Environmental Science and Ecotechnology 20 (2024) 100402

Contents lists available at ScienceDirect

Environmental Science and Ecotechnology

journal homepage: www.journals.elsevier.com/environmental-science-and-ecotechnology/

Original Research

Water quality prediction based on sparse dataset using enhanced machine learning

Sheng Huang [a, b, c], Jun Xia [a, b, d, *], Yueling Wang [d, **], Jiarui Lei [c], Gangsheng Wang [a, b]

[a] State Key Laboratory of Water Resources Engineering and Management, Wuhan University, Wuhan 430072, China
[b] Institute for Water-Carbon Cycles and Carbon Neutrality, Wuhan University, Wuhan 430072, China
[c] Department of Civil and Environmental Engineering, National University of Singapore, 117578 Singapore
[d] Key Laboratory of Water Cycle and Related Land Surface Processes, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China

ARTICLE INFO

Article history:
Received 2 March 2023
Received in revised form
18 February 2024
Accepted 19 February 2024

Keywords:
Water quality modeling
Sparse measurement
River-lake confluence
Long short-term memory
Load estimator
Machine learning

ABSTRACT

Water quality in surface bodies remains a pressing issue worldwide. While some regions have rich water quality data, less attention is given to areas that lack sufficient data. Therefore, it is crucial to explore novel ways of managing source-oriented surface water pollution in scenarios with infrequent data collection such as weekly or monthly. Here we showed sparse-dataset-based prediction of water pollution using machine learning. We investigated the efficacy of a traditional Recurrent Neural Network alongside three Long Short-Term Memory (LSTM) models, integrated with the Load Estimator (LOADEST). The research was conducted at a river-lake confluence, an area with intricate hydrological patterns. We found that the Self-Attentive LSTM (SA-LSTM) model outperformed the other three machine learning models in predicting water quality, achieving Nash-Sutcliffe Efficiency (NSE) scores of 0.71 for COD_{Mn} and 0.57 for NH_3N when utilizing LOADEST-augmented water quality data (referred to as the SA-LSTM-LOADEST model). The SA-LSTM-LOADEST model improved upon the standalone SA-LSTM model by reducing the Root Mean Square Error (RMSE) by 24.6% for COD_{Mn} and 21.3% for NH_3N. Furthermore, the model maintained its predictive accuracy when data collection intervals were extended from weekly to monthly. Additionally, the SA-LSTM-LOADEST model demonstrated the capability to forecast pollution loads up to ten days in advance. This study shows promise for improving water quality modeling in regions with limited monitoring capabilities.

1. Introduction

Surface water is essential for global biogeochemical cycles related to environmental health and human well-being [1,2]. Due to the expansion of human activities (e.g., urbanization, agriculture, and dam construction), the impaired water quality of various surface water bodies (e.g., rivers and lakes) has been a major concern worldwide [3–5]. Many lakes and river networks are directly connected, affecting the accumulation and blend of pollutants in surface water bodies [6,7]. River-lake systems comprise a mosaic of lotic and lentic landscapes [8,9]. The mass exchange of dissolved constituents, sediment, and other loads between rivers and lakes is significantly influenced by the river-lake flow interactions within the system [10,11]. Compared with relatively stable lakes or rivers, the variation of pollution loads at the river-lake confluence is more complicated owing to the drastic hydrologic exchange regimes [12,13]. Therefore, it is challenging to model the river-lake pollutant load exchange for better-integrated surface water quality management [7,14].

Many mature hydrodynamic and water quality models can simulate and predict the advection and diffusion of pollutants under various complex flow regimes [15]. These process-based mechanistic models often require substantial parameter calibration work and detailed topographic data that are difficult to obtain [16–18]. In recent decades, machine learning has shown notable success due to its cost-effectiveness, robustness, high accuracy, and

* Corresponding author. State Key Laboratory of Water Resources Engineering and Management, Wuhan University, Wuhan 430072, China.
** Corresponding author.
E-mail addresses: xiajun666@whu.edu.cn (J. Xia), wangyl@igsnrr.ac.cn (Y. Wang).

https://doi.org/10.1016/j.ese.2024.100402

superiority in nonlinear problems, making it a powerful alternative to traditional water quality modeling techniques [19–22]. Increased frequency of machine learning research has been found in water quality applications of rivers and lakes [23–27]. Long short-term memory (LSTM), a variant of recurrent neural network (RNN), stands out in water quality modeling with architectural advances in sequence learning [28–30]. Recent work has shown that LSTM behaves well in many aspects, such as chlorophyll concentration prediction for algal management [31,32], water quality classification [33], and general water quality parameters or pollution loads simulation [34–36].

Machine learning is a data-driven approach that usually requires many data to train parameters for adequate generalization performance [30,37]. However, collecting and analyzing water quality samples is costly and commonly requires significant work compared to streamflow, which can be monitored on a daily-to-minute scale [35]. In many undeveloped regions or small and medium-sized rivers in the world, water quality data are typically sparse, monitored at a low frequency (e.g., monthly and weekly), and may even be partially missing owing to technology, cost, environmental conditions, or man-made mistakes [38–40]. Previous research on machine learning (e.g., LSTM) modeling of water quality was mainly focused on sufficient-data regions with well-equipped monitoring stations that can obtain continuous high-frequency data but less on insufficient-data regions with sparse water quality measurement [22]. For example, most studies employing LSTM are based on finer temporal scales, including intervals of 15 min [41], 1 h [32], and daily observations [29,31,34,36]. The application of weekly or monthly datasets is restricted to water quality classification or evaluation, as the sparsity of these data presents a significant challenge for water quality prediction [26,33]. A few studies have employed trend-decomposition mathematical methods for monthly-scale autoregressive prediction [110]. However, the autoregressive approach ignores upstream pollutant sources (i.e., boundary conditions). This limitation makes it difficult to apply to various projects (e.g., sudden pollution accidents, best management practices, pollution reduction scenario, and dam regulation) that require source-sink or upstream-downstream pollution modeling [113,114]. The potential of machine learning methods, like LSTM, for source-oriented surface water pollution control with sparse and partially missing data, remains to be better understood, especially in large water bodies with complex flow regimes.

Here, we proposed combining LSTM models and the load estimator (LOADEST) to model the pollution load with sparse water quality measurement. The major objectives of this paper are to (1) verify the effectiveness of the LSTM models combined with LOADEST in pollution load modeling with sparse water quality data; (2) test the application potential of the LOADEST-based machine learning methods with varying water quality sparsity; and (3) explore the pollution load forecasting ability with lead-time days using the optimal LSTM-LOADEST model. We conducted a case study in Dongting Lake, which connects to the Yangtze River, the largest river in China. We attempted to take advantage of machine learning and LOADEST for sparse water quality modeling, thereby facilitating the Yangtze River simulator that China is developing now [42,111].

2. Material and methods

2.1. Study area and data collection

Dongting Lake is the second largest freshwater lake naturally connected to the Yangtze River of China. The area of Dongting Lake covers about 2600 km^2, and its total volume is approximately 2.2×10^{10} m^3. Dongting Lake (28°30′–29°43′ N and 111°40′–113°10′ E) has a subtropical monsoon climate with drastic rainfall variation and distinct seasons. The annual precipitation ranges from 1100 to 1400 mm, and the annual average temperature is about 14–18 °C [43,44]. There are four basin-wide tributaries of Dongting Lake, namely the Xiang, Zi, Yuan, and Li Rivers. Their streamflow and water quality are measured at Xiangtan, Taojiang, Taoyuan, and Shimen stations, respectively (Fig. 1). In addition to the water from the tributaries, part of the Yangtze River water enters the lake through three main inlets, and then flow back into the Yangtze River via the Chenglingji outlet, a pivotal river-lake confluence located at 415 km downstream from the Three Gorges Dam (TGD) [45]. The unique hydrology, climate, and environmental conditions make Dongting Lake an important international wetland under the Ramsar Convention, providing various habitats for flora and fauna [43,46,47]. Typical submerged plants with the widest distribution are *Potamogeton malaianus*, *Hydrilla verticillata*, *Vallisneria natans*, and *Ceratophyllum demersum*, which are important tools for lake ecosystem restoration [48].

With the dual effects of the four tributaries in the Dongting Lake basin and the Yangtze River, a complicated river-lake regime has formed near Chenglingji [49]. Since the mass exchange of pollution loads at the Chenglingji Station was impacted by both the lake and the river, we used the hydrological and water quality data from four stations (Xiangtan, Taojiang, Taoyuan, and Shimen) at the

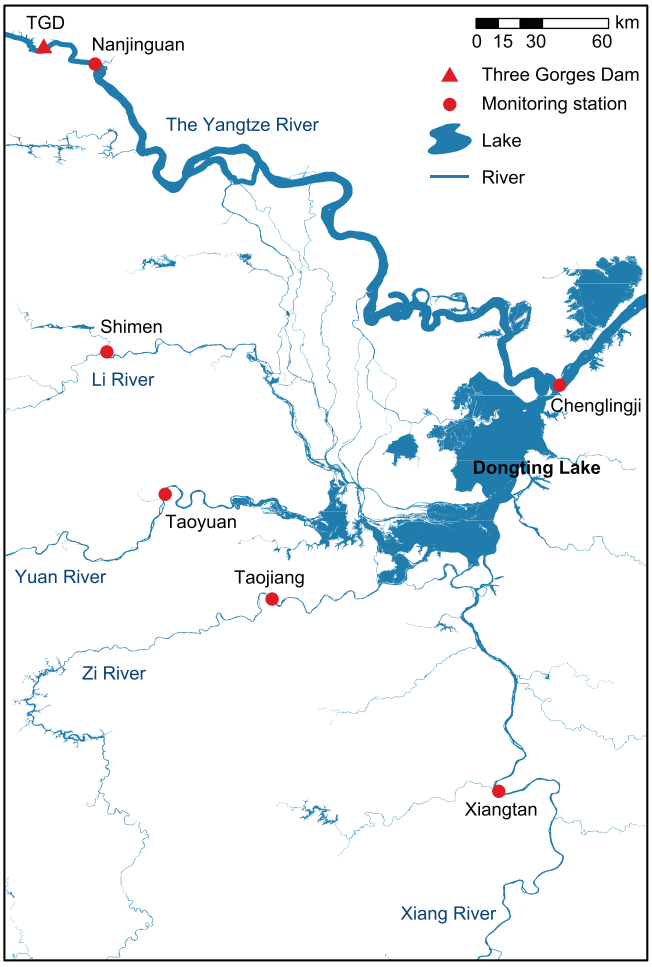

Fig. 1. The distribution of the monitoring stations in Dongting Lake and the Yangtze River. Daily streamflow and weekly water quality were measured at each monitoring station, and the river-lake exchange loads were measured at the Chenglingji Station.

tributaries of Dongting Lake and one (Nanjinguan) on the Yangtze River as model inputs to predict the pollutant output at the river-lake confluence (Fig. 1). The model input-output structure and the upstream-downstream relationship are illustrated in Fig. S1. As permanganate index (COD_{Mn}) and ammonia nitrogen (NH_3N) were key water parameters that could quickly reflect the pollution load condition for government administration to issue water quality early warning [26], we took these two indicators as the main research objects. The weekly COD_{Mn} and NH_3N concentrations of the six sites in Xiangtan, Taojiang, Taoyuan, Shimen, Nanjinguan, and Chenglingji from June 2012 to December 2018 were obtained from the China National Environmental Monitoring Centre (http://www.cnemc.cn/). There were only 293 weeks of data out of 344 weeks, and the missing data were concentrated in 2017 and 2018 (Fig. S2). The daily streamflow data of these six sites during the same period were available from the official website of the Hubei Water Resources Department (https://slt.hubei.gov.cn/sjfb/).

2.2. The load estimator

Developed by the United States Geological Survey, the Load Estimator (LOADEST) is a statistical method to estimate site-specific constituent loads [50–52]. It uses continuous streamflow data and discrete constituent concentrations to establish a regression model to interpolate and supplement the entire constituent load time series [53]. The LOADEST program includes several predefined models with different forms of the regression equation. Also, it provides users with an automated model selection option based on Akaike information criteria (AIC) to pick the best form [51,54,55]. Taking the most generalized form as an example to explain the working principle, the model can be expressed as

$$\ln(L) = a_0 + \sum_{i=1}^{N} a_i X_i \tag{1}$$

where L is an estimate of instantaneous load; a_0 and a_i are model parameters that can be calibrated for different sites; X_i is an explanatory variable, and N is the total number of explanatory variables [51,52,56].

In this study, the daily COD_{Mn} and NH_3N loads were computed from the weekly water quality data using the LOADEST program; that is, the water quality data were expanded and downscaled temporarily. The adjusted maximum likelihood estimator (AMLE) was adopted to calibrate the LOADEST model, and AMLE's coefficient of determination was used to test the model performance [57,58]. Among the six sites, the determination coefficients ranged from 0.73 to 0.88 for COD_{Mn}, and from 0.55 to 0.79 for NH_3N (Table S1). These results indicate that applying the LOADEST program was feasible and reasonable.

2.3. Machine learning methods

Machine learning is a burgeoning data-driven approach to artificial intelligence that includes models of various structures. Among many machine learning models, RNN is commonly used for hydrological and water quality sequence problems because of its memory and parameter-sharing characteristics [37,59]. Traditional RNN consists of an input layer, an output layer, and one or more hidden layers. RNN was used as a benchmark model in this study.

2.3.1. Long short-term memory

LSTM is a variant that solves the problem of vanishing or explosive gradients to learn long-term dependencies between model inputs and outputs for time-series tasks [60,61]. Unlike simple RNNs with one state variable, an LSTM layer comprises two states (i.e., cell and hidden states) and a series of sequentially connected memory cells [62]. Each memory cell mainly contains three gates (i.e., input, forget, and output gates) and other small units [63]. Such characteristics of LSTM are generally considered suitable for exploring the implicit internal relationships of nonlinear systems with hysteretic behavior in nature [64,65]. The structure of data transfer and the steps of model running are illustrated in Fig. 2a. The detailed algorithm for LSTM described above can be expressed as follows [66,67]:

$$\text{Input gate}: i_t = \sigma(W_{ix}x_t + W_{ih}h_{t-1} + b_i) \tag{2}$$

$$\text{Forget gate}: f_t = \sigma\left(W_{fx}x_t + W_{fh}h_{t-1} + b_f\right) \tag{3}$$

$$\text{Output gate}: o_t = \sigma(W_{ox}x_t + W_{oh}h_{t-1} + b_o) \tag{4}$$

$$\text{Cell state}: c_t = f_t \odot c_{t-1} + i_t \odot \tanh(W_{cx}x_t + W_{ch}h_{t-1} + b_c) \tag{5}$$

$$\text{Hidden state}: h_t = o_t \odot \tanh(c_t) \tag{6}$$

$$\text{Output}: y_t = W_{hy}h_t + b_y \tag{7}$$

where the subscript t and $t-1$ denote the time step for time-dependent variables; W and b are the weight matrices and bias vectors for calibration; x_t is the input and y_t is the predicted output; $\odot$ is Hadamard product (i.e., element-wise multiplication); σ is the sigmoidal activation function.

2.3.2. Bidirectional long short-term memory

In time series tasks, sometimes the output at the current moment is jointly determined by the previous state and the subsequent state, which may make it more accurate, so bidirectional RNN is proposed [68]. Bidirectional long short-term memory (Bi-LSTM), based on the concept of bidirectional RNN [69], is developed to encode backward-to-forward information and better capture bidirectional dependencies. Because the diffusion and dispersion of pollutants take time, a latent relationship may exist between the pollution loads of consecutive days. Bi-LSTM could better capture this relationship to improve modeling performance [70]. A Bi-LSTM layer consists of a forward LSTM layer and a backward LSTM layer (Fig. 2b). At each time step, the outputs of the forward and backward layers are saved, and the final output of this Bi-LSTM layer is obtained by combining them [70,71]. As a result, each node in the output layer contains complete bidirectional contextual information [72]. The mathematical expression is as follows:

$$y_t = W_{\overrightarrow{h}y}\overrightarrow{h}_t + W_{\overleftarrow{h}y}\overleftarrow{h}_t + b_y \tag{8}$$

where $\overrightarrow{h}_t$ and $\overleftarrow{h}_t$ denote the outputs or hidden states of the forward LSTM layer and backward LSTM layer, respectively; $W_{\overrightarrow{h}y}$ and $W_{\overleftarrow{h}y}$ are the corresponding weight matrices, and b_y is the bias parameter of the Bi-LSTM layer.

2.3.3. Self-Attentive long short-term memory (SA-LSTM)

Although LSTM can partly alleviate the long-term dependence problem in RNN, the attention mechanism is introduced in various machine learning tasks to further improve the information extraction ability [73,74]. The self-attention mechanism is a variant of the attention mechanism, and it reduces the dependence on external information and is better at capturing the internal

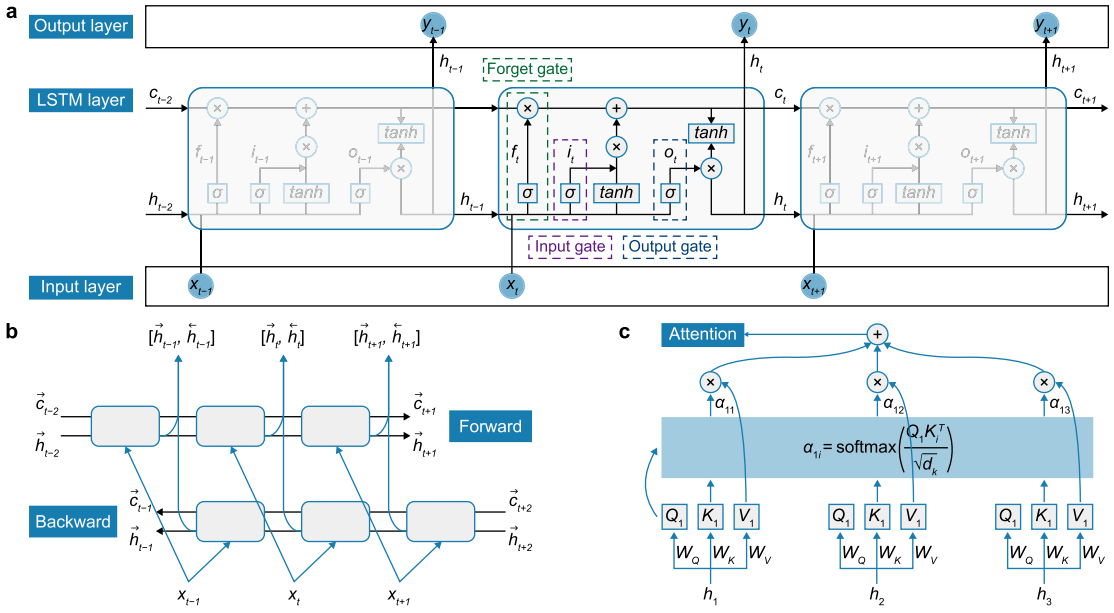

Fig. 2. The structure of three long short-term memory (LSTM) models: **a**, traditional LSTM; **b**, Bidirectional LSTM; **c**, Self-Attentive LSTM. The subscript t denotes the time step for time-dependent variables; x_t is the input and y_t is the predicted output; i_t, f_t, o_t are the input, forget, and output gates; c_t and h_t are the cell and hidden states; σ is the sigmoid activation function; $\overrightarrow{h}_t$ and $\overleftarrow{h}_t$ denote the hidden states of the forward LSTM layer and backward LSTM layers, which is similar for cell states $\overrightarrow{c}_t$ and $\overleftarrow{c}_t$; Q, K, and V are three feature vectors called query, key, and value; d_k is the dimension of the feature vector K; W_Q, W_K, and W_V are the projection matrices; α is the self-attention weight.

correlation of data or features. It allows inputs to interact and determine what they should be paying more attention to. When using LSTM for pollution load prediction, its hidden nodes extract various features of upstream pollution sources. However, LSTM cannot recognize which node is significantly different from the others (e.g., the one that contains the most information of the peak), whereas the self-attention mechanism may assist in focusing more on these critical nodes. Suppose $H = \{h_1, h_2, \cdots, h_t\}$ is the hidden state vector of the last LSTM layer. The self-attention layer converts the hidden state vector to three feature vectors: query Q, key K, and value V, respectively [75].

$$Q = H \times W_Q \tag{9}$$

$$K = H \times W_K \tag{10}$$

$$V = H \times W_V \tag{11}$$

where W_Q, W_K and W_V are the projection matrices. Note that Q in self-attention is a transformation of its own input vector, as shown in equation (9) and Fig. 2c, while it comes from the outside (e.g., target vector) in the traditional attention mechanism. The self-attention weight and the output vector can be expressed as

$$\alpha = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) \tag{12}$$

$$\mathrm{Attention}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{13}$$

where d_k is the dimension of the feature vector K; and softmax is a normalized exponential function. Here we got the value weight by the similarity of query and key and used a scale factor $\sqrt{d_k}$ to keep the gradient stable. And this method is called scaled dot-product attention [74].

2.3.4. Model settings and parameterization

The main objective of this study is to explore a feasible method for predicting pollution loads using sparse and partially missing water quality data rather than simply pursuing the highest accuracy. Therefore, we adopted a general method to select parameters or used default settings and tried to make the parameters have more interpretable physical meanings. All machine learning models in this study were run on the publicly available and standardized library PyTorch in the Spyder environment of Anaconda. All inputs and outputs were normalized using min-max normalization to avoid dimension interference on experimental results. Considering the limited data, the models were conducted using the five-fold cross-validation method to avoid the problem of unbalanced data partitioning [23]. This approach involves randomly partitioning the data into five subsets over five cycles, selecting one-fifth of the data as the testing set and the remaining four-fifths as the training set. The hyperparameters were determined before performing the cross-validation operation, which divided the sequential data into training, validation, and testing sets in a ratio of 6:2:2, aligning with the cross-validation ratio of 8:2. The number of hidden layers was set to 2 through trial and error [76,77]. The learning rate was optimized by adaptive moment estimation (Adam) algorithm [78]. The epoch was controlled by an early stopping method to avoid overfitting [79]. Then, three feature-related parameters, i.e., observation time step, batch size, and hidden-state size, were fine-tuned using the grid search method [80]. Since it takes no more than a week for the water from the TGD and the four stations in the tributaries of Dongting Lake to reach the Chenglingji outlet [81,82], the observation time step (also referred to as memory length) was searched from 3 to 7 days. The search range of batch size and hidden-state size was $\{2^3, 2^4, 2^5, 2^6, 2^7, 2^8\}$. The optimal observation time step, batch size, and hidden-state size obtained by the grid search method are shown in Table S2.

2.4. Framework overview and evaluation

Machine learning models (e.g., LSTM) can take upstream boundary conditions as input to predict water quality at downstream sites for better watershed pollution control [79], unlike site-specific autoregressive prediction (e.g., LOADEST) that cannot forecast in advance based on pollution source variation. Therefore, our first step was to combine the two methods, run machine learning models with LOADEST-expanded daily data, and pick out the best-combined model (Fig. 3). We would compare it with the single machine learning without LOADEST to verify the availability of the combined model. Secondly, to explore the potential of the best combination in sparse water quality modeling, we continuously selected odd-numbered weeks of water quality measured data at equal time intervals to obtain a bi-weekly water quality series. We repeated the operation to obtain a four-weekly (about monthly) water quality series. Then we used LOADEST to expand the bi- and four-weekly water quality data to daily data for LSTM modeling. Finally, we also tested the lead-time forecast ability of the best-combined model. The detailed flow chart of this study is shown in Fig. 3.

Three statistical metrics, namely Nash-Sutcliffe efficiency coefficient (*NSE*), the square of Pearson's correlation coefficient (R^2), and root mean squared error (*RMSE*), were used to evaluate the model performance [35,110]:

$$NSE = 1 - \frac{\sum_{i=1}^{n}\left(y_{o,i} - y_{s,i}\right)^2}{\sum_{i=1}^{n}\left(y_{o,i} - \overline{y_o}\right)^2} \tag{14}$$

$$R^2 = \frac{\left[\sum_{i=1}^{n}\left(y_{s,i} - \overline{y_s}\right)\left(y_{o,i} - \overline{y_o}\right)\right]^2}{\sum_{i=1}^{n}\left(y_{s,i} - \overline{y_s}\right)^2 \sum_{i=1}^{n}\left(y_{o,i} - \overline{y_o}\right)^2}, \tag{15}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}\left(y_{s,i} - y_{o,i}\right)^2}{n}}, \tag{16}$$

where $y_{o,i}$ is the measured pollution load and $y_{s,i}$ is the predicted pollution load for time step i; $\overline{y_o}$ and $\overline{y_s}$ are the measured mean and the predicted mean, respectively; n is the total data length within the analysis period. Although the LOADEST-LSTM model can produce a pollution load value daily, the days without measured data must be excluded from the evaluation. Hence, only the days with available sparse water quality data were extracted for accuracy calculation.

3. Results

3.1. Model performance for COD_{Mn} and NH_3N loads

The performance of RNN and LSTMs combined with LOADEST was significantly better than those without LOADEST as shown in Fig. 4 and Table S3 (i.e., larger *NSE*, larger R^2, and smaller *RMSE*), especially for NH_3N with a more obvious gap. The COD_{Mn} was better predicted than the NH_3N, both for the single or LOADEST-combined machine learning models, which was consistent with the pollution load interpolation performance of LOADEST (Table S1). The accuracy of RNN was slightly inferior to that of LSTMs based on the LOADEST-expanded daily water quality data. As for the three LSTM-LOADEST models for the COD_{Mn} and NH_3N loads, SA-LSTM-LOADEST outperformed the other models, and Bi-LSTM-LOADEST performed similarly as, if not better than, the traditional LSTM-LOADEST model. Furthermore, the error bars show that the SA-LSTM-LOADEST model also had relatively less uncertainty on the data across different cross-validation epochs. Overall, the best-performing SA-LSTM-LOADEST successfully captured the variation trend of the COD_{Mn} and NH_3N pollution loads for the whole study period, although there were a few mismatches at extreme loads (e.g., June 22, 2015 for COD_{Mn} and July 3, 2017 for NH_3N; see Fig. 5). A detailed look indicated that SA-LSTM-LOADEST was superior to the other LSTM-LOADEST models because of its more accurate simulation of COD_{Mn} load peaks (Fig. S3). The pollution load peak of COD_{Mn} and NH_3N in July 2017 was mainly caused by the double peak of pollutant concentration and flooding (Fig. S2). In this case, the SA-LSTM-LOADEST model successfully captured the COD_{Mn} peak but failed to capture the NH_3N peak. In addition, compared with the single SA-LSTM model without LOADEST, the *RMSE* value of SA-LSTM-LOADEST model was 24.6% lower for COD_{Mn} (21.3% lower for NH_3N).

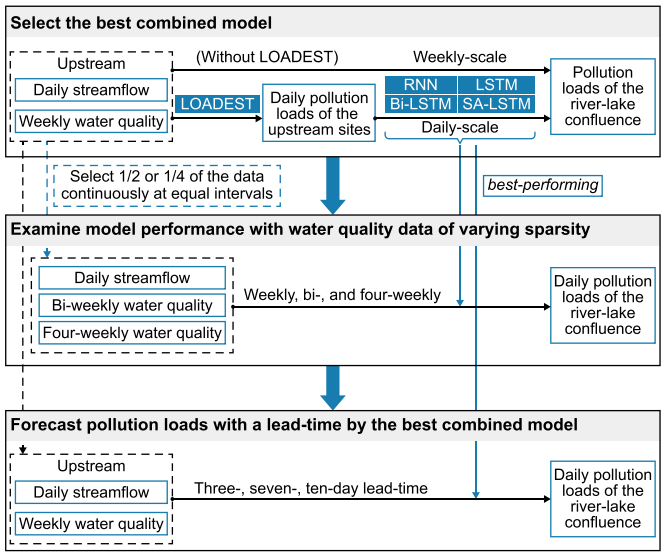

Fig. 3. The flow chart and study framework. The black arrows with start and end points indicate the input and output of a model. The blue arrows indicate model application, and the dashed arrows indicate data transfer.

To better probe into the performance of SA-LSTM-LOADEST in various pollution load ranges, we divided the predicted values into high, intermediate, and low load intervals at a ratio of 1:1:1. As illustrated in Fig. 6, the COD_{Mn} and NH_3N predictions mostly overestimated in the low-load interval and underestimated in the high-load intervals. However, the correlation between predictions and observations was higher in the high-load and low-load intervals than in the intermediate-load interval.

3.2. Pollution load modeling with varying sparsity (from weekly to monthly)

Machine learning usually requires mass data to learn the implicit relationship between variables. LOADEST was used to expand the sparse water quality data into the dense to serve the LSTM models in our research. To investigate the different impacts of sparse and dense water quality data on the prediction accuracy, we tested the SA-LSTM-LOADEST performance with varying sparsity (i.e., weekly, bi-weekly, and four-weekly) of the raw water quality (Fig. 7). When the frequency of water quality monitoring changed from weekly to bi-weekly, or even four-weekly (about monthly), the model still performed well without much loss of accuracy. The

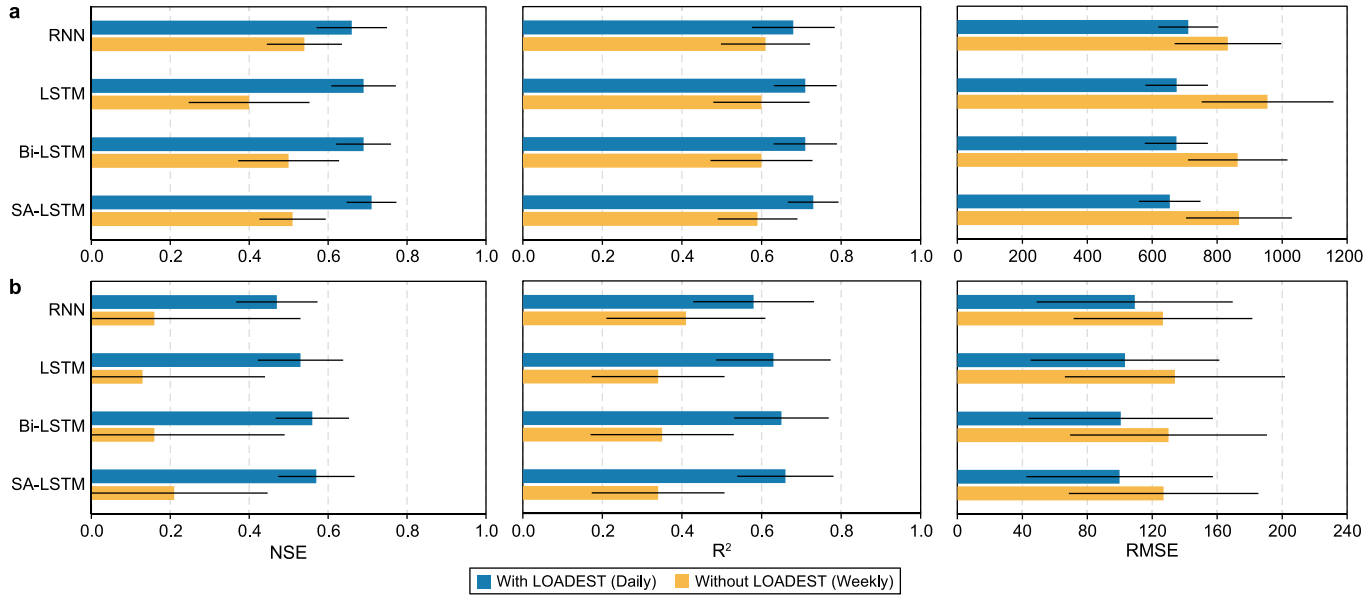

Fig. 4. Accuracy of the COD_{Mn} (**a**) and NH_3N (**b**) pollution loads at the river-lake confluence using various machine learning models (i.e., RNN, LSTM, Bi-LSTM, and SA-LSTM) with and without LOADEST. The error bars are the standard deviation of the five-fold cross-validation accuracy.

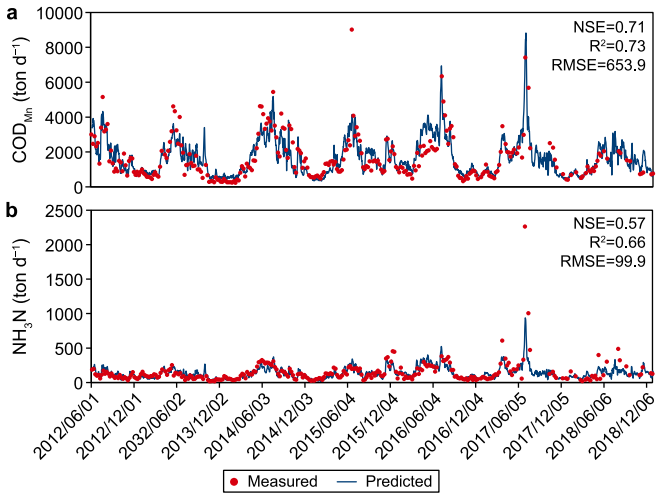

Fig. 5. Predicted and measured COD_{Mn} (**a**) and NH_3N (**b**) pollution load values at the river-lake confluence using the best-performing SA-LSTM-LOADEST model.

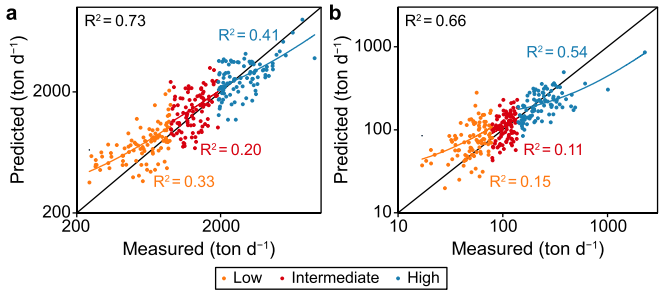

Fig. 6. Predicted pollution loads at the high, intermediate, and low load intervals using the SA-LSTM-LOADEST model: **a**, COD_{Mn}; **b**, NH_3N. The results are plotted with a logarithmic scale to show the model performance for high and low loads. The dotted lines are the corresponding trend lines.

RMSE value of COD_{Mn} load increased by 7.9% when water quality sparsity from weekly to monthly, while it surprisingly decreased by 1.6% for NH_3N load, perhaps because some data with high error were not selected when re-produce the LOADEST-expanded data. Moreover, the uncertainty of three evaluation metrics (i.e., *NSE*, R^2, and *RMSE*) to the data reduced slightly with the data becoming more sparse. In any case, the SA-LSTM-LOADEST model effectively modeled the monthly-scale COD_{Mn} and NH_3N loads.

3.3. Pollution load forecasting with a lead-time by the SA-LSTM-LOADEST model

Accurate and reliable pollution load forecasts with specific lead times are helpful for pollution control and water quality management. We tested the forecast ability of the SA-LSTM-LOADEST model with a three-, seven-, or ten-day lead time (Fig. 8). With an increased lead time, the model's performance became slightly worse. Specifically, the *RMSE* values increased by 1.2%, 13.5%, and 13.8% for COD_{Mn}, and 5.2%, 3.0%, and 14.6% for NH_3N with a three-, seven-, or ten-day lead-time, respectively. Because of the limited loss of short-term forecast accuracy, which might be less than the model error itself, the *RMSE* value of NH_3N with a seven-day lead-time was smaller than that with a three-day lead time. Besides, the stages when the model cumulative error increased rapidly differed for COD_{Mn} and NH_3N. The forecast accuracy of COD_{Mn} decreased more when the lead time went from three to seven days, while it was seven to ten days for NH_3N. Overall, the SA-LSTM-LOADEST model could forecast pollution loads up to ten days in advance with the *RMSE* value increasing no more than 15% than that of the 0-day.

4. Discussion

4.1. Uncertainty analysis of data and models

Due to different monitoring equipment and methods, water quality sequence data are interfered with by some complex noise that can cause a large penalty in the accuracy of water quality modeling [115]. Research from Zhang et al. (2023) [116] and Song

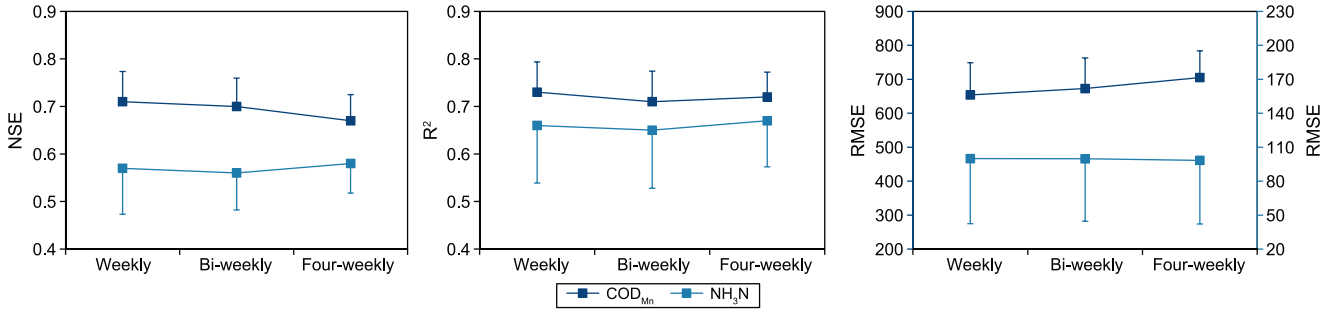

Fig. 7. The prediction results of the SA-LSTM-LOADEST model with varying water quality sparsity. The daily data for SA-LSTM are obtained using LOADEST to re-expand the bi- and four-weekly data continuously selected from raw weekly water quality data. Error bars are only shown half for the clearer figure.

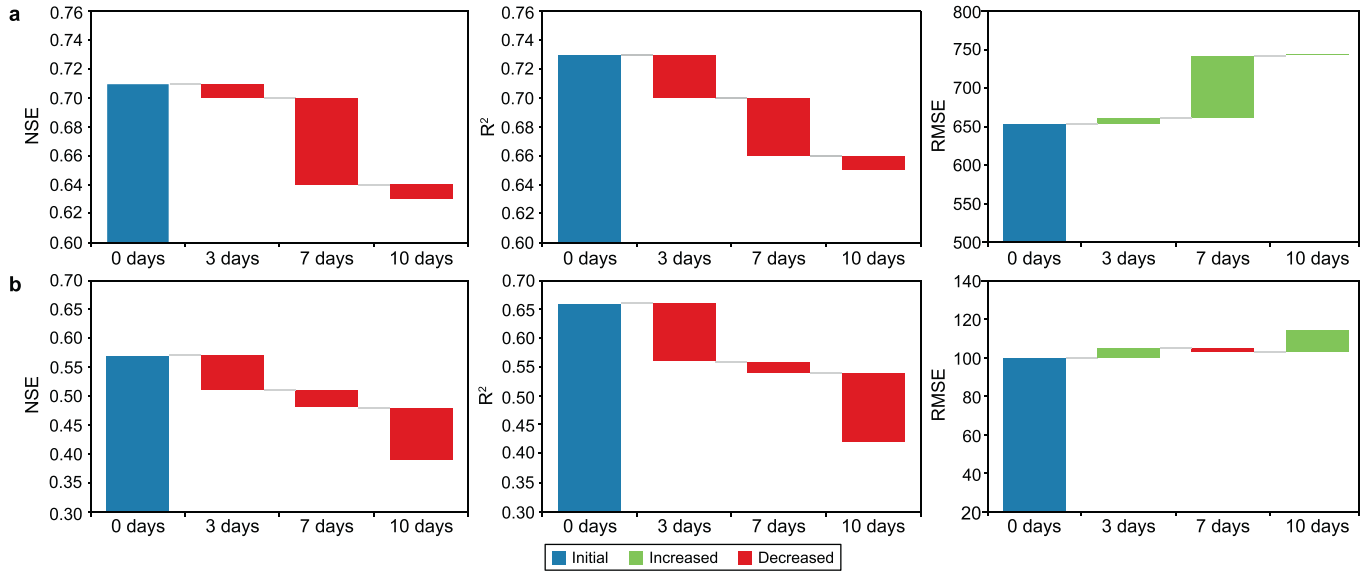

Fig. 8. The waterfall plot of COD_{Mn} (**a**) and NH_3N (**b**) pollution load forecasts with three, seven, and ten-day lead-time steps using the SA-LSTM-LOADEST model.

et al. (2021) [117] showed that wavelet transform could denoise water quality data to facilitate the recent prediction of LSTM models. Therefore, we used wavelet transform to denoise the raw data (Fig. S4) and compared the impact of noise reduction on the SA-LSTM-LOADEST model results (Fig. 9). The denoising algorithm slightly improved the combined model performance in this study, but not much. We subsequently applied the same wavelet transform technology to the LOADEST-expanded data. We found that the denoising effect was very weak (Fig. S5), which showed that LOADEST played a certain role in noise reduction when expanding the water quality data. Additionally, although denoising is expected to enhance model prediction, it is necessary to avoid mistakenly deleting some abnormal peak data caused by severe pollution events, especially in water quality modeling with sparse and limited data.

The model uncertainty mainly comes from the LOADEST program and LSTM. Since LSTMs are driven by LOADEST-generated synthetic data, it is logical that their simulation capability is limited by how accurately the LOADEST program expands water quality data at a specific site, such as the Chenglingji station [53,58]. The errors generated when LOADEST expands the data will propagate to the final results through LSTMs. In this study, the accuracy of the COD_{Mn} predictions was better than that of NH_3N. This discrepancy can be attributed to the measurement data and, more likely, to the weaker correlation between NH_3N and streamflow

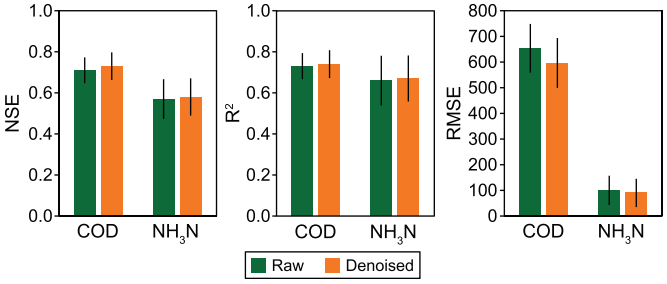

Fig. 9. Model performance based on raw and denoised data. The denoising algorithm is a discrete wavelet transform from the PyWavelets library.

compared to that between COD_{Mn} and streamflow. When some data with larger errors had been removed, the four-weekly NH_3N load simulation was better than the bi-weekly. Therefore, the accuracy of the LOADEST-expanded data explained part of the uncertainty in the LSTM-LOADEST model [51].

Machine learning models like LSTM also have some common errors in fitting. The rivers' peak flow or peak water quality is not always that much, causing the LSTM insufficient peak training and making it difficult to predict the loads that significantly deviate from the mean loads range [35,79]. Such a drawback of machine

learning in water quality modeling has also been confirmed by other studies [32,35]. In this study, the LSTM models could perceive the temporal variation of the pollution loads, but it is difficult to match the magnitude of actual pollution loads. But satisfactorily, self-attention helped LSTM improve the generalization ability, and SA-LSTM achieved better COD_{Mn} performance than the traditional LSTM in the peak prediction. This generalization ability facilitated the short-term lead-time forecasting of SA-SLTM. In fact, when the lead time increased, the temporal gap between input and output would also be increased, thereby generating more noise from the input data and less correlation between the input and output of the model [32,83,84]. However, the cumulative error of SA-LSTM in multi-day lead-time prediction appeared to be less than that of the traditional LSTM used in other studies [83,84]. In general, more complex structures and more parameters also bring greater model uncertainty. Therefore, to reduce the uncertainty of machine learning like SA-LSTM, the model characteristic parameters should be more physically meaningful and explicable [79,85].

4.2. External environmental factors affecting pollution load modeling

External environmental factors affect the accuracy of pollution load modeling. Due to the limited data, in this study, streamflow was taken as the only explanatory variable of the LOADEST program to compute pollution loads, but in fact, the pollution loads in rivers may also be affected by other factors such as water temperature and climate phenomena [86–88]. For instance, El Niño influences the variability of water quality through biogeochemistry, which goes beyond simple flow-load relationships [89]. Changes in water temperature can alter the rate of nutrient release from sediments in rivers and lakes, consequently affecting the flux exchange of pollutant loads at the sediment-water interface [90,91]. Rossi et al. (2021) [112] found that ignoring pH when using LOADEST could cause larger errors for constituents highly controlled by pH-dependent reactions. As a result, the model's accuracy is promising to be improved with more environmental explanatory variables introduced into the LOADEST program. It is convenient to include user-defined variables (e.g., conductance, pH, water temperature, and turbidity) in addition to streamflow in LOADEST, which is already integrated into the application. Furthermore, the LOADEST program provides an example of alkalinity load calibration using streamflow and specific conductance [51].

Water quality monitoring needs to be improved, and the lack of boundary conditions increases the difficulty of pollution load modeling. More automatic water quality monitoring stations are being built [92,93]. However, due to the technical problems of water quality sensors, the water quality data are not so reliable and representative [94,95]. The pollutants in Dongting Lake are monitored in the tributaries considering China's environmental management policies, but it cannot rule out that unmonitored agricultural water or domestic sewage directly entered the lake and interfered with the model results [96,97]. Advances in water quality management will facilitate more accurate simulation and prediction of pollution loads [98,99].

4.3. Application of machine learning in modeling water quality with sparse data

The performance of a machine learning model improves with an increase in training data size [100]. Due to the complexity and high cost of measurements, water quality series are often at weekly or monthly frequencies [38,40,101]. However, some meteorological data and streamflow monitoring have been done on minute to daily frequency [102,103]. Therefore, in our study, the LOADEST program was used to increase the density of the water quality series by utilizing the high-frequency streamflow data, creating more opportunities for machine learning models to train and learn the implicit relationship between input and output while also refining the temporal scale of pollution load modeling. Besides, the self-attention mechanism strengthened the learning ability of LSTM, making it more efficient to perceive important information, especially peaks. The SA-LSTM-LOADEST method provided a new idea for machine learning modeling in sparse water quality measurement. Since the model demonstrated good accuracy at a river-lake confluence with a complex flow regime, it could also be implemented at basin outlets and key cross-sections with sufficient streamflow monitoring. More importantly, this upstream-downstream modeling method can potentially promote flow-controlled water quality management (e.g., dam regulation, sudden water pollution accidents, and pollution reduction scenarios in watersheds) because it can reflect the relationship between downstream pollution loads and variations in upstream pollution sources.

There are also other techniques and methods for machine learning to deal with a small amount of water quality data. Transfer learning techniques can help machine learning apply knowledge learned in domains with sufficient data to related domains with insufficient data [104]. Combining the advantages of LSTM in capturing long-term dependencies and the model transfer ability of transfer learning, the features of a series of small temporal scales can be applied to increase the density of large temporal scale data or impute missing data [105–107]. It is also possible to expand the series of data-deficient sites by transferring the characteristics of long-term high-frequency monitoring data of adjacent sites [70]. However, the limitation of this method is that there must be a suitable water quality series as the source domain; that is, the data volume of the series is large enough and has a certain similarity with the target series [105]. In addition, innovative technologies with specialized functions show potential for integration with machine learning models (e.g., LSTM) to enhance the accuracy of source-sink water quality modeling. For example, few-shot learning is developed for small-sample tasks [108,109], and the seasonal-trend decomposition procedure based on loess (STL) can capture seasonal features more effectively [110]. Whether it is the LOADEST program used in this study, transfer learning, or other specific methods for sparse data [39], errors are inevitably introduced during the processing of water quality data with machine learning. Hence, developing strategies to control data preprocessing errors and improve the water quality prediction accuracy remains a crucial area for further exploration.

5. Conclusion

To conclude, this study explored the traditional RNN and three LSTMs combined with LOADEST to improve the modeling accuracy of pollution loads with sparse water quality data, especially at the complicated river-lake confluence. The best-performing SA-LSTM-LOADEST model was established. We further analyzed its performance in varying sparsity (i.e., weekly, bi-weekly, and four-weekly) and its forecasting ability. The main findings of this study are summarized as follows.

- The SA-LSTM performed best among the RNN and three LSTMs ($NSE_{COD_{Mn}} = 0.71$, $NSE_{NH_3N} = 0.57$) combined with LOADEST program. Compared with the single SA-LSTM without LOADEST,

the SA-LSTM-LOADEST lowered the *RMSE* by 24.6% for COD_{Mn} and 21.3% for NH_3N. Additionally, the prediction accuracy of the SA-LSTM-LOADEST model could be further improved by using wavelet transform to denoise the raw data.

- The SA-LSTM-LOADEST model still performed well when the data sparsity changed from weekly to four-weekly ($NSE_{COD_{Mn}} = 0.67$, $NSE_{NH_3N} = 0.58$), indicating the SA-LSTM-LOADEST could be potentially used for pollution loads modeling with observations at monthly or even longer temporal scales.
- The SA-LSTM-LOADEST model could forecast with a lead time of ten days. The ten-day lead-time *RMSE* value increased no more than 15% than that of the 0-day.

The SA-LSTM-LOADEST method effectively modeled and forecasted pollution loads with sparse water quality data at the river-lake confluence. This work can be further improved by adding explanatory variables, such as water temperature, and upgrading water quality monitoring technology. Moreover, advanced machine learning techniques should be explored to tackle the challenges in water quality modeling effectively with sparse or missing water quality series.

CRediT authorship contribution statement

Sheng Huang: Conceptualization, Methodology, Investigation, Formal Analysis, Visualization, Writing - Original Draft, Writing - Review & Editing. **Jun Xia:** Conceptualization, Resources, Project Administration, Supervision, Formal Analysis, Visualization, Writing - Review & Editing. **Yueling Wang:** Methodology, Investigation, Resources, Formal Analysis, Writing - Review & Editing. **Jiarui Lei:** Methodology, Supervision, Formal Analysis, Visualization, Writing - Review & Editing. **Gangsheng Wang:** Conceptualization, Methodology, Formal Analysis, Visualization, Writing - Review & Editing.

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.

Acknowledgments

This work was supported by the Strategic Priority Research Program of the Chinese Academy of Sciences (XDA23040502), National Natural Science Foundation of China (41890823), and Key Laboratory of Water Cycle and Related Land Surface Processes, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences (No. WL2019003).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ese.2024.100402.

References

[1] A.K. Aufdenkampe, E. Mayorga, P.A. Raymond, J.M. Melack, S.C. Doney, S.R. Alin, R.E. Aalto, K. Yoo, Riverine coupling of biogeochemical cycles between land , oceans , and atmosphere, Front. Ecol. Environ. 9 (2011) 53–60, https://doi.org/10.1890/100014.

[2] G. Grill, B. Lehner, M. Thieme, B. Geenen, D. Tickner, F. Antonelli, S. Babu, L. Cheng, H. Crochetiere, H.E. Macedo, R. Filgueiras, M. Goichot, J. Higgins, Z. Hogan, B. Lip, M.E. Mcclain, J. Meng, M. Mulligan, C.R. Liermann, C. Nilsson, J.D. Olden, J.J. Opperman, A. Van-Soesbergen, C. Zarfl, J. Snider, F. Tan, K. Tockner, Mapping the world's free-flowing rivers, Nature 569 (2019) 215–221, https://doi.org/10.1038/s41586-019-1111-9.

[3] J.A. Downing, S. Polasky, S.M. Olmstead, S.C. Newbold, Protecting local water quality has global benefits, Nat. Commun. (2015) 8–13, https://doi.org/10.1038/s41467-021-22836-3.

[4] F.E. Rowland, C.A. Stow, T.H. Johengen, A.M. Burtner, D. Palladino, D.C. Gossiaux, T.W. Davis, L.T. Johnson, S. Ruberg, Recent patterns in lake erie phosphorus and chlorophyll a concentrations in response to changing loads, Environ. Sci. Technol. 54 (2020) 835–841, https://doi.org/10.1021/acs.est.9b05326.

[5] V. Sagan, K.T. Peterson, M. Maimaitijiang, P. Sidike, J. Sloan, B.A. Greeling, S. Maalouf, C. Adams, Monitoring inland water quality using remote sensing : potential and limitations of spectral indices, bio-optical simulations, machine learning, and cloud computing, Earth Sci. Rev. 205 (2020), https://doi.org/10.1016/j.earscirev.2020.103187.

[6] J.R. Gardner, T.M. Pavelsky, M. Doyle, The abundance, size, and spacing of lakes and reservoirs connected to river networks, Geophys. Res. Lett. 46 (2019) 2592–2601, https://doi.org/10.1029/2018GL080841.

[7] N.M. Schmadel, J.W. Harvey, R.B. Alexander, G.E. Schwarz, R.B. Moore, K. Eng, J.D. Gomez-velez, E.W. Boyer, D. Scott, Thresholds of lake and reservoir connectivity in river networks control nitrogen removal, Nat. Commun. 9 (2018), https://doi.org/10.1038/s41467-018-05156-x.

[8] A. Hillbricht-ilkowska, The dynamics and retention of phosphorus in lentic and lotic patches of two river-lake systems, Hydrobiologia 251 (1993) 257–268, https://doi.org/10.1007/bf00007185.

[9] A.E. Jones, B.R. Hodges, J.W. McClelland, A.K. Hardison, K.B. Moffett, Residence-time-based classification of surface water systems Allan, Water Resour. Res. 53 (2017) 5567–5584, https://doi.org/10.1002/2016WR019928.

[10] X. Lai, J. Jiang, Q. Liang, Q. Huang, Large-scale hydrodynamic modeling of the middle Yangtze River Basin with complex river – lake interactions, J. Hydrol. 492 (2013) 228–243, https://doi.org/10.1016/j.jhydrol.2013.03.049.

[11] G. Yang, Q. Zhang, R. Wan, X. Lai, X. Jiang, L. Li, H. Dai, G. Lei, J. Chen, Y. Lu, Lake hydrology , water quality and ecology impacts of altered river – lake interactions : advances in research on the middle Yangtze river, Nord. Hydrol 47 (2016) 1–7, https://doi.org/10.2166/nh.2016.003.

[12] J. Harvey, M. Gooseff, River corridor science: hydrologic exchange and ecological consequences from bedforms to basins, Water Resour. Res. 51 (2015) 6893–6922, https://doi.org/10.1002/2015WR017617.

[13] A. Kuriata-potasznik, S. Szymczyk, A. Skwierawski, Influence of cascading river – lake Systems on the dynamics of nutrient circulation in catchment areas, Water 12 (2020) 1144, https://doi.org/10.3390/w12041144.

[14] J. Gao, J. Jia, A.J. Kettner, F. Xing, Y. Ping, X. Nan, Y. Yang, Changes in water and sediment exchange between the Changjiang River and Poyang Lake under natural and anthropogenic conditions, China. Sci. Total Environ. 481 (2014) 542–553, https://doi.org/10.1016/j.scitotenv.2014.02.087.

[15] D. Sharma, A. Kansal, W.Á. Wqrrs, Á.B.Á. Epdriv, Assessment of river quality models : a review, Rev. Environ. Sci. Bio-Technology 12 (2013) 285–311, https://doi.org/10.1007/s11157-012-9285-8.

[16] A.N. Ahmed, F.B. Othman, H.A. Afan, A. Elsha, Machine learning methods for better water quality prediction, J. Hydrol. 578 (2019), https://doi.org/10.1016/j.jhydrol.2019.124084.

[17] S. Huang, J. Xia, S. Zeng, Y. Wang, D. She, Effect of Three Gorges Dam on Poyang Lake water level at daily scale based on machine learning, J. Geogr. Sci. 31 (2021) 1598–1614, https://doi.org/10.1007/s11442-021-1913-1.

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

[19] S. Baek, J. Pyo, J.A. Chun, Prediction of water level and water quality using a CNN-LSTM combined deep learning approach, Water 12 (2020), https://doi.org/10.3390/w12123399.

[20] R. Huang, C. Ma, J. Ma, X. Huangfu, Q. He, Machine learning in natural and engineered water systems, Water Res. 205 (2021), https://doi.org/10.1016/j.watres.2021.117666.

[21] K.P. Singh, A. Basant, A. Malik, G. Jain, Artificial neural network modeling of the river water quality — a case study, Ecol. Model. 220 (2009) 888–895, https://doi.org/10.1016/j.ecolmodel.2009.01.004.

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

[23] R. Xia, G. Wang, Y. Zhang, P. Yang, Z. Yang, S. Ding, X. Hou, K. Zhang, X. Gao, P. Duan, C. Qian, River algal blooms are well predicted by antecedent environmental conditions, Water Res. 185 (2020) 116221, https://doi.org/10.1016/j.watres.2020.116221.

[24] J. Pyo, K. Hwa, K. Kim, S. Baek, G. Nam, S. Park, Cyanobacteria cell prediction using interpretable deep learning model with observed , numerical , and sensing data assemblage, Water Res. 203 (2021), https://doi.org/10.1016/j.watres.2021.117483.

[25] L.H. Silva, J. Alexandre, K. Silva, Non - intrusive , real - time deep learning - based pollution analysis applied to open - channels, J. Brazilian Soc. Mech. Sci. Eng. 43 (2021), https://doi.org/10.1007/s40430-021-03096-0.

[26] K. Chen, H. Chen, C. Zhou, Y. Huang, X. Qi, R. Shen, F. Liu, M. Zuo, X. Zou, J. Wang, Y. Zhang, D. Chen, X. Chen, Y. Deng, H. Ren, Comparative analysis of surface water quality prediction performance and identification of key water parameters using different machine learning models based on big data, Water Res. 171 (2020) 115454, https://doi.org/10.1016/j.watres.2019.115454.

[27] M. Liu, Y. Huang, J. Hu, J. He, X. Xiao, Algal community structure prediction by

machine learning, Environmental Science and Ecotechnology (2022) 100233, https://doi.org/10.1016/j.ese.2022.100233.
[28] Y. Lecun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (2015) 436–444, https://doi.org/10.1038/nature14539.
[29] Z. Liang, R. Zou, X. Chen, T. Ren, H. Su, Y. Liu, Simulate the forecast capacity of a complicated water quality model using the long short-term memory approach, J. Hydrol. 581 (2020), https://doi.org/10.1016/j.jhydrol.2019.124432.
[30] R. Markus, G. Camps-valls, B. Stevens, M. Jung, J. Denzler, N. Carvalhais, Deep learning and process understanding for data-driven Earth system science, Nature 566 (2019) 195–204, https://doi.org/10.1038/s41586-019-0912-1.
[31] Z. Yu, K. Yang, Y. Luo, C. Shang, Spatial-temporal process simulation and prediction of chlorophyll-a concentration in Dianchi Lake based on wavelet analysis and long-short term memory network, J. Hydrol. 582 (2020) 124488, https://doi.org/10.1016/j.jhydrol.2019.124488.
[32] L. Zheng, H. Wang, C. Liu, S. Zhang, A. Ding, E. Xie, J. Li, S. Wang, Prediction of harmful algal blooms in large water bodies using the combined EFDC and LSTM models, J. Environ. Manag. 295 (2021) 113060, https://doi.org/10.1016/j.jenvman.2021.113060.
[33] N.H. Than, D.L. Che, P.V. Tat, The performance of classification and forecasting dong nai river water quality for sustainable water resources management using neural network techniques, J. Hydrol. 596 (2021) 126099, https://doi.org/10.1016/j.jhydrol.2021.126099.
[34] P. Liu, J. Wang, A.K. Sangaiah, Y. Xie, X. Yin, Analysis and prediction of water quality using LSTM deep neural networks in IoT environment, Sustainability 11 (2019) 1–14, https://doi.org/10.3390/su11072058.
[35] W. Zhi, D. Feng, W. Tsai, G. Sterle, A. Harpold, C. Shen, L. Li, From hydrometeorology to River water quality: can a deep learning model predict dissolved oxygen at the continental scale? Environ. Sci. Technol. 55 (2021) https://doi.org/10.1021/acs.est.0c06783.
[36] B. Yang, Z. Xiao, Q. Meng, Y. Yuan, W. Wang, H. Wang, Y. Wang, X. Feng, Deep learning-based prediction of effluent quality of a constructed wetland, Environmental Science and Ecotechnology 13 (2023) 100207, https://doi.org/10.1016/j.ese.2022.100207.
[37] I. Goodfellow, A. Courville, Y. Bengio, Deep Learning, MIT Press, 2016.
[38] J.W. Kirchner, C. Neal, Universal fractal scaling in stream chemistry and its implications for solute transport and water quality trend detection, Proc. Natl. Acad. Sci. U.S.A. 110 (2013), https://doi.org/10.1073/pnas.1304328110.
[39] J. Ma, Y. Ding, J.C.P. Cheng, F. Jiang, Z. Xu, Soft detection of 5-day BOD with sparse matrix in city harbor water using deep learning techniques, Water Res. 170 (2020) 115350, https://doi.org/10.1016/j.watres.2019.115350.
[40] C. Minaudo, R. Dupas, C. Gascuel-Odoux, O. Fovet, P.-E. Mellander, P. Jordan, M. Shore, F. Moatar, Nonlinear empirical modeling to estimate phosphorus exports using continuous records of turbidity and discharge, Water Resour. Res. 53 (2017) 7590–7606, https://doi.org/10.1002/2017WR020590.
[41] R. Barzegar, M.T. Aalami, J. Adamowski, Short-term water quality variable prediction using a hybrid CNN–LSTM deep learning model, Stoch. Environ. Res. Risk Assess. 34 (2020) 415–433, https://doi.org/10.1007/s00477-020-01776-2.
[42] A.M. Michalak, J. Xia, D. Brdjanovic, A.-N. Mbiyozo, D. Sedlak, T. Pradeep, U. Lall, N. Rao, J. Gupta, The frontiers of water and sanitation, Nature Water 1 (2023) 10–18, https://doi.org/10.1038/s44221-022-00020-1.
[43] M. Geng, K. Wang, N. Yang, F. Li, Y. Zou, Evaluation and variation trends analysis of water quality in response to water regime changes in a typical river-connected lake (Dongting Lake), China, Environ. Pollut. 268 (2021) 115761, https://doi.org/10.1016/j.envpol.2020.115761.
[44] X. Long, F. Liu, X. Zhou, J. Pi, W. Yin, F. Li, S. Huang, Estimation of spatial distribution and health risk by arsenic and heavy metals in shallow groundwater around Dongting Lake plain using GIS mapping, Chemosphere 269 (2021) 128698, https://doi.org/10.1016/j.chemosphere.2020.128698.
[45] Y. Yu, X. Mei, Z. Dai, J. Gao, J. Li, J. Wang, Y. Lou, Hydromorphological processes of Dongting Lake in China between 1951 and 2014, J. Hydrol. 562 (2018) 254–266, https://doi.org/10.1016/j.jhydrol.2018.05.015.
[46] H. Ru, X. Liu, X. Huang, Y. Ning, H. Wang, Diversity of fish species and its spatio-temporal variations in Lake Dongting, a large Yangtze-connected lake, J. Lake Sci. 20 (2008) 93–99, https://doi.org/10.18307/2008.0114.
[47] Y. Zou, P. Zhang, S. Zhang, X. Chen, F. Li, Z. Deng, S. Yang, H. Zhang, F. Li, Y. Xie, Crucial sites and environmental variables for wintering migratory waterbird population distributions in the natural wetlands in East, Sci. Total Environ. 655 (2019) 147–157, https://doi.org/10.1016/j.scitotenv.2018.11.185.
[48] X. Liu, Z. Hou, Y. Xie, X. Yu, X. Li, Influence of water level on four typical submerged plants in wetlands of Lake Dongting, J. Lake Sci. 33 (2021) 181–191, https://doi.org/10.18307/2021.0113.
[49] X. Dai, G. Yang, R. Wan, Y. Li, The effect of the changjiang river on water regimes of its tributary lake east dongting, J. Geogr. Sci. 28 (2018) 1072–1084, https://doi.org/10.1016/j.jglr.2015.06.008.
[50] T.G. Huntington, W.M. Balch, G.R. Aiken, J. Sheffield, L. Luo, C.S. Roesler, P. Camill, Climate change and dissolved organic carbon export to the Gulf of Maine, J. Geophys. Res. Biogeosciences (2016), https://doi.org/10.1002/2015JG003314.
[51] B.R.L. Runkel, C.G. Crawford, T.A. Cohn, Load Estimator (LOADEST): A FORTRAN Program for Estimating Constituent Loads in Streams and Rivers, Tech. Methods B. 4-A5 U.S. Geological Survey, Reston, VA, 2004, https://doi.org/10.3133/tm4A5.
[52] J.R. Stewart, B. Livneh, J.R. Kasprzyk, B. Rajagopalan, J.T. Minear, W.J. Rasemen, A multialgorithm approach to land surface modeling of suspended sediment in the Colorado front range, J. Adv. Model. Earth Syst. (2017) 2526–2544, https://doi.org/10.1002/2017MS001120.
[53] L. Chen, C. Sun, G. Wang, H. Xie, Z. Shen, Event-based nonpoint source pollution prediction in a scarce data catchment, J. Hydrol. 552 (2017) 13–27, https://doi.org/10.1016/j.jhydrol.2017.06.034.
[54] H. Akaike, A new look at the statistical model identification, IEEE Trans. Automat. Control 19 (1974) 716–723, https://doi.org/10.1109/TAC.1974.1100705.
[55] X. Gao, N. Chen, D. Yu, Y. Wu, B. Huang, Hydrological controls on nitrogen (ammonium versus nitrate) fluxes from river to coast in a subtropical region : observation and modeling, J. Environ. Manag. 213 (2018) 382–391, https://doi.org/10.1016/j.jenvman.2018.02.051.
[56] D. Chen, M. Hu, Y. Guo, R.A. Dahlgren, Reconstructing historical changes in phosphorus inputs to rivers from point and nonpoint sources in a rapidly developing watershed in eastern, Sci. Total Environ. 533 (2015) 196–204, https://doi.org/10.1016/j.scitotenv.2015.06.079.
[57] B.A. Pellerin, B.A. Bergamaschi, R.J. Gilliom, C.G. Crawford, J. Saraceno, C.P. Frederick, B.D. Downing, J.C. Murphy, Mississippi river nitrate loads from high frequency sensor measurements and regression-based load estimation, Environ. Sci. Technol. 48 (2014) 12612–12619. https://doi-org.libproxy1.nus.edu.sg/10.1021/es504029c.
[58] Y. Zhu, L. Chen, G. Wei, S. Li, Z. Shen, Uncertainty assessment in base flow nonpoint source pollution prediction : the impacts of hydrographic separation methods , data sources and base flow period assumptions, J. Hydrol. 574 (2019) 915–925, https://doi.org/10.1016/j.jhydrol.2019.05.010.
[59] S. Yang, D. Yang, J. Chen, B. Zhao, Real-time reservoir operation using recurrent neural networks and inflow forecast from a distributed hydrological model, J. Hydrol. 579 (2019) 124229, https://doi.org/10.1016/j.jhydrol.2019.124229.
[60] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (1997) 1735–1780, https://doi.org/10.1162/neco.1997.9.8.1735.
[61] F. Kratzert, D. Klotz, C. Brenner, K. Schulz, M. Herrnegger, Rainfall – runoff modelling using long short-term memory (LSTM) networks, hydrol, Earth Syst. Sci. 22 (2018) 6005–6022, https://doi.org/10.5194/hess-22-6005-2018.
[62] D. Feng, K. Fang, C. Shen, Enhancing stream flow forecast and extracting insights using long - short term memory networks with data integration at continental scales, Water Resour. Res. 56 (2020) 1–24, https://doi.org/10.1029/2019WR026793.
[63] F.A. Gers, J. Schmidhuber, F. Cummins, Learning to forget : continual prediction with LSTM, Neural Comput. 12 (2000) 2451–2471, https://doi.org/10.1049/cp:19991218.
[64] Z. Xiang, J. Yan, I. Demir, A rainfall-runoff model with LSTM-based sequence-to-sequence learning, Water Resour. Res. 56 (2019), https://doi.org/10.1029/2019WR025326.
[65] G. Young, J. Ngarambe, P. Nzivugira, G. Ulpiani, R. Paolini, S. Haddad, K. Vasilakopoulou, M. Santamouris, Predicting the magnitude and the characteristics of the urban heat island in coastal cities in the proximity of desert landforms, The case of Sydney. Sci. Total Environ. 709 (2020) 136068, https://doi.org/10.1016/j.scitotenv.2019.136068.
[66] K. Greff, R.K. Srivastava, J. Koutník, B.R. Steunebrink, J. Schmidhuber, Lstm : a search space odyssey, IEEE Transact. Neural Networks Learn. Syst. 28 (2017) 2222–2232. http://arxiv.org/abs/1503.04069.
[67] F. Kratzert, D. Klotz, M. Herrnegger, A.K. Sampson, Toward improved predictions in ungauged basins : exploiting the power of machine learning, Water Resour. Res. 55 (2019) 11344–11354, https://doi.org/10.1029/2019WR026065.
[68] M. Schuster, K.K. Paliwal, Bidirectional recurrent neural network, IEEE Trans. Signal Process. 45 (1997) 2673–2681, https://doi.org/10.1109/78.650093.
[69] A. Graves, Framewise phoneme classification with bidirectional LSTM and other neural network architectures, Neural Network. 18 (2005) 602–610, https://doi.org/10.1016/j.neunet.2005.06.042.
[70] J. Ma, Z. Li, J.C.P. Cheng, Y. Ding, C. Lin, Z. Xu, Air quality prediction at new stations using spatially transferred bi- directional long short-term memory network, Sci. Total Environ. 705 (2020) 135771, https://doi.org/10.1016/j.scitotenv.2019.135771.
[71] A. Ullah, S. Member, J. Ahmad, S. Member, Action recognition in video sequences using deep Bi-directional LSTM with CNN features, IEEE Access 6 (2018) 1155–1166, https://doi.org/10.1109/ACCESS.2017.2778011.
[72] J. Yin, Z. Deng, A.V.M. Ines, J. Wu, E. Rasu, Forecast of short-term daily reference evapotranspiration under limited meteorological variables using a hybrid bi-directional long short-term memory model (Bi-LSTM), Agric. Water Manag. 242 (2020) 106386, https://doi.org/10.1016/j.agwat.2020.106386.
[73] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, Proc. Int. Conf. Learn. Represent (2015) 1–15. http://arxiv.org/abs/1409.0473.
[74] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, Adv. Neural Inf. Process. Syst. 30 (2017) 5998–6008. http://arxiv.org/abs/1706.03762.
[75] G. Shi, Y. Leung, J. She, T. Fung, F. Du, Y. Zhou, A novel method for identifying hotspots and forecasting air quality through an adaptive utilization of spatio-temporal information of multiple factors, Sci. Total Environ. 759 (2021), https://doi.org/10.1016/j.scitotenv.2020.143513.
[76] L. Wen, X. Li, L. Gao, S. Member, A new reinforcement learning based

learning rate scheduler for convolutional neural network in fault classification, IEEE Trans. Ind. Electron. (2020), https://doi.org/10.1109/TIE.2020.3044808.
[77] X. Ye, C.Y. Xu, Q. Zhang, J. Yao, X. Li, Quantifying the human induced water level decline of China's largest freshwater lake from the changing underlying surface in the lake region, Water Resour. Manag. 32 (2018) 1467–1482, https://doi.org/10.1007/s11269-017-1881-5.
[78] D.P. Kingma, J.L. Ba, Adam: a method for stochastic optimization, Comput. Sci. 1–15 (2015). http://arxiv.org/abs/1412.6980.
[79] S. Huang, J. Xia, Y. Wang, W. Wang, S. Zeng, D. She, G. Wang, Coupling machine learning into hydrodynamic models to improve river modeling with complex boundary conditions, Water Resour. Res. 58 (2022), https://doi.org/10.1029/2022WR032183.
[80] H.A. Fayed, A.F. Atiya, Speed up grid-search for parameter selection of support vector machines, Appl. Soft Comput. J. 80 (2019) 202–210, https://doi.org/10.1016/j.asoc.2019.03.037.
[81] Y. Guo, X. Lai, Water level prediction of Lake Poyang based on long short-term memory neural network, J. Lake Sci. 32 (2020) 865–876, https://doi.org/10.18307/2020.0325.
[82] X. Lai, J. Jiang, Q. Huang, Pattern of impoundment effects and influencing mechanism of Three Gorges Project on water regime of Lake Dongting, J. Lake Sci. 24 (2012) 178–184, https://doi.org/10.18307/2012.0202.
[83] M. Cheng, F. Fang, T. Kinouchi, I.M. Navon, C.C. Pain, Long lead-time daily and monthly streamflow forecasting using machine learning methods, J. Hydrol. 590 (2020), https://doi.org/10.1016/j.jhydrol.2020.125376.
[84] Y. Lian, J. Luo, J. Wang, G. Zuo, N. Wei, Climate - driven model based on long short - term memory and bayesian optimization for multi - day - ahead daily streamflow forecasting, Water Resour. Manag. (2021), https://doi.org/10.1007/s11269-021-03002-2.
[85] A. Mcgovern, R. Lagerquist, D.J. Gagne, G.E. Jergensen, K.L. Elmore, C.R. Homeyer, T. Smith, Making the black box more transparent: understanding the physical implications of machine learning, Bull. Am. Meteorol. Soc. 100 (2019) 2175–2200, https://doi.org/10.1175/BAMS-D-18-0195.1.
[86] Z. Li, X. Li, Y. Li, Study of the COD release in the sediment of tang He reservoir, Adv. Mater. Res. 613 (2013) 1134–1138. https://doi.org/10.4028/www.scientific.net/AMR.610-613.1134.
[87] H.E.M. Meier, R. Hordoir, H.C. Andersson, C. Dieterich, K. Eilola, B.G. Gustafsson, S. Schimanke, A. Ho, Modeling the combined impact of changing climate and changing nutrient loads on the Baltic Sea environment in an ensemble of transient simulations for 1961 – 2099, Clim. Dynam. (2012) 2421–2441, https://doi.org/10.1007/s00382-012-1339-7.
[88] J. Torrecilla, J.P. Galve, L.G. Zaera, J.F. Retamar, N.A.A. Alejandro, Nutrient sources and dynamics in a mediterranean fluvial regime (Ebro river , NE Spain) and their implications for water management, J. Hydrol. 304 (2005) 166–182, https://doi.org/10.1016/j.jhydrol.2004.07.029.
[89] A.P. Smits, C.M. Ruf, T.V. Royer, A.P. Appling, N.A. Grif, Detecting signals of large - scale climate phenomena in discharge and nutrient loads in the Mississippi - atchafalaya river basin, Geophys. Res. Lett. (2019) 3791–3801, https://doi.org/10.1029/2018GL081166.
[90] K.J. Gibbons, T.B. Bridgeman, Effect of temperature on phosphorus flux from anoxic western Lake Erie sediments, Water Res. 182 (2020) 116022, https://doi.org/10.1016/j.watres.2020.116022.
[91] J. Zhong, J. Yu, J. Wang, D. Liu, C. Chen, C. Fan, The co-regulation of nitrate and temperature on denitrification at the sediment-water interface in the algae-dominated ecosystem of Lake Taihu , China, J. Soils Sediments 20 (2020) 2277–2288, https://doi.org/10.1007/s11368-019-02558-2.
[92] J. Kahiluoto, J. Hirvonen, T. Näykki, Automatic real-time uncertainty estimation for online measurements : a case study on water turbidity, Environ. Monit. Assess. 191 (2019), https://doi.org/10.1007/s10661-019-7374-7.
[93] A. Mentzafou, Y. Panagopoulos, E. Dimitriou, Designing the national network for automatic monitoring of water quality parameters in Greece, Water 11 (2019) 1310, https://doi.org/10.3390/w11061310.
[94] S.A. Jaywant, K.M. Arif, A comprehensive review of microfluidic water quality monitoring sensors, Sensors 19 (2019), https://doi.org/10.3390/s19214781.
[95] P. Kruse, Review on water quality sensors, J. Phys. D Appl. Phys. 51 (2018), https://doi.org/10.1088/1361-6463/aabb93.
[96] Y. Hou, W. Chen, Y. Liao, Y. Luo, Scenario analysis of the impacts of socio-economic development on phosphorous export and loading from the Dongting Lake watershed , China, Environ. Sci. Pollut. Res. 25 (2017) 26706–26723, https://doi.org/10.1007/s11356-017-0138-4.
[97] X. Wang, F. Hao, H. Cheng, Estimating non-point source pollutant loads for the large-scale basin of the Yangtze River in China, Environ. Earth Sci. 63 (2011) 1079–1092, https://doi.org/10.1007/s12665-010-0783-0.
[98] Y. Cai, X. Fu, X. Gao, L. Li, Research progress of on-line automatic monitoring of chemical oxygen demand (COD) of water, IOP Conf. Ser. Earth Environ. Sci. 121 (2018) 022039, https://doi.org/10.1088/1755-1315/121/2/022039.
[99] Y. Zhuang, W. Wen, S. Ruan, F. Zhuang, B. Xia, S. Li, Real-time measurement of total nitrogen for agricultural runoff based on multiparameter sensors and intelligent algorithms, Water Res. 210 (2022), https://doi.org/10.1016/j.watres.2021.117992.
[100] J. Booz, W. Yu, G. Xu, D. Griffith, N. Golmie, A deep learning-based weather forecast system for data volume and recency analysis, in: International Conference on Computing, Networking and Communications (ICNC), IEEE, 2019, pp. 697–701, https://doi.org/10.1109/ICCNC.2019.8685584.
[101] V. Vandenberghe, P.L.M. Goethals, A.V.A.N. Griensven, J. Meirlaen, N.D.E. Pauw, P. Vanrolleghem, W. Bauwens, Application of automated measurement stations for continuous water quality monitoring of the Dender River in Flanders, Belgium, Environ. Monit. Assess. 108 (2005) 85–98, https://doi.org/10.1007/s10661-005-3964-7.
[102] K. Kawanisi, M. Bahrainimotlagh, M. Basel, A. Sawaf, M. Razaz, High-frequency stream flow acquisition and bed level/flow angle estimates in a mountainous river using shallow-water acoustic tomography, Hydrol. Process. 2254 (2016) 2247–2254, https://doi.org/10.1002/hyp.10796.
[103] L.G. Lanza, E. Vuerich, The WMO field intercomparison of rain intensity gauges, Atmos. Res. 94 (2009) 534–543, https://doi.org/10.1016/j.atmosres.2009.06.012.
[104] S.J. Pan, Q. Yang, A survey on transfer learning, IEEE Trans. Knowl. Data Eng. 22 (2010) 1345–1359, https://doi.org/10.1109/TKDE.2009.191.
[105] Z. Chen, H. Xu, P. Jiang, S. Yu, G. Lin, I. Bychkov, A. Hmelnov, G. Ruzhnikov, N. Zhu, Z. Liu, A transfer Learning-Based LSTM strategy for imputing Large-Scale consecutive missing data and its application in a water quality prediction system, J. Hydrol. 602 (2021), https://doi.org/10.1016/j.jhydrol.2021.126573.
[106] J. Ma, J.C.P. Cheng, C. Lin, Y. Tan, J. Zhang, Improving air quality prediction accuracy at larger temporal resolutions using deep learning and transfer learning techniques, Atmos. Environ. 214 (2019), https://doi.org/10.1016/j.atmosenv.2019.116885.
[107] Y. Zhou, Real-time probabilistic forecasting of river water quality under data missing situation : deep learning plus post-processing techniques, J. Hydrol. 589 (2020), https://doi.org/10.1016/j.jhydrol.2020.125164.
[108] A. Santoro, S. Bartunov, M. Botvinick, D. Wierstra, T. Lillicrap, One-shot learning with memory-augmented neural networks, in: International Conference on Machine Learning, 2016. http://arxiv.org/abs/1605.06065.
[109] O. Vinyals, C. Blundell, T. Lillicrap, K. Kavukcuoglu, Matching networks for one shot learning, in: 30th Conference on Neural Information Processing Systems, NIPS), 2016. http://arxiv.org/abs/1606.04080.
[110] W. Dong, Y. Zhang, L. Zhang, W. Ma, L. Luo, What will the water quality of the Yangtze River be in the future? Sci. Total Environ. 857 (2023) https://doi.org/10.1016/j.scitotenv.2022.159714.
[111] J. Xia, C. Zhan, S. Zeng, L. Zou, D. She, Q. Zuo, Theoretical method and practical exploration of Yangtze River Simulator construction, J. Hydraul. Eng. 53 (2022), https://doi.org/10.13243/j.cnki.slxb.20220077.
[112] C. Rossi, J. Oyarzún, P. Pastén, R.L. Runkel, J. Núñez, D. Duhalde, H. Maturana, E. Rojas, J.L. Arumí, D. Castillo, R. Oyarzún, Assessment of a conservative mixing model for the evaluation of constituent behavior below river confluences, Elqui River Basin, Chile, River Res. Appl. 37 (2021), https://doi.org/10.1002/rra.3823.
[113] Z. Chang, W. Lu, Z. Wang, Study on source identification and source-sink relationship of LNAPLs pollution in groundwater by the adaptive cyclic improved iterative process and Monte Carlo stochastic simulation, J. Hydrol. 612 (2022), https://doi.org/10.1016/j.jhydrol.2022.128109.
[114] G. Mao, X. Duan, Z. Niu, J. Xu, X. X, X. Huang, H. Chen, F. Mehr, R. Moti, Z. Qiao, Application of source-sink theory and MCR model to assess hydrochemical change risk in Lhasa River basin, Tibet, China, Environ. Impact Assess. Rev. 101 (2023), https://doi.org/10.1016/j.eiar.2023.107124.
[115] S. Kang, H. Lin, Wavelet analysis of hydrological and water quality signals in an agricultural watershed, J. Hydrol 338 (2007), https://doi.org/10.1016/j.jhydrol.2007.01.047.
[116] W. Zhang, J. Zhao, P. Quan, J. Wang, X. Meng, Q. Li, Prediction of influent wastewater quality based on wavelet transform and residual LSTM, Appl. Soft Comput. 148 (2023), https://doi.org/10.1016/j.asoc.2023.110858.
[117] C. Song, L. Yao, C. Hua, Q. Ni, A novel hybrid model for water quality prediction based on synchrosqueezed wavelet transform technique and improved long short-term memory, J. Hydrol 603 (2021), https://doi.org/10.1016/j.jhydrol.2021.126879.