Environmental Science and Pollution Research
https://doi.org/10.1007/s11356-022-22758-7

RESEARCH ARTICLE

Comparative analysis of water quality prediction performance based on LSTM in the Haihe River Basin, China

Qiang Li[1] · Yinqun Yang[2] · Ling Yang[1] · Yonggui Wang[1]

Received: 16 March 2022 / Accepted: 24 August 2022

Abstract

As the most water shortage and water polluted area in China, the water quality prediction is of utmost needed and important in Haihe River Basin for its water resource management. The long short-term memory (LSTM) has been a widely used tool for water quality forecast in recent years. The performance and adaptability of LSTM for water quality prediction of different indicators needs to be discussed before it adopted in a specific basin. However, literature contains very few studies on the comparative analysis of the various prediction accuracy of different water quality indicators and the causes, especially in Haihe River Basin. In this study, LSTM was employed to predict biochemical oxygen demand (BOD), permanganate index (COD_{Mn}), dissolved oxygen (DO), ammonia nitrogen (NH_3–N), total phosphorus (TP), hydrogen ion concentration (pH), and chemical oxygen demand digested by potassium dichromate (COD_{Cr}). According to results under 24 different input conditions, it is demonstrated that LSTMs present better predicting on BOD, COD_{Mn}, COD_{Cr}, and TP (median Nash–Sutcliffe efficiency reaching 0.766, 0.835, 0.837, and 0.711, respectively) than NH_3–N, DO, and pH (median Nash–Sutcliffe efficiency of 0.638, 0.625, and 0.229, respectively). Besides, the performance of LSTM to predict water quality is linearly related to the maximum value of temporal autocorrelation and cross-correlation coefficients of water quality indicators calculated by maximal information coefficient with the coefficients of determination of 0.79 to approximately 0.80. This study would provide new knowledge and support for the practical application and improvement of the LSTM in water quality prediction.

Keywords Water quality prediction · Long short-term memory (LSTM) · Haihe River Basin · Performance comparative analysis

Introduction

The Haihe River Basin (HRB) occupies an extremely important strategic position in the political, economic, and cultural development of China (Song et al. 2021; Wang et al. 2014; Zhu et al. 2010). Under the background of rapid urbanization and economic development, the Haihe River Basin receives large capacities of sewage and waste discharged from diverse origins (Dang et al. 2017; Zheng et al. 2015), suffering the most water shortages and water pollution among all seven river basins in China (Cao et al. 2021). The ecological resources and regulatory functions have been destroyed on a large scale due to environmental pollution, which affects the healthy, sustainable, and coordinated development of the basin area (Song et al. 2021). Hence, it is urgent to carry out scientific and comprehensive water quality prediction in HRB and provide insights for its subsequent pollution control.

Responsible Editor: Xianliang Yi

Highlights
1. Differentwater quality indicators in the Haihe River Basinare predicted using LSTM.
2. LSTM is more accurate when increasing thetime steps of input variables properly.
3. LSTM performs better forBOD, COD_{Mn}, COD_{Cr}, and TPthan NH_3–N, DO, and pH.
4. LSTM performance is linearly related to autocorrelation/cross-correlation of inputs.

✉ Yonggui Wang
wangyg@cug.edu.cn

[1] Key Laboratory of Regional Ecology and Environmental Change, School of Geography and Information Engineering, China University of Geosciences, Wuhan, China

[2] Changjiang Water Resources Protection Institute, Wuhan 430051, China

Published online: 30 August 2022

Process-based models, such as Environmental Fluid Dynamics Code (EFDC) (Hamrick 1992) and Soil and Water Assessment Tool (SWAT) (Santhi et al. 2006), are primary tools for supporting water quality predictions and have been widely applied in environmental management. Nevertheless, a lot of input data and boundary conditions (such as topographic data, weather conditions, and load input) are required but usually unavailable, which makes it difficult and uncertain to establish process-based models (Jiang et al. 2021). However, data-driven techniques provide an effective alternative (Palani et al. 2008), such as ANN (Kim et al. 2010; Najah Ahmed et al. 2019), RNN (Antanasijević et al. 2013; Li et al. 2019). Data-driven models can efficiently establish the relationships among water quality variables, thus rendering the forecast of boundary conditions unnecessary (Liang et al. 2020; Maier and Dandy 1996). Among deep learning algorithms, the LSTM (Hochreiter and Schmidhuber 1997) has a unique gate structure, which enables it to capture long-term dependencies in time series and thus is particularly good at processing time series (Zhang et al. 2022), such as natural language processing (Mikolov et al. 2010), speech recognition (Graves and Jaitly 2014), and machine translation (Sutskever et al. 2014). Besides, LSTM has been introduced into the field of hydrology forecast, including predicting water table depth (Zhang et al. 2018) and streamflow(Hu et al. 2018; Kratzert et al. 2019; Le et al. 2019; Sudriani et al. 2019). Furthermore, some studies also reported the successfully application of LSTM in water quality prediction (Jiang et al. 2021; Song et al. 2021; Wang et al. 2017; Zhang et al. 2022). In particular, the ability of LSTM to simulate the forecast capacity of EFDC has been confirmed (Liang et al. 2020). Using LSTM neural network as the modeling algorithm is an effective way to improve the accuracy of water quality prediction (Zhang et al. 2022).

Recently, water quality in HRB has been predicted using different models, such as a comprehensive hydrodynamic and water quality model package (Liu et al. 2008), principal component regression (PCR), and artificial neural networks (ANN) (Zhang et al. 2012), chaotic prediction model based on wavelet transform (Zhang et al. 2016), improved fuzzy time series model (Li 2018), support vector regression (SVR) combined with empirical mode decomposition and fast independent component analysis noise reduction (Liang et al. 2019), and a hybrid LSTM model that recruits synchro-squeezed wavelet transform (SWT) (Song et al. 2021). It is obvious that the same model performed differently in predicting different water quality indicators. Although both the process-based model and data-driven model were applied to predict the water quality of HRB, these studies still have limitations. Firstly, dissolved oxygen (DO) and potassium permanganate index (COD_{Mn}) were the most popular water quality indicators to be predicted in HRB, which is a similar trend all over the world (Tiyasha et al. 2020). However, according to the *Bulletin on ecological and Environmental Conditions of China 2020* from the Ministry of Ecology and Environment of the People's Republic of China, chemical oxygen demand digested by potassium dichromate (COD_{Cr}), COD_{Mn}, biochemical oxygen demand (BOD) are the primary pollution indicators and total phosphorus (TP) for several tributaries of HRB. It is significant to predict and analyze these water quality indicators, whereas very few studies focus on predicting these water quality indicators, especially using data-driven models.

Secondly, the performance of LSTM varies greatly when predicting different indicators in different basins. For example, the RMSEs (root mean square error) of LSTM to predict DO and TP in Taihu Lake are respectively 0.046 and 0.041 (Wang et al. 2017), while those to predict DO in Yongding River and Gangnan gauging station of Haihe River Basin are 1.5588 and 0.9281 (Song et al. 2021), respectively. Accordingly, evaluation of feasibility and comparative analysis of LSTM is essential before it is applied to predict different water quality indicators in a specific basin.

Thirdly, few studies concentrate on the comparative analysis of the prediction accuracy of different water quality indicators, especially the performance of LSTM. Even though some studies found model performance would be various in predicting different water quality indicators (Najah Ahmed et al. 2019), little attention was paid to the causes. As the basic approach of water quality prediction models is to establish the sequential relationship between model input variables and output variables (Maier et al. 2010; Maier and Dandy 2000; Tiyasha et al. 2020), the time series features, which are mainly temporal autocorrelation/cross-correlation, of water quality indicators, affect model performance. Therefore, the relationship between temporal autocorrelation/cross-correlation of water quality indicators and model performance should be considered in practical application, which has not been analyzed yet.

The aims and significance of this paper are summarized as follows: Firstly, LSTMs were employed to predict 7 different water quality indicators of HRB. Water quality indicators considered are ammonia nitrogen (NH_3–N), hydrogen ion concentration (pH), BOD, COD_{Mn}, DO, TP, and COD_{Cr}. Secondly, model performances of different water quality indicators in HRB were compared and analyzed. Thirdly, the relationship between autocorrelation/cross-correlation of water quality indicators and LSTM performance was analyzed.

Materials and methods

Study area and data collection

The Haihe River Basin (HRB, shown in Fig. 1), is the largest river catchment in Northern China (Dang et al. 2017) with

a catchment area of about 2.6×10^5 km^2 and water resource of about 1.05×10^5 m^3/km^2. Located in the continental monsoon climate zone, the HRB (nearly 112–120°E, 35–43°N) belongs to the semi-humid and semi-arid regions. Approximately 75–85% of the annual precipitation is concentrated in the rainy season from June to September (Bao et al. 2012). Monthly water quality data of 76 monitoring stations in HRB from January 2010 to September 2014 were obtained from Hebei Provincial Academy of Ecological Environmental Science, China (http://www.hebhky.cn/).

In this study, water quality indicators considered are biochemical oxygen demand (BOD), permanganate index (COD_{Mn}), dissolved oxygen (DO), ammonia nitrogen (NH_3–N), total phosphorus (TP), hydrogen ion concentration (pH), and chemical oxygen demand digested by potassium dichromate (COD_{Cr}). The basic statistical parameters, i.e., mean, minimum, maximum, standard deviation (SD), and coefficient of variation (CV) of these water quality indicators are depicted in Table 1. Large changes can be seen in some water quality indicators with a high coefficient of variation (i.e., 1.796, 2.01, 1.4, 1.2, and 1.703 for NH_3–N, BOD, COD_{Mn}, COD_{Cr}, and TP, respectively). The existence of large disparity in the indicators' concentrations can be attributed to the types (non-point and point) and nature of sources that have been distributed in the river basin's wide geographical area, and large geographical variations in climate as well as seasonal effects pertaining to the study region (Najah Ahmed et al. 2019).

Methodology framework

The methodology framework is shown in Fig. 2. Firstly, the temporal autocorrelation coefficients (ACF) and cross-correlation coefficients (CCF) of water quality monitoring data were calculated with lag times from 1 to 12. In this study, the ACF and CCF examine the correlation between the current concentration and historical concentration of the same water quality indicator or different indicators, respectively. Secondly, a number of LSTM models and artificial neural network (ANN, as control experiment) models with different model structures were developed, whose performances were thereafter evaluated based on Nash–Sutcliffe

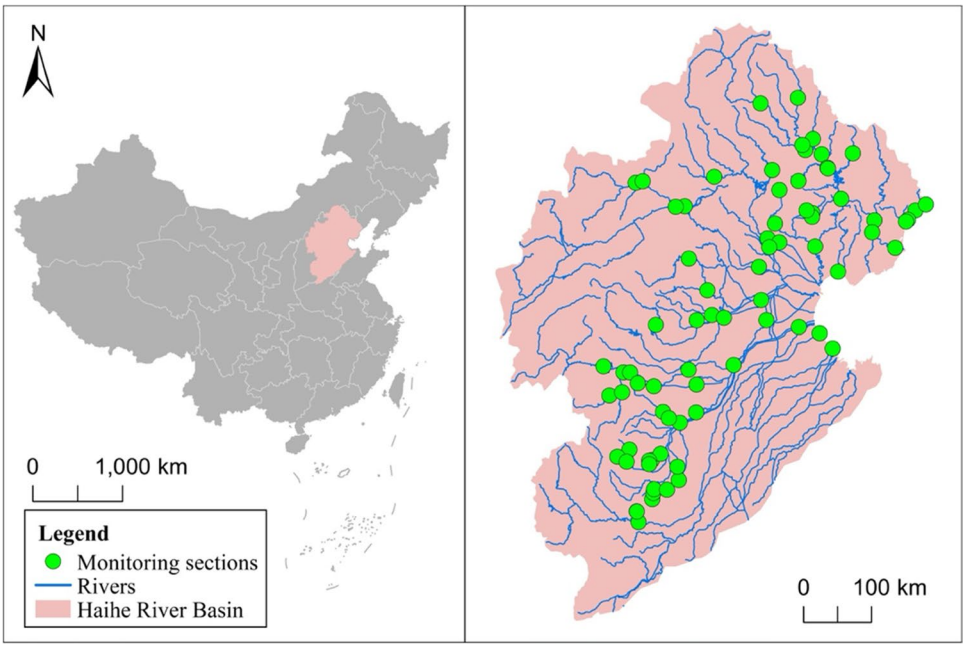

Fig. 1 Study areas and locations of water quality monitoring stations

Table 1 Basic statistical analysis for seven water quality indicators

Indicators	Unit	Mean	Minimum	Maximum	SD	CV
NH_3–N	mg/l	8.104	0.012	122.000	14.554	1.796
BOD	mg/l	12.3	0.2	220.0	24.7	2.01
COD_{Cr}	mg/l	45.2	2.0	381.0	56.3	1.2
COD_{Mn}	mg/l	11.0	0.6	127.0	16.0	1.4
DO	mg/l	6.75	0.02	18.80	3.50	0.52
pH	-	7.89	6.42	8.99	0.38	0.05
TP	mg/l	0.730	0.005	8.880	1.243	1.703

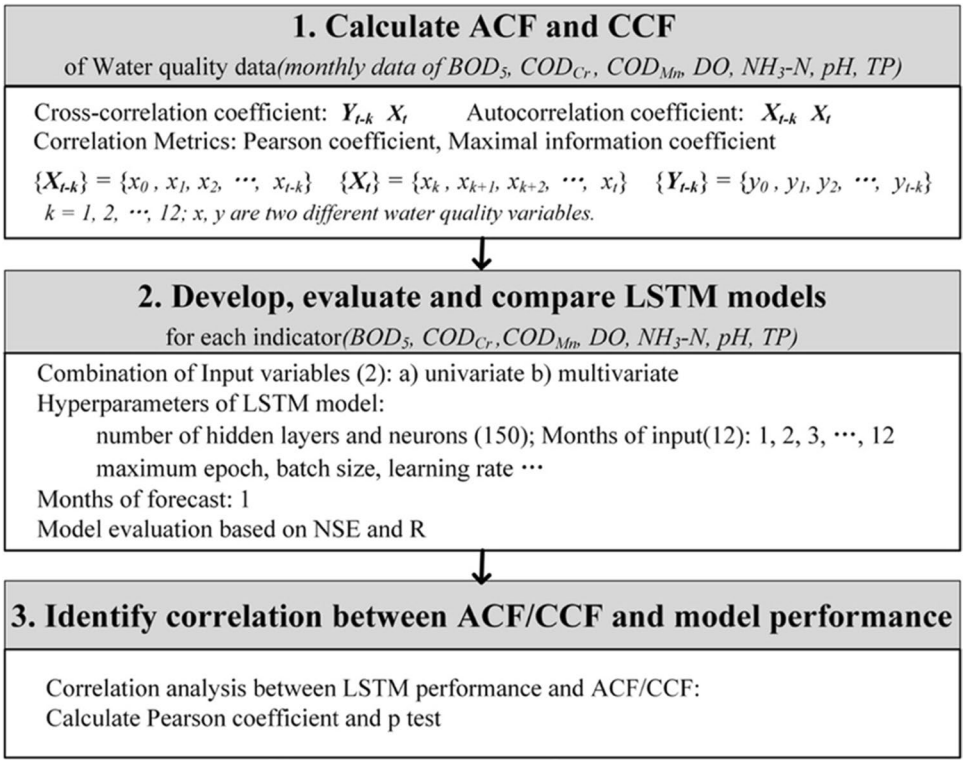

Fig. 2 Methodology framework of study. ACF and CCF represent temporal autocorrelation coefficients and cross-correlation coefficients, respectively

efficiency (NSE) and correlation of coefficient (R). LSTM performances of different water quality indicators were compared and analyzed. Finally, this study identified and analyzed the relationship between temporal autocorrelation/cross-correlation and LSTM performance.

Temporal autocorrelation/cross-correlation calculation

Temporal autocorrelation and cross-correlation coefficients are calculated respectively by Eq. (1) and (2):

$$\mathrm{AC}(x,k) = ACF\left(\boldsymbol{X_t}, \boldsymbol{X_{t-k}}\right) \tag{1}$$

$$\mathrm{CC}(x,y,k) = CCF\left(\boldsymbol{X_t}, \boldsymbol{Y_{t-k}}\right) \tag{2}$$

where AC and CC represent autocorrelation and cross-correlation respectively, and ACF and CCF denote correlation metric formulas (Eqs. (3)–(4)); k is the lag time ranging from 1 to 12; x and y are two different water quality indicators; $\boldsymbol{X_t} = \{x_k, x_{k+1}, x_{k+2}, \ldots, x_t\}$, $\boldsymbol{X_{t-k}} = \{x_0, x_1, x_2, \ldots, x_{t-k}\}$, and $\boldsymbol{Y_{t-k}} = \{y_0, y_1, y_2, \ldots, y_{t-k}\}$.

This study utilized two common correlation metrics, the Pearson coefficient and maximal information coefficient, to calculate temporal autocorrelation and cross-correlation coefficients of water quality indicators. Pearson coefficient (Pearson and Lee 1900), defined as Eq. (3), is one of the most famous relationship metrics with values varying between − 1 and 1. The closer its absolute value is to 1, the stronger the linear correlation is. Note that only the absolute values of the Pearson coefficient were considered. Furthermore, the maximal information coefficient (MIC, Eqs. (4) and (5)) can reflect the dependence between two variables no matter whether a linear or other functional relationship between them (Reshef et al. 2011). The measurement MIC is symmetric and normalized into a range [0, 1]. A high MIC value indicates a strong dependency between the investigated variables, whereas MIC = 0 describes the independent relationship between two variables.

$$r(\boldsymbol{x},\boldsymbol{y}) = \frac{\sum_{i=1}^{n}\left(x_i-\bar{\boldsymbol{x}}\right)\left(y_i-\bar{\boldsymbol{y}}\right)}{\sqrt{\sum_{i=1}^{n}\left(x_i-\bar{\boldsymbol{x}}\right)^2\sum_{i=1}^{n}\left(y_i-\bar{\boldsymbol{y}}\right)^2}} \tag{3}$$

$$\mathrm{MIC} = \max\left\{\frac{I(\boldsymbol{x},\boldsymbol{y})}{\log_2\min\left(n_x,n_y\right)}\right\} \tag{4}$$

$$\begin{aligned}(\boldsymbol{x},\boldsymbol{y}) = H(\boldsymbol{x}) + H(\boldsymbol{y}) - H(\boldsymbol{x},\boldsymbol{y}) = \sum_{i=1}^{n_x} p\left(x_i\right)\log_2\frac{1}{p(x_i)} \\ + \sum_{j=1}^{n_y} p\left(y_j\right)\log_2\frac{1}{p(y_j)} - \sum_{i=1}^{n_x}\sum_{j=1}^{n_y} p\left(x_i,y_j\right)\log_2\frac{1}{p(x_i,y_j)}\end{aligned} \tag{5}$$

where $x = \{x_1, x_2, \ldots, x_n\}$ and $y = \{y_1, y_2, \ldots, y_n\}$; $\bar{x}$ and $\bar{y}$ are the mean value of x and y, respectively; H(x) and $H(y)$

are the entropy of x and y, respectively, and $H(x, y)$ is their joint entropy; $n_x n_y < B(n)$ and $B(n) = n^{0.6}$.

Development and evaluation of LSTMs

LSTM (described in Appendix A) is an advanced recurrent neural network (RNN) that includes specialized memory blocks (shown in Fig. 3b) which can capture multi-time-step relationships (Read et al. 2019). The difference between LSTM and other ANNs is that the hidden layer in LSTM is constituted of an internal self-looped unit (shown in Fig. 3b) and continuously iterates input of t time steps to extract information (shown in Fig. 3a) (Zhang et al. 2022), which makes LSTM able to learn from the long-term (static) and short-term (dynamic) dependencies raised in time series and can conquer the exploding/vanishing gradient bottlenecks owing to the gradient propagation of the recurrent network over multi-layers (Zhou 2020).

Nash–Sutcliffe efficiency (NSE; Eq. (6)) (Nash and Sutcliffe 1970), one of the most widely used criteria for hydrological modeling (Bennett et al. 2013), and correlation of coefficient (R, Eq. (8)) are models evaluation criteria in this study. Trying to minimize the variance unexplained (FVU, the residual sum of squares divided by the total sum of squares, Eq. (7)), models used the FVU (Eq. (7)) as the loss function (Xiang et al. 2020; Xiang and Demir 2020). The lower the FVU, the higher the NSE. NSE ranges from $-\infty$ to 1, and the closer its value is to 1, the better the model performs. The model performance is acceptable when NSE ≥ 0.65 (Ritter and Muñoz-Carpena 2013), and it performs well when NSE ≥ 0.75 (Moriasi et al. 2007; Tiyasha et al. 2020).

$$\mathrm{NSE} = 1 - \frac{\sum_{i=1}^{n}\left(y_{m,i} - y_{p,i}\right)^2}{\sum_{i=1}^{n}\left(y_{m,i} - \overline{y_m}\right)^2} \tag{6}$$

$$\mathrm{FVU} = \frac{\sum_{i=1}^{n}\left(y_{m,i} - y_{p,i}\right)^2}{\sum_{i=1}^{n}\left(y_{m,i} - \overline{y_m}\right)^2} \tag{7}$$

$$R = \frac{\sum_{i=1}^{n}\left(y_{m,i} - \overline{y_m}\right)\left(y_{p,i} - \overline{y_p}\right)}{\sqrt{\sum_{i=1}^{n}\left(y_{m,i} - \overline{y_m}\right)^2 \sum_{i=1}^{n}\left(y_{p,i} - \overline{y_p}\right)^2}} \tag{8}$$

where n is the number of observations; $y_{m,i}$ and $y_{p,i}$ define the i_{th} observations and the corresponding values predicted by LSTM, respectively; $\overline{y_m}$ represents the average of observations.

Data from each station have been stochastically split into three parts: training (80%), validation (10%), and test (10%) subset. And the same subsets (training/validation/test) of

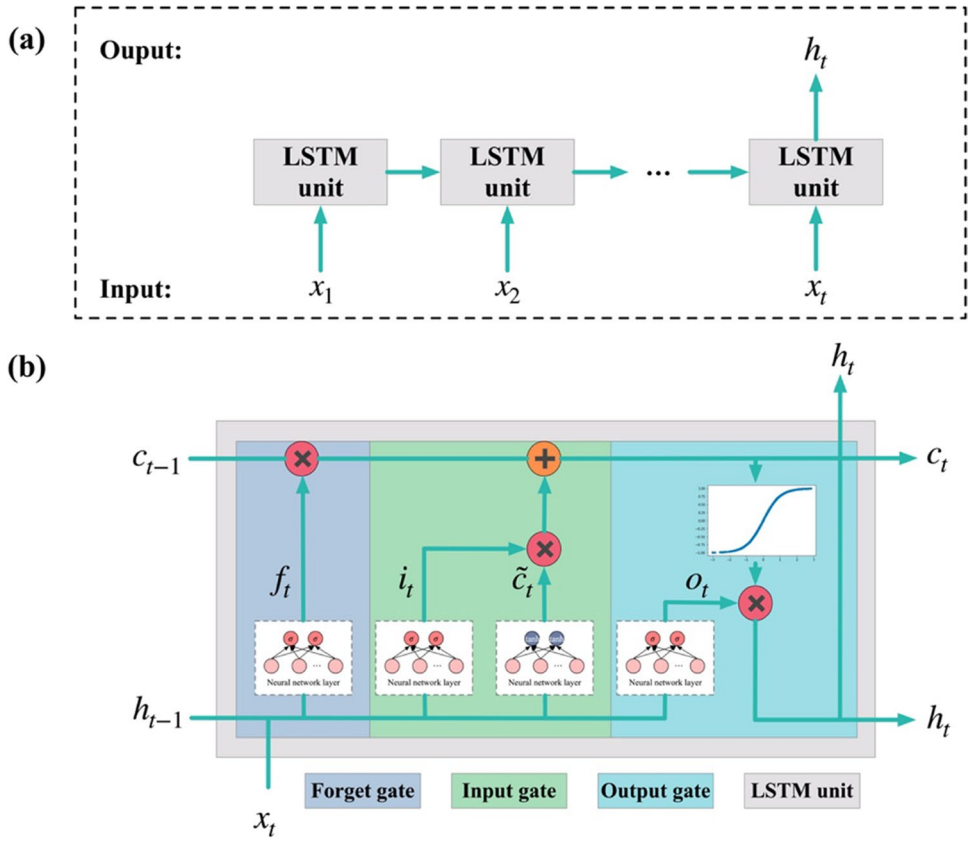

Fig. 3 The workflow and structure of LSTM neural network. **a** One by one spread of LSTM in the time step dimension; **b** conceptual illustration of memory block in LSTM

different stations were combined to form the training, validation, and test dataset. All input variables are standardized according to Eq. (9) to ensure input variables remain on the same scale and to guarantee a stable convergence of parameters in the LSTM.

$$\widetilde{x_i} = \frac{x_i - \bar{x}}{\text{std}} \tag{9}$$

where $\bar{x}$ and std signify the average and standard deviation of the raw observations, respectively; x_i defines every single value of the raw data, while $\widetilde{x_i}$ represents the standardized value.

In this study, LSTM models were developed respectively to forecast the future one-month data of a single water quality indicator. There were two different combinations of input variables (input patterns, shown in Table S2): (a) univariate inputs: only the variable to be predicted could be used as input variable; (b) multivariate inputs: all water quality variables are utilized as input variables. The time steps (k) of model inputs range from 1 to 12 months, which means series data of input variables from the previous 1 to k were used as model inputs. For each LSTM model, Keras Tuner was employed to optimize model hyperparameters by 150 times random search (10 repetitive times training for each random search) from the searching space of hidden layers and neurons (shown in Table S2). In this stage, models were trained on the training dataset, then the optimized model structures were automatically selected by Keras Tuner according to NSEs on the validation dataset. Finally, optimized LSTM models would be trained and tested respectively on the training dataset and testing dataset repeatedly for 100 times to eliminate the effects of uncertainty caused by random initialization of model parameters.

Other hyperparameters of LSTM were determined using trial and error. The batch size and the dropout rate are 256 and 0.5, respectively. The maximal epoch is set to 200, which means model training will end that upon convergence or the epoch reaches 200. Besides, the learning rate is initialized to 0.1 and will be automatically decreased while there is no reduction with the value of loss function on validation dataset for 5 continuous epochs.

Computation environment

All computations were conducted in Python. TensorFlow was employed to build LSTM, while MIC was calculated using the minepy library (Albanese et al. 2013) (alpha = 0.6, $c = 15$). Other Python libraries used are NumPy, pandas, Seaborn, SciencePlots (https://github.com/garrettj403/SciencePlots), Matplotlib (Hunter 2007), and KerasTuner (https://keras-team.github.io/keras-tuner/), which is a python library that can automatically select the optimized hyperparameters of models built with TensorFlow or Keras.

Results and discussion

LSTM performances of prediction for different water quality indicators

The performances of LSTMs and ANNs to predict different water quality indicators are shown in Table 2. The optimal input patterns of LSTMs for predicting NH_3–N, BOD, COD_{Cr}, COD_{Mn}, pH, and TP are univariate inputs, while multivariate inputs for predicting DO. But for ANN, the optimal input patterns are univariate inputs. For predicting NH_3–N, BOD, and COD_{Cr}, LSTM and ANN have the same optimal number of input months, which are respectively 8, 5, and 3 months. However, LSTM needs only 3 months of data of input variables to predict DO well, but ANN needs 11-month data.

Among all LSTM models, LSTMs for predicting COD_{Cr} and COD_{Mn} performed the best with median NSE of 0.837 and 0.835, whereas LSTMs to predict pH are the worst with median NSE of 0.229 among seven water quality indicators.

Table 2 Summary of model performances and optimal hyperparameters for different water quality indicators

Models	Hyperparameters and performances	NH_3–N	BOD	COD_{Cr}	COD_{Mn}	DO	pH	TP
LSTM	Input patterns	a	a	a	a	b	a	a
	Months of input	8	5	3	7	5	6	6
	Optimal neurons	8–20-16	8–16-8	20–12-12	16–8-20	4–12-12	12–20	16–16-12
	Median of NSE	0.638	0.766	0.837	0.835	0.625	0.229	0.711
	Median of R	0.809	0.883	0.919	0.932	0.819	0.495	0.862
ANN	Input patterns	a	a	a	-	a	-	a
	Months of input	8	5	3	-	11	-	5
	Optimal neurons	20–16-8	20–16-8	20–16-12	-	20–20-8	-	20–20-8
	Median of NSE	0.607	0.768	0.839	-	0.563	-	0.542
	Median of R	0.793	0.878	0.920	-	0.794	-	0.876

Other water quality indicators, such as NH_3–N, BOD, DO, and TP, get median NSE of 0.638, 0.766, 0.625, and 0.711, respectively. Comparing the performance of LSTM and ANN, LSTM outperformed ANN when predicting NH_3–N, COD_{Mn}, DO, pH, and TP, while little difference between them in predicting BOD and COD_{Cr}. Additionally, ANNs for predicting COD_{Cr} and pH did not converge. The inconsistency of ANN indicates that it has poor performance compared to the LSTM model. This is because water quality data are time series data, and LSTMs are generally superior to ANN in terms of long-term dependence (Jiang et al. 2021).

Model performances are affected by their input conditions, which decide the information to input into the model (Lv et al. 2020; Maier and Dandy 2000; Tiyasha et al. 2020). Therefore, model performances of different water quality indicators need to be analyzed and compared from the perspective of the input parameters and their lag times.

The influence of different input conditions on LSTM performances

The distributions of NSE values of LSTMs for each water quality indicator in all input conditions are shown in Fig. 4. Model performances measured by correlation of coefficient (Fig. S1) are generally consistent with performances measured by NSE. Models to predict BOD, COD_{Mn}, COD_{Cr}, TP are acceptable since most of their NSE values are bigger than 0.65, and some models performed well (NSE > 0.75). However, the NSEs for NH_3–N, DO, and pH are all below 0.65. This findings show that it is challenging to develop a consistent model for all water quality indicators using LSTM models due to high variations and intrinsic nonlinear correlation among the parameters of the water quality because of the probabilistic nature and chemical procedure of water environment (Najah Ahmed et al. 2019). Thus, different models or more advanced models are supposed to be considered for predicting different variables. On the one hand, when inputting data of more time steps to some extent, model performances for nearly all seven water quality indicators get significantly better than that with 1-time-step input. For example, the optimal number of months of input data for all seven indicators are above 1 (shown in Table 2 and Fig. 4). Additionally, for predicting NH_3–N, BOD, and COD_{Cr}, models with multivariate inputs outperform models with univariate models when inputting observation of only 1 or 2 months. Therefore, providing more historical data, including more time steps and more water quality indicators, enables the model to learn the temporal and internal features of water quality indicators.

On the other hand, inputting more historical data may worsen model performance. For example, comparing multivariate inputs models with univariate inputs models for each indicator except DO, model performances get slightly worse. Besides, models for BOD and COD_{Cr} with 12-month historical data underperform models with 3-month historical data. When more data were input into LSTM, water quality variables with low correlation to outputs variables would bring redundant information, which would interfere with the models from finding the patterns among variables and then weaken the prediction performances (Galelli et al. 2014; Maier et al. 2010; Maier and Dandy 2000). Thus, the identification of key drivers of the model performance may be advantageous for building LSTM models with a simple structure and high forecast accuracy (Liang et al. 2020).

As the learning process of deep learning is to capture the correlation between model inputs variables and outputs variables (Reichstein et al. 2019), the relationship between the temporal autocorrelation/cross-correlation of water quality indicators and the LSTM performance needs further research.

Relationship between autocorrelation/cross-correlation of water quality indicators and LSTM performances

The autocorrelation and cross-correlation of seven water quality variables measured by maximal information coefficient and Pearson coefficient are shown in Fig. S2 and Fig. S3, respectively. Figure 5 shows the influence of the maximum temporal autocorrelation and cross-correlation on the performances of models. The autocorrelation and cross-correlation coefficients of NH_3–N COD_{Mn}, COD_{Cr}, and TP are bigger than these for BOD, DO, and pH, which is generally consistent with the relative relation of model performances among these indicators.

In all two input patterns, NSEs are linearly dependent on the maximum values of temporal autocorrelation and cross-correlation coefficients of water quality indicators measured by MIC, with the R^2 of 0.79 and 0.80. Water quality indicator with more significant autocorrelation or more significant cross-correlation with other indicators can be predicted more accurately by LSTM. Different input patterns of the same correlation metric have almost the same slope in Fig. 5. For example, slopes for input pattern a and b are 1.069 and 0.965, respectively. Nevertheless, much difference exists in intercepts (Fig. 5) which may be caused by these differences between different input patterns.

To some extent, this study also indicated the statistical nature of deep learning algorithms (LSTM in this study) that capture the relationship between model inputs variables and outputs variables (Reichstein et al. 2019). As what model performances are significantly linear related to is the maximum temporal autocorrelation and cross-correlation, the principles and behavior of LSTM were also explained that it can store information of model inputs over long time periods and "remember" the information of the most related

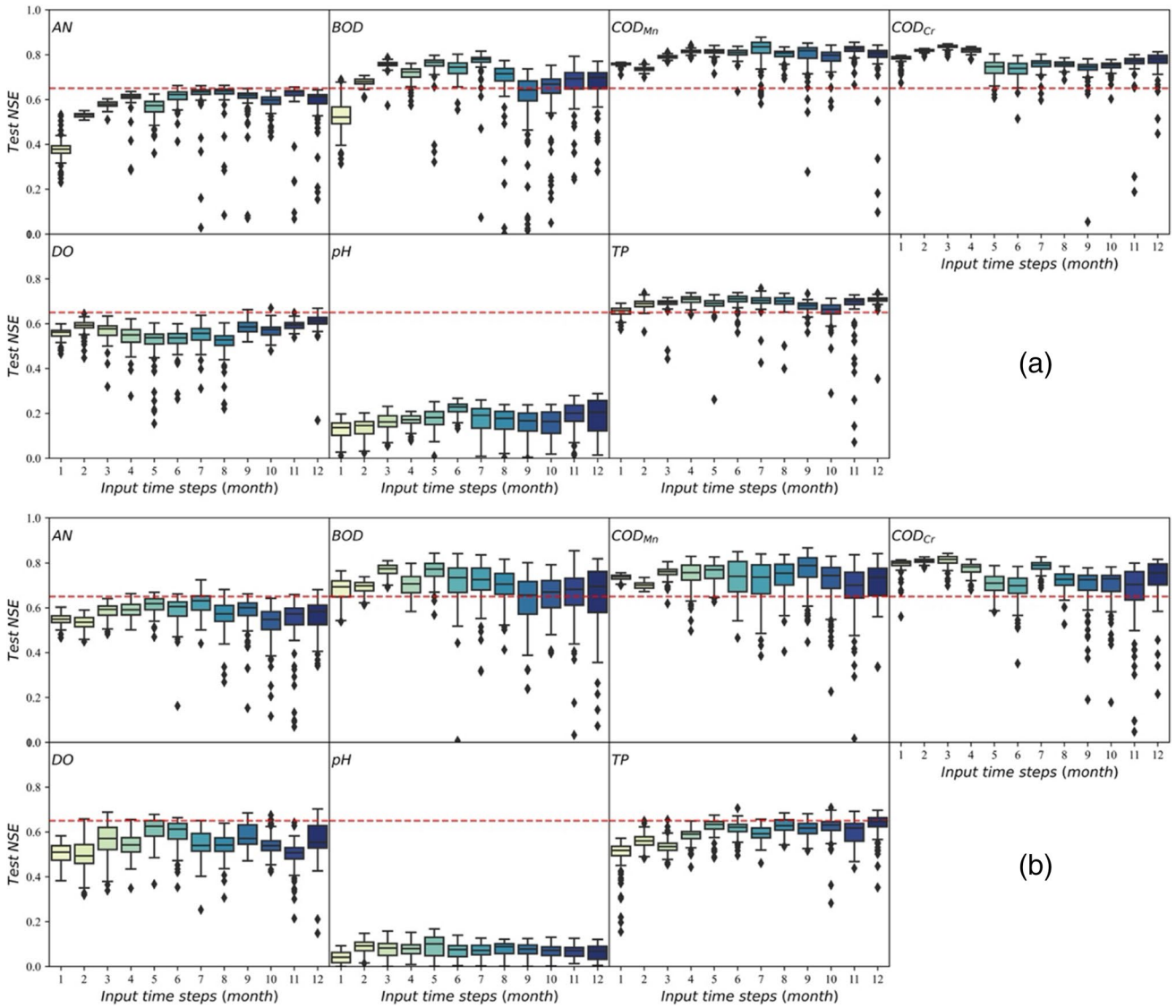

Fig. 4 Distribution of NSE values of LSTMs for each water quality indicator predicted with univariate input (**a**) and multivariate inputs (**b**). The red dotted lines represent the threshold of acceptable model performance (NSE = 0.65)

input values (Zhang et al. 2018). Furthermore, statistical principles such as stochasticity and uncertainty, which have a vital role to play in improving the performance of data-driven models and in defining their areas of applicability, should be considered in the data-driven model-building process (Maier and Dandy 2000).

MIC is a more stable correlation metric than the Pearson coefficient to calculate temporal autocorrelation/cross-correlation coefficients of water quality indicators. Compared to the autocorrelation and cross-correlation of water quality indicators measured by MIC, these correlations calculated by Pearson coefficient are less significantly linear related to model performances (0.79 and 0.77 VS 0.79 and 0.80). It may be caused by that water-related data have properties of nonlinearity, nonstationary, and vagueness due to the unpredictable natural changes, interdependent relationship, and human interference (Najah Ahmed et al. 2019; Tiyasha et al. 2020; Votruba 1988). Since Pearson coefficient is sensitive to singular values (Gnanadesikan and Kettenring 1972) and could only capture the association limited to linear function well, Pearson coefficient is not as stable or suitable as MIC to measure these autocorrelation/cross-correlations in all cases. Hence, with its property of nonlinear statistical dependence (Reshef et al. 2011), MIC are competent for calculating autocorrelation and cross-correlations of water quality variables, which is consistent with results of this study. In addition, mutual information (Babel et al. 2015; Lv et al. 2020) and partial mutual information (Fernando et al. 2009; May et al. 2008; Quilty et al. 2016; Zhou 2020) was also employed widely to determine the dominant input

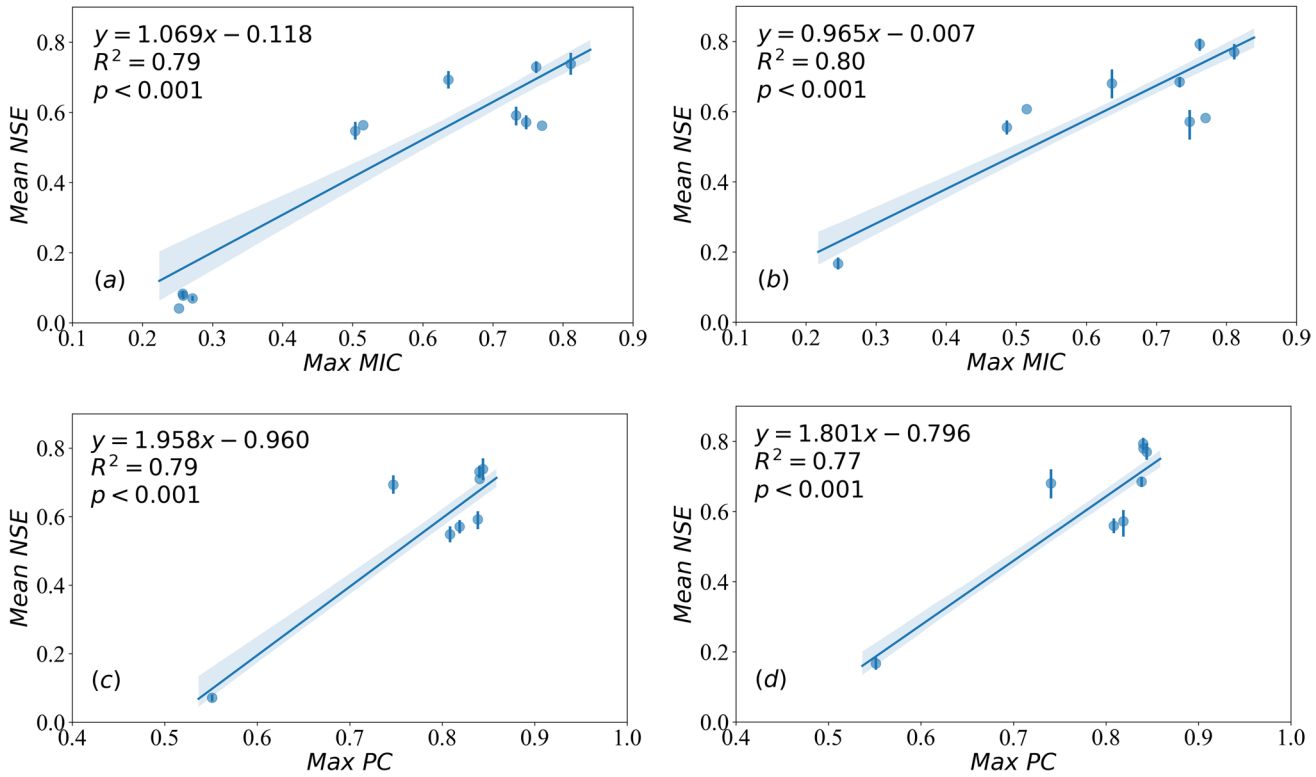

Fig. 5 Scatter plots demonstration of the relationship between performances and the maximum autocorrelation and cross-correlation of water quality indicators calculated by **a** MIC with multivariate inputs; **b** MIC with univariate inputs; **c** Pearson coefficient with multivariate inputs; **d** Pearson coefficient with univariate inputs. R^2, coefficient of determination; p, p-value of statistical significance test for linear regression

variables and their appropriate lag times in water resources. MIC, a better version of MI with higher accuracy, is also suited for LSTM-based model input features selection and their appropriate lag times determination.

Conclusion

This study described the implementation of LSTM to predict water quality indicators in HRB based on monthly observations of water quality indicators. LSTM performance for predicting NH_3–N, BOD, COD_{Mn}, COD_{Cr}, DO, pH, and TP are compared and analyzed. The reasons for the differences in model performance between different water quality indicators were preliminarily explored. Results show that LSTMs for predicting BOD, COD_{Mn}, COD_{Cr}, and TP generally outperform LSTMs for NH_3–N, DO, and pH in HRB. Additionally, LSTM performances are linearly dependent on the maximum temporal autocorrelation and cross-correlation coefficients of water quality indicators measured by MIC.

Although the average NSE of some models could not satisfy the required acceptable level, it is noteworthy that the model performance may be improved by including other water quality or meteorology variables that are not explored in this study or by optimizing the hyperparameters of LSTM. This study focuses on the most basic application of LSTM. Because, whether in research or in application, the higher limit of competence of LSTM to predict water quality is hard to reach, thus, studying the measurement to predict the lower limit of LSTM's predictive power is more meaningful. However, the highest competence of LSTM to predict water quality also needs to be explored specifically. There are also a lot of points which cannot be explained clearly. For example, some models with high input–output correlations performed poorly, and results may not be convincing enough. They are limited by available data and experiments and will be studied in the future work.

Appendix A LSTM model structure

A memory block, whose state at time t is illustrated in Fig. 3, consists of a forget gate, an input gate, a memory cell, and an output gate (Hochreiter and Schmidhuber 1997). In the last computation at time $(t-1)$, both cell state (C_{t-1}) and output (h_{t-1}) are stored by the memory block, and the initial values of C_0 and h_0 are zero. At time t, new inputs (X_t) are available. First, the forget gate, which determines what information to remove,

generates a value (f_t) between 0 and 1 as a basis for determining the extent of allowing C_{t-1} to pass by combing h_{t-1} and X_t into sigmoid function (Eq. (10)). Meanwhile, a new candidate cell state ($\widetilde{c_t}$) and its coefficient can be generated by Eqs. (11) and (12), respectively. Thereafter, the new cell state (C_t) is determined according to Eq. (13). Next, the output gate produces a value (o_t) to determine the parts of the cell state to output based on Eq. (14). Finally, the output is calculated by Eq. (15).

$$f_t = \sigma\left(W_{xf}X_t + W_{hf}h_{t-1} + b_f\right) \tag{10}$$

$$\widetilde{c_t} = \tanh\left(W_{xc}X_t + W_{hc}h_{t-1} + b_c\right) \tag{11}$$

$$i_t = \sigma\left(W_{xi}X_t + W_{hi}h_{t-1} + b_t\right) \tag{12}$$

$$c_t = f_t * c_{t-1} + i_t * \widetilde{c_t} \tag{13}$$

$$o_t = \sigma\left(W_{xo}X_t + W_{ho}h_{t-1} + b_0\right) \tag{14}$$

$$h_t = o_t * \tanh(c_t) \tag{15}$$

In Eqs. (10), (11), (12), and (14), W denotes the matrices of weights for the gates or cells with the corresponding subscripts; b represents learnable biases. Besides, σ and tan$h$ denotes the sigmoid function and the tan$h$ function, respectively.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11356-022-22758-7.

Acknowledgements The authors would like to acknowledge the Hebei Provincial Academy of Ecological Environmental Science, China (http://www.hebhky.cn/) for their data.

Author contribution All authors contributed to the study's conception and design. QL and YW designed all experiments. QL conducted all experiments and analyzed the results. The first draft of the manuscript was written by QL, and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding This work was supported by the National Nature Science Foundation of China (no. 41807471) and the Open Research Fund Program of MNR Key Laboratory for Geo-Environmental Monitoring of Great Bay Area (SZU51029202010).

Data availability Not applicable.

Declarations

Ethics approval The authors express their ethical approval of the contents of the submitted work.

Consent to participate The authors express their consent to have participated in the submitted work.

Consent for publication The authors state that the data used is in the public domain and may be published.

Competing interests The authors declare no competing interests.

References

Albanese D, Filosi M, Visintainer R, Riccadonna S, Jurman G, Furlanello C (2013) Minerva and minepy: a C engine for the MINE suite and its R, Python and MATLAB wrappers. Bioinformatics 29:407–408. https://doi.org/10.1093/bioinformatics/bts707

Antanasijević D, Pocajt V, Povrenović D, Perić-Grujić A, Ristić M (2013) Modelling of dissolved oxygen content using artificial neural networks: Danube River, North Serbia, case study. Environ Sci Pollut Res 20:9006–9013. https://doi.org/10.1007/s11356-013-1876-6

Babel MS, Badgujar GB, Shinde VR (2015) Using the mutual information technique to select explanatory variables in artificial neural networks for rainfall forecasting. Meteorol Appl 22:610–616. https://doi.org/10.1002/met.1495

Bao Z, Zhang J, Wang G, Fu G, He R, Yan X, Jin J, Liu Y, Zhang A (2012) Attribution for decreasing streamflow of the Haihe River basin, northern China: climate variability or human activities? J Hydrol 460–461:117–129. https://doi.org/10.1016/j.jhydrol.2012.06.054

Bennett ND, Croke BFW, Guariso G, Guillaume JHA, Hamilton SH, Jakeman AJ, Marsili-Libelli S, Newham LTH, Norton JP, Perrin C, Pierce SA, Robson B, Seppelt R, Voinov AA, Fath BD, Andreassian V (2013) Characterising performance of environmental models. Environ Model Softw 40:1–20. https://doi.org/10.1016/j.envsoft.2012.09.011

Cao Q, Yu G, Sun S, Dou Y, Li H, Qiao Z (2021) Monitoring water quality of the Haihe River Based on ground-based hyperspectral remote sensing. Water 14:22. https://doi.org/10.3390/w14010022

Dang B, Mao D, Xu Y, Luo Y (2017) Conjugative multi-resistant plasmids in Haihe River and their impacts on the abundance and spatial distribution of antibiotic resistance genes. Water Res 111:81–91. https://doi.org/10.1016/j.watres.2016.12.046

Fernando TMKG, Maier HR, Dandy GC (2009) Selection of input variables for data driven models: an average shifted histogram partial mutual information estimator approach. J Hydrol 367:165–176. https://doi.org/10.1016/j.jhydrol.2008.10.019

Galelli S, Humphrey GB, Maier HR, Castelletti A, Dandy GC, Gibbs MS (2014) An evaluation framework for input variable selection algorithms for environmental data-driven models. Environ Model Softw 62:33–51. https://doi.org/10.1016/j.envsoft.2014.08.015

Gnanadesikan R, Kettenring JR (1972) Robust estimates, residuals, and outlier detection with multiresponse data. Biometrics 28:81. https://doi.org/10.2307/2528963

Graves A, Jaitly N (2014) Towards end-to-end speech recognition with recurrent neural networks. In: Xing EP, Jebara T (eds) Proceedings of the 31st International Conference on Machine Learning. PMLR, Bejing, China, pp 1764–1772

Hamrick JM (1992) A three-dimensional environmental fluid dynamics computer code: theoretical and computational aspects. Special report in applied marine science and ocean engineering ; no. 317.. Virginia Institute of Marine Science, College of William and Mary. 64.

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

Hu C, Wu Q, Li H, Jian S, Li N, Lou Z (2018) Deep learning with a long short-term memory networks approach for rainfall-runoff

simulation. Water (Switzerland) 10. https://doi.org/10.3390/w10111543
Hunter JD (2007) Matplotlib: a 2D graphics environment. Comput Sci Eng 9:90–95. https://doi.org/10.1109/MCSE.2007.55
Jiang Y, Li C, Sun L, Guo D, Zhang Y, Wang W (2021) A deep learning algorithm for multi-source data fusion to predict water quality of urban sewer networks. J Clean Prod 318:128533. https://doi.org/10.1016/j.jclepro.2021.128533
Kim M, Gerba CP, Choi CY (2010) Assessment of physically-based and data-driven models to predict microbial water quality in open channels. J Environ Sci 22:851–857. https://doi.org/10.1016/S1001-0742(09)60188-1
Kratzert F, Klotz D, Herrnegger M, Sampson AK, Hochreiter S, Nearing GS (2019) Toward improved predictions in Ungauged Basins: exploiting the power of machine learning. Water Resour Res 55:11344–11354. https://doi.org/10.1029/2019WR026065
Le XH, Ho HV, Lee G, Jung S (2019) Application of long short-term memory (LSTM) neural network for flood forecasting. Water (Switzerland) 11. https://doi.org/10.3390/w11071387
Li L, Jiang P, Xu H, Lin G, Guo D, Wu H (2019) Water quality prediction based on recurrent neural network and improved evidence theory: a case study of Qiantang River, China. Environ Sci Pollut Res 26:19879–19896. https://doi.org/10.1007/s11356-019-05116-y
Li R (2018) Water quality forecasting of Haihe River based on improved fuzzy time series model. Dwt 106:285–291. https://doi.org/10.5004/dwt.2018.22085
Liang N, Zou Z, Wei Y (2019) Regression models (SVR, EMD and FastICA) in forecasting water quality of the Haihe River of China. DWT 154:147–159. https://doi.org/10.5004/dwt.2019.24034
Liang Z, Zou R, Chen X, Ren T, Su H, Liu Y (2020) Simulate the forecast capacity of a complicated water quality model using the long short-term memory approach. J Hydrol 581. https://doi.org/10.1016/j.jhydrol.2019.124432
Liu X-b, Peng W-q, He G-j, Liu J-l, Wang Y-c (2008) A coupled model of hydrodynamics and water quality for Yuqiao Reservoir in Haihe River Basin. J Hydrodyn 20:574–582. https://doi.org/10.1016/S1001-6058(08)60097-9
Lv N, Liang X, Chen C, Zhou Y, Li J, Wei H, Wang H (2020) A long short-term memory cyclic model with mutual information for hydrology forecasting: a case study in the xixian basin. Adv Water Resour 141. https://doi.org/10.1016/j.advwatres.2020.103622
Maier HR, Dandy GC (2000) Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications. Environ Model Softw 15:101–124. https://doi.org/10.1016/S1364-8152(99)00007-9
Maier HR, Dandy GC (1996) The use of artificial neural networks for the prediction of water quality parameters. Water Resour Res 32:1013–1022. https://doi.org/10.1029/96WR03529
Maier HR, Jain A, Dandy GC, Sudheer KP (2010) Methods used for the development of neural networks for the prediction of water resource variables in river systems: current status and future directions. Environ Model Softw 25:891–909. https://doi.org/10.1016/j.envsoft.2010.02.003
May RJ, Dandy GC, Maier HR, Nixon JB (2008) Application of partial mutual information variable selection to ANN forecasting of water quality in water distribution systems. Environ Model Softw 23:1289–1299. https://doi.org/10.1016/j.envsoft.2008.03.008
Mikolov T, Karafiát M, Burget L, Jan C, Khudanpur S (2010) Recurrent neural network based language model, in: Proceedings of the 11th Annual Conference of the International Speech Communication Association, INTERSPEECH 2010. pp 1045–1048. https://doi.org/10.21437/interspeech.2010-343
Moriasi DN, Arnold JG, Van Liew MW, Bingner RL, Harmel RD, Veith TL (2007) Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. Trans ASABE 50:885–900. https://doi.org/10.13031/2013.23153
Najah Ahmed A, Binti Othman F, Abdulmohsin Afan H, Khaleel Ibrahim R, Ming Fai C, Shabbir Hossain M, Ehteram M, Elshafie A (2019) Machine learning methods for better water quality prediction. J Hydrol 578. https://doi.org/10.1016/j.jhydrol.2019.124084
Nash JE, Sutcliffe JV (1970) River flow forecasting through conceptual models part I - a discussion of principles. J Hydrol 10:282–290. https://doi.org/10.1016/0022-1694(70)90255-6
Palani S, Liong SY, Tkalich P (2008) An ANN application for water quality forecasting. Mar Pollut Bull 56:1586–1597. https://doi.org/10.1016/j.marpolbul.2008.05.021
Pearson K, Lee A (1900) Mathematical contributions to the theory of evolution. VIII. On the Inheritance of Characters not Capable of Exact Quantitative Measurement. Part I. Introductory. Part II. On the Inheritance of Coat-Colour in Horses. Part III. On the Inheritance of Eye-Co. Philosophical Transactions of the Royal Society of London. Series a, Containing Papers of a Mathematical or Physical Character 195:79–150
Quilty J, Adamowski J, Khalil B, Rathinasamy M (2016) Bootstrap rank-ordered conditional mutual information (broCMI): a nonlinear input variable selection method for water resources modeling. Water Resour Res 52:2299–2326. https://doi.org/10.1002/2015WR016959
Read JS, Jia X, Willard J, Appling AP, Zwart JA, Oliver SK, Karpatne A, Hansen GJA, Hanson PC, Watkins W, Steinbach M, Kumar V (2019) Process-guided deep learning predictions of lake water temperature. Water Resour Res 55:9173–9190. https://doi.org/10.1029/2019WR024922
Reichstein M, Camps-Valls G, Stevens B, Jung M, Denzler J, Carvalhais N, Prabhat (2019) Deep learning and process understanding for data-driven Earth system science. Nature 566:195–204. https://doi.org/10.1038/s41586-019-0912-1
Reshef DN, Reshef YA, Finucane HK, Grossman SR, McVean G, Turnbaugh PJ, Lander ES, Mitzenmacher M, Sabeti PC (2011) Detecting novel associations in large data sets. Science 334:1518–1524. https://doi.org/10.1126/science.1205438
Ritter A, Muñoz-Carpena R (2013) Performance evaluation of hydrological models: statistical significance for reducing subjectivity in goodness-of-fit assessments. J Hydrol 480:33–45. https://doi.org/10.1016/j.jhydrol.2012.12.004
Santhi C, Srinivasan R, Arnold JG, Williams JR (2006) A modeling approach to evaluate the impacts of water quality management plans implemented in a watershed in Texas. Environ Model Softw 21:1141–1157. https://doi.org/10.1016/j.envsoft.2005.05.013
Song C, Yao L, Hua C, Ni Q (2021) A novel hybrid model for water quality prediction based on synchrosqueezed wavelet transform technique and improved long short-term memory. J Hydrol 603:126879. https://doi.org/10.1016/j.jhydrol.2021.126879
Sudriani Y, Ridwansyah I, Rustini HA (2019) Long short term memory (LSTM) recurrent neural network (RNN) for discharge level prediction and forecast in Cimandiri river, Indonesia, in: IOP Conference Series: Earth and Environmental Science. Institute of Physics Publishing, p 012037. https://doi.org/10.1088/1755-1315/299/1/012037
Sutskever I, Vinyals O, Le QV (2014) Sequence to sequence learning with neural networks, in: Advances in Neural Information Processing Systems. Neural information processing systems foundation, pp 3104–3112.
Tiyasha, Minh Tung T, MundherYaseen Z (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
Votruba L. (Ed.) (1988) Systems in water resouce management, in: Developments in Water Science, Developments in Water Science. Elsevier, pp 38–86. https://doi.org/10.1016/S0167-5648(08)70921-3

Wang C, Shan B, Zhang H, Zhao Y (2014) Limitation of spatial distribution of ammonia-oxidizing microorganisms in the Haihe River, China, by heavy metals. J Environ Sci 26:502–511. https://doi.org/10.1016/S1001-0742(13)60443-X

Wang Y, Zhou J, Chen K, Wang Y, Liu L (2017) Water quality prediction method based on LSTM neural network, In: Li T, Lopez LM, Li Y (Ed.), 2017 12th International Conference On INTELLigent Systems And Knowledge Engineering (Ieee Iske).

Xiang Z, Demir I (2020) Distributed long-term hourly streamflow predictions using deep learning – a case study for State of Iowa. Environ Model Softw 131:104761. https://doi.org/10.1016/j.envsoft.2020.104761

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

Zhang J, Zhu Y, Zhang X, Ye M, Yang J (2018) Developing a long short-term memory (LSTM) based model for predicting water table depth in agricultural areas. J Hydrol 561:918–929. https://doi.org/10.1016/j.jhydrol.2018.04.065

Zhang L, Zou ZH, Zhao YF (2016) Application of chaotic prediction model based on wavelet transform on water quality prediction. IOP Conf Ser Earth Environ Sci 39:012001. https://doi.org/10.1088/1755-1315/39/1/012001

Zhang X, Jiang HL, Zhang YZ (2012) The hybrid method to predict biochemical oxygen demand of Haihe River in China. AMR 610–613:1066–1069. https://doi.org/10.4028/www.scientific.net/AMR.610-613.1066

Zhang Y, Li C, Jiang Y, Sun L, Zhao R, Yan K, Wang W (2022) Accurate prediction of water quality in urban drainage network with integrated EMD-LSTM model. J Clean Prod 354:131724. https://doi.org/10.1016/j.jclepro.2022.131724

Zheng M, Zheng H, Wu Y, Xiao Y, Du Y, Xu W, Lu F, Wang X, Ouyang Z (2015) Changes in nitrogen budget and potential risk to the environment over 20years (1990–2010) in the agroecosystems of the Haihe Basin, China. J Environ Sci 28:195–202. https://doi.org/10.1016/j.jes.2014.05.053

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

Zhu Y, Drake S, Lü H, Xia J (2010) Analysis of temporal and spatial differences in eco-environmental carrying capacity related to water in the Haihe river basins, China. Water Resour Manage 24:1089–1105. https://doi.org/10.1007/s11269-009-9487-1