warning at any monitoring site or section in need. It may be particularly beneficial for research teams with long-term monitoring plans, as it would form a complete and comprehensive database and model library, allowing for accurate predictions of short-term, medium-term, and long-term water quality forecasts.

CRediT authorship contribution statement

Lizi Xie: Writing – original draft, Visualization, Software, Methodology, Conceptualization. **Yanxin Zhao:** Supervision, Software, Investigation, Data curation. **Pan Fang:** Supervision, Software, Data curation. **Meiling Cheng:** Visualization, Software. **Zhuo Chen:** Writing – review & editing, Software. **Yonggui Wang:** Writing – review & editing, Supervision, Software, Conceptualization.

Software and data availability

Name of the software: WQPMS.
Developers: Authors of this paper.
Year first available: 2022.
Hardware Requirements: Smartphone.
Software Requirements: WeChat on Android 4.0 and above or IOS6 and above.
Programming Language: Python, Java.
Availability: The data used in this study can be obtained through web crawling the surface water quality module on https://www.mee.gov.cn/.The model training, prediction code, and the front-end and back-end code for the WeChat mini program can be obtained at https://github.com/Moontreenut/LSTM_seq2seq-for-water-quality.git.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was joint supported by the National Key R&D Program of China, China (No. 2022YFC3203502); Hubei Provincial Natural Science Foundation and Three Gorges Innovative Development Foundation of China, China (2024AFD371); Wuhan Knowledge Innovation Special Project, China (No. 2022020801020385); Science and Technology Research Project of Education Department of Hubei Province, China (B2023244); the Open Fund of State Key Laboratory of Hydraulic Engineering Intelligent Construction and Operation, China (HSEE-2311); Science and Technology Major Project of Hubei Province, China (2023BCA003).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.envsoft.2024.106290.

Data availability

The authors do not have permission to share data.

References

Alfieri, L., Salamon, P., Pappenberger, F., Wetterhall, F., Thielen, J., 2012. Operational early warning systems for water-related hazards in Europe. Environ. Sci. Pol. 21, 35–49. https://doi.org/10.1016/j.envsci.2012.01.008.

Azrour, M., Mabrouki, J., Fattah, G., Guezzaz, A., Aziz, F., 2022. Machine learning algorithms for efficient water quality prediction. Model. Earth Syst. Environ. 8, 2793–2801. https://doi.org/10.1007/s40808-021-01266-6.

Bai, H., Chen, Y., Wang, Y., Song, Z., Tong, H., Wei, Y., Yu, Q., Xu, Z., Yang, S., 2021. Contribution rates analysis for sources apportionment to special river sections in Yangtze River Basin. J. Hydrol. 600, 126519. https://doi.org/10.1016/j.jhydrol.2021.126519.

Beven, K.J., 2000. Uniqueness of place and process representations in hydrological modelling. Hydrol. Earth Syst. Sci. 4, 203–213. https://doi.org/10.5194/hess-4-203-2000.

Cai, H., Shi, H., Zhou, Z., Liu, S., Babovic, V., 2024. Explaining the mechanism of multiscale groundwater drought events: a new perspective from interpretable deep learning model. Water Resour. Res. 60, e2023WR035139. https://doi.org/10.1029/2023WR035139.

Chadalawada, J., Herath, H.M.V.V., Babovic, V., 2020. Hydrologically informed machine learning for rainfall-runoff modeling: a genetic programming-based toolkit for automatic model induction. Water Resour. Res. 56, e2019WR026933. https://doi.org/10.1029/2019WR026933.

Chen, Y., Song, L., Liu, Y., Yang, L., Li, D., 2020. A review of the artificial neural network models for water quality prediction. Appl. Sci. 10, 5776. https://doi.org/10.3390/app10175776.

Docheshmeh Gorgij, A., Askari, G., Taghipour, A.A., Jami, M., Mirfardi, M., 2023. Spatiotemporal forecasting of the groundwater quality for irrigation purposes, using deep learning method: Long Short-Term Memory (LSTM). Agric. Water Manag. 277, 108088. https://doi.org/10.1016/j.agwat.2022.108088.

Drage, B.E., Upton, J.E., Purvis, M., 1998. On-line monitoring of micropollutants in the river trent (UK) with respect to drinking water abstraction. Water Sci. Technol. 38, 123–130. https://doi.org/10.2166/wst.1998.0451.

Fang, Z., Crimier, N., Scanu, L., Midelet, A., Alyafi, A., Delinchant, B., 2021. Multi-zone indoor temperature prediction with LSTM-based sequence to sequence model. Energy Build. 245, 111053. https://doi.org/10.1016/j.enbuild.2021.111053.

Gad, M., Abou El-Safa, M.M., Farouk, M., Hussein, H., Alnemari, A.M., Elsayed, S., Khalifa, M.M., Moghanm, F.S., Eid, E.M., Saleh, A.H., 2021. Integration of water quality indices and multivariate modeling for assessing surface water quality in qaroun lake, Egypt. Water 13, 2258. https://doi.org/10.3390/w13162258.

Gers, F.A., Schmidhuber, J., Cummins, F., 1999. Learning to forget: continual prediction with LSTM. In: 1999 Ninth International Conference on Artificial Neural Networks ICANN 99. (Conf. Publ. No. 470). Presented at the 1999 Ninth International Conference on Artificial Neural Networks ICANN 99. (Conf. Publ. No. 470), vol. 2, pp. 850–855. https://doi.org/10.1049/cp:19991218.

Gullick, R.W., Gaffney, L.J., Crockett, C.S., Schulte, J., Gavin, A.J., 2004. Developing regional early warning systems FOR US SOURCE WATERS. J. Am. Water Works Assoc. 96, 68–82.

He, J., Wu, X., Zhang, Y., Zheng, B., Meng, D., Zhou, H., Lu, L., Deng, W., Shao, Z., Qin, Y., 2020. Management of water quality targets based on river-lake water quality response relationships for lake basins – a case study of Dianchi Lake. Environ. Res. 186, 109479. https://doi.org/10.1016/j.envres.2020.109479.

Herath, H.M.V.V., Chadalawada, J., Babovic, V., 2021. Genetic programming for hydrological applications: to model or to forecast that is the question. J. Hydroinf. 23, 740–763. https://doi.org/10.2166/hydro.2021.179.

Hochreiter, S., 1998. The vanishing gradient problem during learning recurrent neural nets and problem solutions. Int. J. Uncertain. Fuzziness Knowledge-Based Syst. 6, 107–116. https://doi.org/10.1142/S0218488598000094.

Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural Comput. 9, 1735–1780.

Hou, D., Song, X., Zhang, G., Zhang, H., Loaiciga, H., 2013. An early warning and control system for urban, drinking water quality protection: China's experience. Environ. Sci. Pollut. Res. 20, 4496–4508. https://doi.org/10.1007/s11356-012-1406-y.

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, 1420. https://doi.org/10.3390/s19061420.

Jia, P., Cao, N., Yang, S., 2021. Real-time hourly ozone prediction system for Yangtze River Delta area using attention based on a sequence to sequence model. Atmos. Environ. 244, 117917. https://doi.org/10.1016/j.atmosenv.2020.117917.

Jin, T., Cai, S., Jiang, D., Liu, J., 2019. A data-driven model for real-time water quality prediction and early warning by an integration method. Environ. Sci. Pollut. Res. 26, 30374–30385. https://doi.org/10.1007/s11356-019-06049-2.

Jónsdóttir, G.M., Milano, F., 2019. Data-based continuous wind speed models with arbitrary probability distribution and autocorrelation. Renew. Energy 143, 368–376. https://doi.org/10.1016/j.renene.2019.04.158.

Keller, A.A., Garner, K., Rao, N., Knipping, E., Thomas, J., 2023. Hydrological models for climate-based assessments at the watershed scale: a critical review of existing hydrologic and water quality models. Sci. Total Environ. 867, 161209. https://doi.org/10.1016/j.scitotenv.2022.161209.

Khan, A.H., Cao, X., Li, S., Katsikis, V.N., Liao, L., 2020. BAS-ADAM: an ADAM based approach to improve the performance of beetle antennae search optimizer. IEEECAA J. Autom. Sin. 7, 461–471. https://doi.org/10.1109/JAS.2020.1003048.

Kouadri, S., Pande, C.B., Panneerselvam, B., Moharir, K.N., Elbeltagi, A., 2022. Prediction of irrigation groundwater quality parameters using ANN, LSTM, and MLR models. Environ. Sci. Pollut. Res. 29, 21067–21091. https://doi.org/10.1007/s11356-021-17084-3.

Kow, P.-Y., Liou, J.-Y., Sun, W., Chang, L.-C., Chang, F.-J., 2024. Watershed groundwater level multistep ahead forecasts by fusing convolutional-based autoencoder and LSTM models. J. Environ. Manag. 351, 119789. https://doi.org/10.1016/j.jenvman.2023.119789.

Kratzert, F., Klotz, D., Brenner, C., Schulz, K., Herrnegger, M., 2018. Rainfall–runoff modelling using Long Short-Term Memory (LSTM) networks. Hydrol. Earth Syst. Sci. 22, 6005–6022. https://doi.org/10.5194/hess-22-6005-2018.

Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., Nearing, G., 2019. Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets. Hydrol. Earth Syst. Sci. 23, 5089–5110. https://doi.org/10.5194/hess-23-5089-2019.

Li, L., Knapp, J.L.A., Lintern, A., Ng, G.-H.C., Perdrial, J., Sullivan, P.L., Zhi, W., 2024. River water quality shaped by land–river connectivity in a changing climate. Nat. Clim. Change 14, 225–237. https://doi.org/10.1038/s41558-023-01923-x.

Li, Q., Yang, Y., Yang, L., Wang, Y., 2023. Comparative analysis of water quality prediction performance based on LSTM in the Haihe River Basin, China. Environ. Sci. Pollut. Res. 30, 7498–7509. https://doi.org/10.1007/s11356-022-22758-7.

Li, Y., Tong, Z., Tong, S., Westerdahl, D., 2022. A data-driven interval forecasting model for building energy prediction using attention-based LSTM and fuzzy information granulation. Sustain. Cities Soc. 76, 103481. https://doi.org/10.1016/j.scs.2021.103481.

Li, Z., Liu, H., Zhang, C., Fu, G., 2024. Real-time water quality prediction in water distribution networks using graph neural networks with sparse monitoring data. Water Res. 250, 121018. https://doi.org/10.1016/j.watres.2023.121018.

Liu, J., Wang, P., Jiang, D., Nan, J., Zhu, W., 2020. An integrated data-driven framework for surface water quality anomaly detection and early warning. J. Clean. Prod. 251, 119145. https://doi.org/10.1016/j.jclepro.2019.119145.

Liu, X., Lu, D., Zhang, A., Liu, Q., Jiang, G., 2022. Data-driven machine learning in environmental pollution: gains and problems. Environ. Sci. Technol. 56, 2124–2133. https://doi.org/10.1021/acs.est.1c06157.

Liu, Y., Zhang, Q., Song, L., Chen, Y., 2019. Attention-based recurrent neural networks for accurate short-term and long-term dissolved oxygen prediction. Comput. Electron. Agric. 165, 104964. https://doi.org/10.1016/j.compag.2019.104964.

Maciąg, P.S., Bembenik, R., Piekarzewicz, A., Del Ser, J., Lobo, J.L., Kasabov, N.K., 2023. Effective air pollution prediction by combining time series decomposition with stacking and bagging ensembles of evolving spiking neural networks. Environ. Model. Software 170, 105851. https://doi.org/10.1016/j.envsoft.2023.105851.

Man, Y., Yang, Q., Shao, J., Wang, G., Bai, L., Xue, Y., 2023. Enhanced LSTM model for daily runoff prediction in the upper huai River Basin, China. Engineering 24, 229–238. https://doi.org/10.1016/j.eng.2021.12.022.

Mohammed, H., Michel Tornyeviadzi, H., Seidu, R., 2022. Emulating process-based water quality modelling in water source reservoirs using machine learning. J. Hydrol. 609, 127675. https://doi.org/10.1016/j.jhydrol.2022.127675.

Mohammed, H., Tornyeviadzi, H.M., Seidu, R., 2021. Modelling the impact of weather parameters on the microbial quality of water in distribution systems. J. Environ. Manag. 284, 111997. https://doi.org/10.1016/j.jenvman.2021.111997.

Mosimann, M., Kauzlaric, M., Schick, S., Martius, O., Zischg, A.P., 2024. Evaluation of surrogate flood models for the use in impact-based flood warning systems at national scale. Environ. Model. Software 173, 105936. https://doi.org/10.1016/j.envsoft.2023.105936.

Mu, Y., Wang, M., Zheng, X., Gao, H., 2023. An improved LSTM-Seq2Seq-based forecasting method for electricity load. Front. Energy Res. 10, 1093667. https://doi.org/10.3389/fenrg.2022.1093667.

Najah, A.A., El-Shafie, A., Karim, O.A., Jaafar, O., 2012. Water quality prediction model utilizing integrated wavelet-ANFIS model with cross-validation. Neural Comput. Appl. 21, 833–841. https://doi.org/10.1007/s00521-010-0486-1.

Rasheed Abdul Haq, K.P., Harigovindan, V.P., 2022. Water quality prediction system based on Adam optimised LSTM neural network for aquaculture: a case study in Kerala, India. J. Inst. Eng. India Ser. B 103, 2177–2188. https://doi.org/10.1007/s40031-022-00806-7.

Recuero, L., Wiese, K., Huesca, M., Cicuéndez, V., Litago, J., Tarquis, A.M., Palacios-Orueta, A., 2019. Fallowing temporal patterns assessment in rainfed agricultural areas based on NDVI time series autocorrelation values. Int. J. Appl. Earth Obs. Geoinformation 82, 101890. https://doi.org/10.1016/j.jag.2019.05.023.

Sabzipour, B., Arsenault, R., Troin, M., Martel, J.-L., Brissette, F., Brunet, F., Mai, J., 2023. Comparing a long short-term memory (LSTM) neural network with a physically-based hydrological model for streamflow forecasting over a Canadian catchment. J. Hydrol. 627, 130380. https://doi.org/10.1016/j.jhydrol.2023.130380.

Saeed, A., Alsini, A., Amin, D., 2024. Water quality multivariate forecasting using deep learning in a West Australian estuary. Environ. Model. Software 171, 105884. https://doi.org/10.1016/j.envsoft.2023.105884.

Sangiorgio, M., Guariso, G., 2024. Transfer learning in environmental data-driven models: a study of ozone forecast in the Alpine region. Environ. Model. Software 177, 106048. https://doi.org/10.1016/j.envsoft.2024.106048.

Scarbrough, K., Persaud, P., Fletcher, I., Akin, A.A., Hathaway, J., Khojandi, A., 2023. Real-time sensor-based prediction of soil moisture in green infrastructure: a case study. Environ. Model. Software 162, 105638. https://doi.org/10.1016/j.envsoft.2023.105638.

Servat, E., Dezetter, A., 1991. Selection of calibration objective fonctions in the context of rainfall-ronoff modelling in a Sudanese savannah area. Hydrol. Sci. J. 36, 307–330. https://doi.org/10.1080/02626669109492517.

Song, X.X., Hou, D.B., Huang, P.J., Zhang, G.X., 2013. Research on dynamic early warning technology of water pollution emergency based on pollutant dispersion simulation. Appl. Mech. Mater. 316–317, 682–685. https://doi.org/10.4028/www.scientific.net/AMM.316-317.682.

Storey, M.V., van der Gaag, B., Burns, B.P., 2011. Advances in on-line drinking water quality monitoring and early warning systems. Water Res. 45, 741–747. https://doi.org/10.1016/j.watres.2010.08.049.

Tan, R., Hu, Y., Wang, Z., 2023. A multi-source data-driven model of lake water level based on variational modal decomposition and external factors with optimized bi-directional long short-term memory neural network. Environ. Model. Software 167, 105766. https://doi.org/10.1016/j.envsoft.2023.105766.

Tiyasha, Tung, T.M., Yaseen, Z.M., 2020. A survey on river water quality modelling using artificial intelligence models: 2000–2020. J. Hydrol. 585, 124670. https://doi.org/10.1016/j.jhydrol.2020.124670.

Vörösmarty, C.J., McIntyre, P.B., Gessner, M.O., Dudgeon, D., Prusevich, A., Green, P., Glidden, S., Bunn, S.E., Sullivan, C.A., Liermann, C.R., Davies, P.M., 2010. Global threats to human water security and river biodiversity. Nature 467, 555–561. https://doi.org/10.1038/nature09440.

Wan, H., Xu, R., Zhang, M., Cai, Y., Li, J., Shen, X., 2022. A novel model for water quality prediction caused by non-point sources pollution based on deep learning and feature extraction methods. J. Hydrol. 612, 128081. https://doi.org/10.1016/j.jhydrol.2022.128081.

Wang, P., Yao, J., Wang, G., Hao, F., Shrestha, S., Xue, B., Xie, G., Peng, Y., 2019. Exploring the application of artificial intelligence technology for identification of water pollution characteristics and tracing the source of water quality pollutants. Sci. Total Environ. 693, 133440. https://doi.org/10.1016/j.scitotenv.2019.07.246.

Wang, X., Zhang, J., Babovic, V., 2016. Improving real-time forecasting of water quality indicators with combination of process-based models and data assimilation technique. Ecol. Indicat. 66, 428–439. https://doi.org/10.1016/j.ecolind.2016.02.016.

Wang, Y., Ding, X., Chen, Y., Zeng, W., Zhao, Y., 2023. Pollution source identification and abatement for water quality sections in Huangshui River basin, China. J. Environ. Manag. 344, 118326. https://doi.org/10.1016/j.jenvman.2023.118326.

Wang, Y., Engel, B.A., Huang, P., Peng, H., Zhang, X., Cheng, M., Zhang, W., 2018. Accurately early warning to water quality pollutant risk by mobile model system with optimization technology. J. Environ. Manag. 208, 122–133. https://doi.org/10.1016/j.jenvman.2017.12.006.

Wang, Y., Zhang, W., Engel, B.A., Peng, H., Theller, L., Shi, Y., Hu, S., 2015. A fast mobile early warning system for water quality emergency risk in ungauged river basins. Environ. Model. Software 73, 76–89. https://doi.org/10.1016/j.envsoft.2015.08.003.

Wang, Yuanyuan, Zhou, J., Chen, K., Wang, Yunyun, Liu, L., 2017. Water quality prediction method based on LSTM neural network. In: 2017 12th International Conference on Intelligent Systems and Knowledge Engineering (ISKE). Presented at the 2017 12th International Conference on Intelligent Systems and Knowledge Engineering (ISKE). IEEE, Nanjing, pp. 1–5. https://doi.org/10.1109/ISKE.2017.8258814.

Wu, G., Zhang, J., Xue, H., 2023. Long-term prediction of hydrometeorological time series using a PSO-based combined model composed of EEMD and LSTM. Sustainability 15, 13209. https://doi.org/10.3390/su151713209.

Wunsch, A., Liesch, T., Broda, S., 2021. Groundwater level forecasting with artificial neural networks: a comparison of long short-term memory (LSTM), convolutional neural networks (CNNs), and non-linear autoregressive networks with exogenous input (NARX). Hydrol. Earth Syst. Sci. 25, 1671–1687. https://doi.org/10.5194/hess-25-1671-2021.

Xiang, Z., Yan, J., Demir, I., 2020. A rainfall-runoff model with LSTM-based sequence-to-sequence learning. Water Resour. Res. 56. https://doi.org/10.1029/2019WR025326.

Zamani, M.G., Nikoo, M.R., Rastad, D., Nematollahi, B., 2023. A comparative study of data-driven models for runoff, sediment, and nitrate forecasting. J. Environ. Manage. 341, 118006. https://doi.org/10.1016/j.jenvman.2023.118006.

Zhan, Y., Guo, Z., Yan, B., Chen, K., Chang, Z., Babovic, V., Zheng, C., 2024. Physics-informed identification of PDEs with LASSO regression, examples of groundwater-related equations. J. Hydrol. 638, 131504. https://doi.org/10.1016/j.jhydrol.2024.131504.

Zhang, Y., Ragettli, S., Molnar, P., Fink, O., Peleg, N., 2022. Generalization of an Encoder-Decoder LSTM model for flood prediction in ungauged catchments. J. Hydrol. 614, 128577. https://doi.org/10.1016/j.jhydrol.2022.128577.

Zhao, J., Wei, S., Wen, X., Qiu, X., 2020. Analysis and prediction of big stream data in real-time water quality monitoring system. J. Ambient Intell. Smart Environ. 12, 393–406. https://doi.org/10.3233/AIS-200571.

Zheng, Z., Ding, H., Weng, Z., Wang, L., 2024. Research on out-of-sample prediction method of water quality parameters based on dual-attention mechanism. Environ. Model. Software 176, 106020. https://doi.org/10.1016/j.envsoft.2024.106020.

Zhi, W., Appling, A.P., Golden, H.E., Podgorski, J., Li, L., 2024. Deep learning for water quality. Nat. Water. https://doi.org/10.1038/s44221-024-00202-z.

Zhou, J., Wang, Yuanyuan, Xiao, F., Wang, Yunyun, Sun, L., 2018. Water quality prediction method based on IGRA and LSTM. Water 10, 1148. https://doi.org/10.3390/w10091148.

Zhou, Y., 2020. Real-time probabilistic forecasting of river water quality under data missing situation: deep learning plus post-processing techniques. J. Hydrol. 589, 125164. https://doi.org/10.1016/j.jhydrol.2020.125164.

Zhou, Y., Li, Y., Wang, D., Liu, Y., 2023. A multi-step ahead global solar radiation prediction method using an attention-based transformer model with an interpretable mechanism. Int. J. Hydrogen Energy 48, 15317–15330. https://doi.org/10.1016/j.ijhydene.2023.01.068.

Zou, Q., Xiong, Q., Li, Q., Yi, H., Yu, Y., Wu, C., 2020. A water quality prediction method based on the multi-time scale bidirectional long short-term memory network. Environ. Sci. Pollut. Res. 27, 16853–16864. https://doi.org/10.1007/s11356-020-08087-7.