water

MDPI

Supplementary Materials

A Data-Driven Framework for Spatiotemporal Analysis and Prediction of River Water Quality: A Case Study in Pearl River, China

Text S1: The structure of the Long Short-Term Memory network (LSTM)

LSTM is a deep neural network architecture dedicated to sequential data, which avoids the gradient vanishing problem when dealing with long-term dependence tasks. An LSTM layer consists of a set of recurrently connected blocks (i.e., memory cells) to store and pass sequential information. Each LSTM memory cell contains three gate units: the forget gate, the input gate and the output gate. The core of the LSTM model is the "cell state", C_t defined in the following equations; this state represents the temporal variation in the memory storage space and enables information to flow (Fig S1). The cell state is controlled by these three gates for the flow of information; the LSTM unit is accessed, written, and cleared by the three self-parameterized control gates. The forward pass of the LSTM model is described by the following key equations (Wu et al., 2021; Zhang et al., 2020):

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

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (2)$$

$$\widetilde{C_t} = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \quad (3)$$

$$C_t = f_t * C_{t-1} + i_t * \widetilde{C_t} \quad (4)$$

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

$$h_t = o_t * tanh(C_t) \quad (6)$$

where f_t, i_t, o_t, and $\widetilde{C_t}$ represent the output vectors of the sigmoid layer with values ranging from 0 to 1; W_f, W_i, W_c, W_o, b_f, b_i, b_c, and b_o define the set of training parameters for the gates; C_t and C_{t-1} represent the new and old cell states (carrying the memory of the current moment and last moment), respectively; x_t is the current input information of this cell; h_{t-1} is the final output of the previous cell; h_t is the final output of the new cell; and tanh is the hyperbolic tangent. Eq. (1) first calculates the forgotten degree f_t at the forget gate, after which Eq. (2) determines which value is to be updated at the input gate; Eq. (3) computes the cell state, Eq. (4) updates the cell state, Eq. (5) decides the output at the output gate and Eq. (6) finally outputs the new hidden state h_t by combining Eqs. (4), (5).

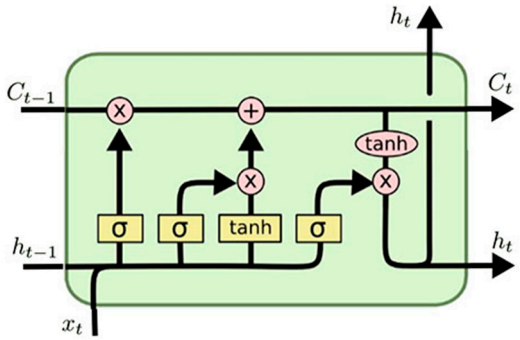

Figure S1. Structure of the LSTM module. Adapted from (Zhang et al., 2020).

The three gates control the information through the cell state. The forget gate determines which information is allowed to pass through and which information is retained. If the output of f_t (representing the forgotten degree) is 0, the state information of the

previous cell will be forgotten and will not be input to the current cell; if the output of f_t is 1, the state information of the previous cell is reserved by the current cell. The input gate determines which new input information can be retained in the cell state, and the output i_t is a control signal to control which new input information can be updated to the cell state. The output gate determines what information to output, where the output o_t, controls the information that needs to be output to the next cell, and the output information is determined by the cell state. Multiple LSTMs can be superimposed and concatenated in time to form more complex structures, thus solving many complex practical sequence problems.

Table S1. The normalization values and weights of water quality parameters used in the WQI calculation, based on the *Environmental Quality Standards for Surface Water of China* (GB 3838-2002) in this study.

Weight[a] **(P_i)**	**Parameters**	**Surface water environmental quality standards**				
		Ⅰ	Ⅱ	Ⅲ	Ⅳ	Ⅴ
		$I_{i,1}$ = 20[b]	**$I_{i,2}$ = 40**[b]	**$I_{i,3}$ = 60**[b]	**$I_{i,4}$ = 80**[b]	**$I_{i,5}$ = 100**[b]
4	DO (mg/L) ≥	7.5	6	5	3	2
3	NH_3-N (mg/L) ≤	0.15	0.5	1	1.5	2
4	TP (mg/L) ≤	0.02	0.1	0.2	0.3	0.4
3	COD_{Mn}(mg/L) ≤	2	4	6	10	15
1	pH			6~9		

[a] The weights are adopted from (Koçer and Sevgili, 2014; Nong et al., 2020), where "1" and "4" presented the parameter that had the lowest and greatest impact on water quality, respectively.

[b] The range of normalized values are adopted from the Environmental Quality Standards for Surface Water (China, 2002).

Table S2. Hyperparameters of LSTM-attention model.

Hyperparameter	**value**
Batch size	16
Dropout rate	0.1
Learning rate	0.001
No. of training epochs	60
No. of LSTM layers	2
No. of neurons in LSTM	128
Loss function	Mean square error (MSE)
Optimization	Adaptive moment estimation (Adam)

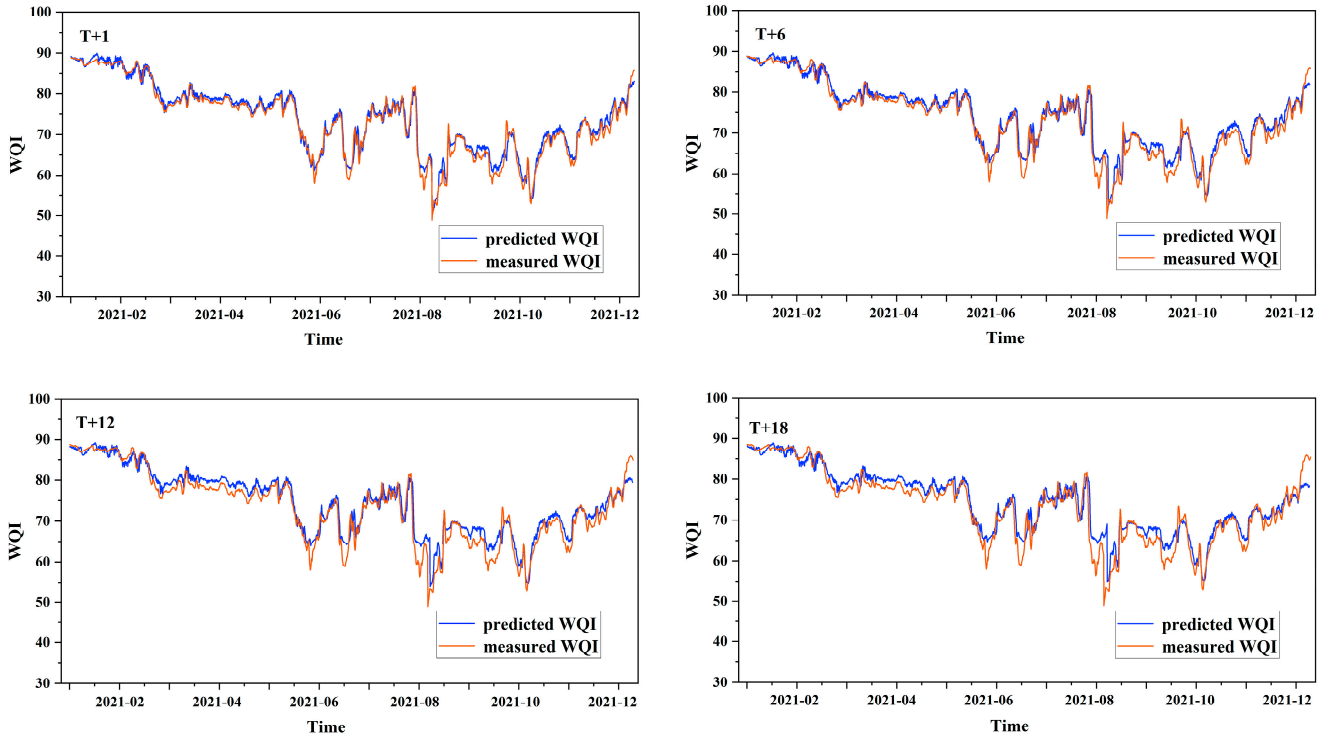

Figure S2. Prediction of the LSTM-attention model at a forecast lead-time of T+1, T+6, T+12, T+18 (Scenario 1).

References:

1. Ministry of Ecology and Environment, PRC. Environmental Quality Standards for Surface Water. 2002. Available online: https://english.mee.gov.cn/Resources/standards/water_environment/quality_standard/200710/t20071024_111792.shtml (accessed on 16 November, 2022).
2. Koçer, M.A.T., Sevgili, H., 2014. Parameters selection for water quality index in the assessment of the environmental impacts of land-based trout farms. Ecol. Indic. 36, 672-681.
3. Nong, X., Shao, D., Zhong, H., Liang, J., 2020. Evaluation of water quality in the South-to-North Water Diversion Project of China using the water quality index (WQI) method. Water Res. 178, 115781.
4. Wu, C., Zhang, X., Wang, W., Lu, C., Zhang, Y., Qin, W., Tick, G.R., Liu, B., Shu, L., 2021. Groundwater level modeling framework by combining the wavelet transform with a long short-term memory data-driven model. Sci. Total Environ. 783, 146948.
5. Zhang, K., Thé, J., Xie, G., Yu, H., 2020. Multi-step ahead forecasting of regional air quality using spatial-temporal deep neural networks: A case study of Huaihai Economic Zone. J. Clean. Prod. 277, 123231.