A new framework for water quality forecasting coupling causal inference, time-frequency analysis, and uncertainty quantification

Chi Zhang [a], Xizhi Nong [a, b, c, *], Kourosh Behzadian [d, e], Luiza C. Campos [d], Lihua Chen [b], Dongguo Shao [a, *]

[a] State Key Laboratory of Water Resources Engineering and Management, Wuhan University, Wuhan 430072, China

[b] College of Civil Engineering and Architecture, Guangxi University, Nanning 530004, China

[c] The National Key Laboratory of Water Disaster Prevention, Nanjing Hydraulic Research Institute, Nanjing 210029, China

[d] Centre for Urban Sustainability and Resilience, Department of Civil, Environmental and Geomatic Engineering, University College London, London WC1E 6BT, United Kingdom

[e] School of Computing and Engineering, University of West London, London W5 5RF, United Kindom

* Corresponding author: Dr. Xizhi Nong; Professor Dongguo Shao

E-mail addresses: nongxizhi@gxu.edu.cn, dongguoshao@163.com

Abstract

Accurate forecasting of water quality variables in river systems is crucial for relevant administrators to identify potential water quality degradation issues and take countermeasures promptly. However, pure data-driven forecasting models are often insufficient to deal with the highly varying periodicity of water quality in today's more complex environment. This study presents a new holistic framework for time-series forecasting of water quality parameters by combining advanced deep learning algorithms (i.e., Long Short-Term Memory (LSTM) and Informer) with causal inference, time-frequency analysis, and uncertainty quantification. The framework was demonstrated for total nitrogen (TN) forecasting in the largest artificial lakes in Asia (i.e., the Danjiangkou Reservoir, China) with six-year monitoring data from January 2017 to June 2022. The results showed that the pre-processing techniques based on causal inference and wavelet decomposition can significantly improve the performance of deep learning algorithms. Compared to the individual LSTM and Informer models, wavelet-coupled approaches diminished well the apparent forecasting errors of TN concentrations, with 24.39%, 32.68%, and 41.26% reduction at most in the average, standard deviation, and maximum values of the errors, respectively. In addition, a post-processing algorithm based on the Copula function and Bayesian theory was designed to quantify the uncertainty of predictions. With the help of this algorithm, each deterministic prediction of our model can correspond to a range of possible outputs. The 95% forecast confidence interval covered almost all the observations, which proves a measure of the reliability and robustness of the predictions. This study provides rich

scientific references for applying advanced data-driven methods in time-series forecasting tasks and a practical methodological framework for water resources management and similar projects.

Keywords: causal inference; Copula function; deep learning algorithms; time-series forecasting; water resources management.

1. Introduction

With the increasing influence of natural events and human activities, water bodies are more vulnerable to drastic changes, making monitoring and protecting water resources particularly critical for the health of humans and the stability of ecosystems (Nong et al., 2020). Accurate forecasting of time-series data related to water quality enables relevant agencies and administrators to comprehend the shifting patterns of water quality parameters and identify potential adverse threats to water bodies (Glibert et al., 2010). Moreover, time-series data forecasting can also help to optimize monitoring programs and resource allocation, improving monitoring efficiency and resource utilization benefits (Li et al., 2018). Therefore, developing and applying reliable models for time-series forecasting is crucial for effective water resources management and environmental protection.

The models widely used for time-series forecasting in water quality management can be generally separated into process-driven and data-driven models. The process-driven models are based on the physical understanding of hydrological processes and water resource systems, using mathematical equations to describe variations in hydrological and water quality processes. Until now, many relevant models have been built, developed, and applied, such as the Water Quality Analysis Simulation Program (WASP), the Environmental Fluid Dynamics Code (EFDC), and the River and Stream Water Quality model (QUAL2K) (Santy et al., 2020). Although process-driven models can provide the understanding and explanatory power of the intrinsic mechanisms of the systems, it is still challenging to determine the boundary condition and calibrate the time-series data for them. Researchers need rich experience with numerical models and comprehensive knowledge of the physic-chemical relationships among water systems (Banerjee et al., 2019). Besides, process-driven models often require detailed geographic and environmental data and rely on the physical assumptions of the system (Wellen et al., 2015). All these factors make such complicated models always data-demanding and time-consuming characteristics to develop in practice.

In recent decades, data-driven models have received more attention due to increasing measurement data and improving computational efforts of computer performance. These models do not rely on a detailed understanding of the physical processes but make predictions by learning patterns and trends in the data (Reichstein et al., 2019). Unlike process-driven models, data-driven models can efficiently establish relationships among different variables. Popular algorithms, including Multiple Linear Regression (MLR), Neural Networks (NN), Support Vector Machine (SVM), and Random Forests (RF), have been widely used for various tasks and have made reliable achievements (He et al., 2020, Xia et al., 2020). Regarding time-series forecasting tasks, deep learning techniques showed remarkable performance due to their adaptability and generalizability to high-dimensional data sequences. Whether the classical structures (e.g., LSTM) or the novel structures (e.g., Informer) leverage the power to capture both short-term and long-term dependencies in data, making them suitable for complex time-series forecasting. As an advanced recurrent network, LSTM has unique memory units and gating mechanisms that enable it to capture long-term dependencies and patterns in data while avoiding the "gradient exploding" problems in the traditional recurrent network (Sit et al., 2019). The application of LSTM in water quality management has been very mature and fruitful. Informer is another advanced deep-learning approach for time-series forecasting tasks. By incorporating self-attention mechanisms and encoder-decoder structure, Informer can effectively model temporal and spatial dependencies in data (Cai et al., 2023). It has demonstrated ability in various domains, such as financial forecasting and energy load prediction (Huang and Jiang, 2022). However, under today's conditions of more detailed requirements and a more complex environment, pure data-driven approaches may often be insufficient (Xiao et al., 2017). A predictive framework integrating multiple and suitable methods is needed. For instance, appropriate data pre-processing techniques are beneficial for harnessing the advantages of the models. In the study on the prediction framework of dissolved oxygen, (Nong et al., 2023) pointed out that feature selection methods can significantly improve the accuracy and robustness of the prediction model. To capture seasonal information in the hydro-climate time series, two types of seasonal LSTM were proposed to simulate the runoff-sediment process (Nourani and Behfar, 2021), showing that the outperformance of seasonal LSTM compared to the individual one in both daily and monthly scales.

Furthermore, relying solely on deterministic predictions may be inadequate for practical water resources management, given the inherent presence of uncertainty. Many researchers have proposed various methods to cope with uncertainty to enhance the ability of predictive models, such as sensitivity analysis or confidence intervals (Hamed et al., 2016, Salimi and Hammad, 2020). In the study of biogas generation, some researchers applied sensitivity analysis to identify the significant factors influencing the biogas, so as to understand and reduce the uncertainty of prediction (Offie et al., 2023). To evaluate the performance of the conceptual basin model, the sensitivity analysis was conducted to determine the uncertain parameters (Tibangayuka et al., 2022). Probabilistic forecasting models with confidence intervals are also one of the common approaches to quantifying the uncertainty of predictions. It can provide a probability distribution for each prediction output instead of just a single deterministic value. For instance, based on a multivariate Bayesian uncertainty processor, (Zhou, 2020) developed a post-processing technique for probabilistic forecasting conditional on point forecasts. Aiming at describing the uncertainty of precipitation forecasts, some studies proposed a new model coupling fuzzy probability and Bayesian theory, which improved the generalization ability of the baseline prediction (Cai et al., 2019). These researchers have quantified the uncertainty well and achieved good results in practice. Decision-makers can better assess the risk and develop strategies by considering uncertainty.

Considering the above gaps and factors, this study developed a predictive framework for time-series tasks based on deep learning approaches coupling various advanced data-processing techniques. The objectives of this study are (1) to explore the applicability of the two state-of-the-art deep learning approaches (i.e., LSTM and Informer) for forecasting of water quality parameters in river systems, (2) to demonstrate the effectiveness of coupling advanced pre-processing techniques, i.e., the causal inference and wavelet decomposition, in improving the performance of forecasting models, (3) to develop a reliable post-processing algorithm for uncertainty quantification of predictions, as a measure for robustness analysis of water quality forecasting. The data matrices comprised of 11 parameters at three stations in the largest artificial lake of Asia (i.e., the Danjiangkou Reservoir in China), were taken as the study cases. The proposed hybrid time-series forecasting framework could also serve as a cost-effective and reliable water quality forecasting tool for water management in the future.

2. Methodology

This study developed a hybrid time-series forecasting framework integrating deep learning approach, causal inference, wavelet decomposition, and Copula function. Of which, causal inference and wavelet decomposition were used as pre-processing tools for time-series data. The LSTM and Informer algorithms were chosen as the models to make predictions, and the Copula function was applied as post-processing technique for uncertainty quantification of outputs. The detailed theoretical introduction of the methodology involved in the framework was shown in **Fig. 1**.

< Fig. 1>

2.1 Causal inference method

This research used the Peter and Clark Momentary Conditional Independence (PCMCI) to identify the causal relationships between variables and conduct feature selection for deep learning models based on the above information. The PCMCI was proposed by (Runge et al., 2015) to assess causal links for a set of temporal lags (τ).

Compared to traditional causal inference methods, the significant advancement of PCMCI is its incorporation of time-varying and autocorrelated relationships. Potential time-dependent system $\boldsymbol{X}_t^j$ for variable *j* at time *t* can be calculated as in **eq. (1)**:

$$X_t^j = f_j(\mathcal{P}(X_t^j), \eta_t^j), \tag{1}$$

where f_j represents the potential nonlinear functional dependency and η_t^j is mutually independent dynamical noise; $\mathcal{P}\left(X_t^j\right) \subset \boldsymbol{X}_t^- = (\boldsymbol{X}_{t-1}, \boldsymbol{X}_{t-2}, \dots \boldsymbol{X}_{t-\tau})$ represents the causal parents of variable X_t^j among the past of all variables. The PCMCI consists of a two-step algorithm as follows:

(1) PC_1 condition selection: PC_1 is a Markov set discovery algorithm based on the PC-stable algorithm (Colombo and Maathuis, 2014), and this method is used to select relevant conditions $\mathcal{P}\left(X_t^j\right)$ for all time-series variables. Specifically, the preliminary parents $\hat{\mathcal{P}}\left(X_t^j\right) = \left(\mathbf{X}_{t-1}, \mathbf{X}_{t-2}, \dots, \mathbf{X}_{t-\tau_{max}}\right)$ are firstly initialised for each variable X_t^j. In the first iteration ($p = 0$), unconditional independence tests are conducted, and $X_{t-\tau}^i$ is removed from $\hat{\mathcal{P}}\left(X_t^j\right)$ if the null hypothesis $X_{t-\tau}^i \perp\!\!\!\perp X_t^j$ cannot be rejected at a significance level α_{PC}. In each next iteration, conditional independence tests ($X_{t-\tau}^i \perp\!\!\!\perp X_t^j | S$, where S is the strongest parents in $\hat{\mathcal{P}}\left(X_t^j\right) \backslash \{X_{t-\tau}^i\}$), are conducted, and all independent parents are removed from $\hat{\mathcal{P}}\left(X_t^j\right)$. If no more conditions can be tested, the algorithm will reach convergence.

(2) Momentary conditional independence (MCI) test: This step addresses false-positive control for the cases where the time series exhibit high interdependence. More precisely, the link $X_{t-\tau}^i \rightarrow X_t^j$ is established if and only if $X_{t-\tau}^i$ and X_t^j are not independent under the condition of $\hat{\mathcal{P}}\left(X_t^j\right) \backslash X_{t-\tau}^i, \hat{\mathcal{P}}_{pX}\left(X_{t-\tau}^i\right)$, where $\hat{\mathcal{P}}_{pX}\left(X_{t-\tau}^i\right) \subseteq$

$\hat{\mathcal{P}}(X_{t-\tau}^{i})$ represents the pX strongest parents based on the sorting in the first step. The MCI test identifies the co-drivers, indirect relationships, and autocorrelation by all selected lagged parents together with contemporaneous pairs. In addition, the significance of each link can be determined based on the p values of the MCI test.

More details about PCMCI can be seen in (Runge et al., 2019b). All the calculations about PCMCI in this study were performed with the help of the Python package **Tigramite** (https://github.com/jakobrunge/tigramite/).

2.2 The development of Wavelet-LSTM and Wavelet-Informer models

2.2.1 The deep learning algorithms

This study applied two popular time-series deep learning algorithms, i.e., the LSTM and Informer. The forms, structures, and characteristics of the algorithms are shown as follows.

2.2.1.1 Long Short-Term Memory network

Long Short-Term Memory is a special-designed recurrent neural network (RNN) architecture that has gained significant popularity in deep learning for time-series analysis. It was initially established to mitigate the vanishing gradient problem of standard RNNs and has demonstrated its powerful capability in capturing long-term dependencies. In an LSTM network, memory cells are used as a replacement for hidden neurons to connect hidden layers. Each memory cell consists of a cell state (C) and three multiplicative gates: the input gate (i), output gate (o), and forget gate (f) (**Fig. S1(a)**). The input gate regulates the new information stored in the current cell based on the current input and the previous hidden state. The output gate determines how much information should be transferred from the current memory cell to the next time step. The forget gate controls the retention of information from the previous state and decides whether information should be retained or be discarded. The information flow regulation of the gates within the network and the detailed algorithms are shown in **eq. (2)** to **eq. (7)**:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f), \tag{2}$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i), \tag{3}$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C), \tag{4}$$

$$C_t = f_t \times C_{t-1} + i_t \times \tilde{C}_t, \tag{5}$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o), \tag{6}$$

$$h_t = o_t \times \tanh(C_t), \tag{7}$$

where W_f, W_i, W_C, and W_o are the weight matrices; b_f, b_i, b_C, and b_o are the bias vectors; σ is the sigmoid function. The LSTM networks can effectively capture the patterns of information over long sequences based on these intricate gating mechanisms, making them particularly suitable for complex time-series forecasting tasks.

2.2.1.2 Informer network

Informer is an improvement of the Transformer model developed by Google for language translation (Vaswani et al., 2017). It combined the strengths of both Transformer networks and convolutional neural networks (CNNs) and was specifically designed to address the challenges of modelling long-term dependencies. Like other competitive neural sequence transduction models, Informer has a multi-layered encoder-decoder structure (**Fig. S1(b)**). The encoder module consists of a stack of self-attention layers, which enables the model to capture global and local dependencies in the input sequence. Each self-attention layer simultaneously attends to different parts of the input sequence through multi-head ProbSparse self-attention mechanisms, which can be briefly described by **eq. (8)**:

$$i\text{-th query's sparsity measurement:} \quad M(\boldsymbol{q}_i, \boldsymbol{K}) = ln \sum_{j=1}^{L_K} e^{\frac{\boldsymbol{q}_i \boldsymbol{k}_j^{\mathrm{T}}}{\sqrt{d}}} - \frac{1}{L_K} \sum_{j=1}^{L_K} \frac{\boldsymbol{q}_i \boldsymbol{k}_j^{\mathrm{T}}}{\sqrt{d}}, \tag{8}$$

where $\boldsymbol{q}_i$ and $\boldsymbol{k}_j$ represent the *i*-th and *j*-th row in query matrix $\boldsymbol{Q}$ and key matrix $\boldsymbol{K}$, respectively. L_K is the size of row for *K*, *d* is the input dimension. The first term stands for the Log-Sum-Exp (LSE) of $\boldsymbol{q}_i$ on all the keys, while the second is their arithmetic mean. The higher $M(\boldsymbol{q}_i, \boldsymbol{K})$ that the *i*-th query has, the more important it is for attention.

Based on the calculated measurement, each key could be allowed to only attend to the *u* dominant queries based on **eq. (9)**:

$$\text{ProbSparse Self-attention:} \quad \mathcal{A}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = Softmax\left(\frac{\bar{\boldsymbol{Q}}\boldsymbol{K}^T}{\sqrt{d}}\right)\boldsymbol{V}, \tag{9}$$

where $\bar{\boldsymbol{Q}}$ is the sparse matrix only containing the Top-*u* queries based on $M(\boldsymbol{q}_i, \boldsymbol{K})$, $\boldsymbol{V}$ is the value matrix.

The decoder module of Informer also utilizes self-attention layers but with an additional cross multi-head attention mechanism. The cross multi-head attention mechanism allows the decoder to interact with the encoder's outputs, enabling it to connect the global context and employ the learned representations from the encoder, which further facilitates accurate and context-aware predictions in the decoding process. Residual connections and layer normalization are designed in both encoder and decoder modules, which help improve the flow of gradients and stabilize the training process. In addition, a feed-forward neural network and a positional encoding component are also involved in Informer to strengthen its modelling capacity. Therefore, the comprehensive combinations of transformer networks and CNNs within the Informer maintain the model's versatile and powerful forecasting capacity, capturing both short-term and long-term patterns. Those unique combinations and the incorporation of ProbSparse self-attention make the Informer a promising approach for various time-series forecasting tasks.

2.2.2 Wavelet decomposition

Wavelet decomposition is a powerful mathematical tool in signal theory. It is used for decomposing signals into different frequency components for analysis and overcomes the limitations of Fourier transformation in non-stationary time series (Labat, 2005). By decomposing the main time series into the time-frequency space, several sub-series could be obtained to extract particular time and frequency characteristics simultaneously. The sub-series are typically derived from a predefined template called the “mother wavelet”, in which these decomposed wavelets are obtained by scaling and translating the mother wavelet. For the calculations, continuous wavelet decomposition (CWD) requires integral operations in continuous time, which may result in computational complexity and memory consumption. In contrast, discrete wavelet decomposition (DWD) utilizes a fixed-length filter, which has the advantages of high computational efficiency and low memory consumption, making it more adopted in practical applications (Cannas et al., 2006). The discrete wavelet decomposition for series $f(t)$ is organized based on **eq. (10)** and **eq. (11)**:

$$\text{DWD coefficients:} \quad W_f(i,j) = \sum_{i,j\in Z} f(t)\Psi_{i,j}^*(t), \tag{10}$$

$$\text{Wavelet function:} \quad \Psi_{i,j}^*(t) = a_0^{-\frac{i}{2}}\Psi\left(a_0^{-j}t - b_0 k\right), a_0 > 1, b_0 > 0, \tag{11}$$

where i and j are the integers which control the decomposition level and translation, respectively. a_0 and b_0 are the constant scale factor of decomposition and position factor of translation, respectively. $\Psi(t)$ is the mother wavelet. Then the main series can be decomposed into a low-frequency approximation sub-series (A_n) and some high-frequency detail sub-series (D_1, D_2, …, D_n) based on low-pass filter and the high-pass filter.

2.2.3 Model development

The hybrid Wavelet-LSTM (WLSTM) and Wavelet-Informer (WInformer) were developed by combining LSTM and Informer with the wavelet decomposition, which refers to (Liu et al., 2022). The process is divided to three steps: (1) the wavelet decomposition of the original series of the predictand; (2) the prediction of each sub-series using LSTM and Informer individually; and (3) the re-composition of each output series for the final results.

To appropriately train the deep-learning models within the WLSTM and Winformer structure, our procedure involved two phases: (1) calibration and (2) evaluation. In the calibration phase, the first 70% of original data were used to develop the deep-learning models, while the following 10% were used as a validation set to avoid over-fitting. After the calibration phase, the parameters with the model performance within the validation were saved for the evaluation phase, in which the trained model performance is tested based on the remaining 20% of the data. The model performances for in-sample and out-of-sample datasets were evaluated in the calibration phase (i.e., the entire establishing data) and the evaluation phase (i.e., the unused data), respectively.

In this study, the LSTM and Informer models were implemented in *Python*. The grid-search method was used to tune the hyperparameters of deep-learning algorithms (all the results were listed in **Table S1** and **S2** in Supplementary Materials). As for wavelet decomposition, we selected the Daubechies-4 (db4) as a mother wavelet to decompose the main series into three levels due to its high-efficiency spectral properties (Nourani et al., 2014b). The DWD procedures were performed with the help of **Wavelet Toolbox** in *Matlab*.

2.3 Uncertainty forecast based on Copula function and Bayesian theory

According to (Challinor et al., 2013), uncertainty refers to the lack of predictive accuracy due to inherent limitations in predictability or a lack of predictive skills. In practice, estimating prediction uncertainty means estimating how predictions are distributed around the observations. In the last step of the prediction framework, we employed the Copula function and Bayesian theory to conduct uncertainty forecasts.

The Copula function is a widely used statistical tool for modelling and analyzing dependencies between random variables. The main idea of the Copula function is to treat the marginal distribution of variables and their correlation structure separately, thus providing a flexible way to describe their interrelations. According to the Sklar theory (Sklar, 1959), if the marginal distributions of the bivariate joint distribution H are F_x and F_y, respectively, there is a Copula function for any $x, \ y \in R$ as expressed by **eq. (12)**:

$$H(x, y) = C(F_x(x), F_y(y)), \tag{12}$$

Based on this theoretical foundation, the joint distribution of two variables can be constructed in just two steps. Firstly, determining the marginal distributions of the variables, and secondly, selecting the optimal Copula function to depict the dependency structure between the variables accurately. More details about Copula theory can be found in (Größer and Okhrin, 2021).

This study established the joint distribution of predictions and observations based on the Copula function. Then the probabilistic forecasting could be conducted according to Bayesian theory. The process to achieve the uncertainty forecast is described as follows:

(1) Fitting the marginal distributions of the *Prediction* $\boldsymbol{X}$ and *Observation* $\boldsymbol{Y}$ based on the predictions $\boldsymbol{X_{cali}} = (x_1, x_2, \dots, x_n)$ and observations $\boldsymbol{Y_{cali}} = (y_1, y_2, \dots, y_n)$ in the calibration phase. Then, the cumulative probability u of data in different sets can be obtained by probability transformation based on **eq. (13)**:

$$u_{set,1i} = F_{x,set}(x_i) \ or \ u_{set,2i} = F_{y,set}(y_i), \tag{13}$$

Where $set = (cali, eval)$ denotes calibration or evaluation phase; $F(\cdot)$ refers to the marginal distribution of the corresponding object (*Prediction* $\boldsymbol{X}$ or *Observation* $\boldsymbol{Y}$).

(2) Constructing the joint distribution of the *Prediction* $\boldsymbol{X}$ and *Observation* $\boldsymbol{Y}$ by using Copula function to connect the cumulative probability $u_{cali,1i}$ and $u_{cali,2i}$. Several types of bivariate Copula function used in this work are presented in **Table S3**.

(3) Given the probability value p the conditional distribution function of a bivariate Copula by **eq. (14)**:

$$H_1(u_2|u_1) = \frac{\partial C(u_1,u_2)}{\partial u_1}, \tag{14}$$

The probabilistic forecasting values $\tilde{y}_j$ in the evaluation phase was calculated based on inverse conditional probability function $\tilde{u}_{eval,2j} = H_1^{-1}(u_{eval,1j}, p)$ and inverse cumulative probability function $\tilde{y}_j = F_y^{-1}(\tilde{u}_{eval,2j})$. In other words, if we calculate the probabilistic forecasting values corresponding to the conditional probability of 2.5% and 97.5%, the 95% forecast confidence interval for the deterministic predicted value could be obtained.

3. Case study

3.1 Study area and data collection

The Danjiangkou Reservoir (DJKR) is located at the junction of Hubei and Henan provinces, China, covering the areas of 32°36′-33°48′ N and 110°59′-111°49′ E (**Fig. 2**). It serves as a vital drinking water source of the Middle Route of the South-to-North Water Diversion Project of China (MRSNWDPC) since December 2014, providing 9.5×10^9 m^3 of freshwater water resources through the main canal of the MRSNWDPC to North China every year. The DJKR currently stands at a height of 176.6 m, maintaining an average impounded level of 170 m and possessing a storage capacity of 29.05 billion m^3. The reservoir falls within the northern subtropical zone and experiences a subtropical monsoon climate, with the average annual air temperature ranging from 15-16 °C, and the annual precipitation ranging from 800-1,000 mm.

In order to effectively monitor and protect the water resources in the DJKR, the Chinese government has undertaken national water quality monitoring programs. The data of this study was obtained from three key national automatic water quality monitoring stations, i.e., the Taocha (TC), Qingshan (QS), and Madeng (MD) stations. The TC is located at the starting point of the MRSNWDPC, and the QC and MD are located at the entrance point of the two main tributaries of the DJKR, i.e., Hanjiang River and Danjiang River, respectively (**Fig. 2**). The daily data used in this analysis were collected for seven water quality parameters, including water temperature (WT, °C), pH, dissolved oxygen (DO, mg/L), conductivity (Cond, μS /cm), chlorophyll-a (Chl-a, mg/L), total phosphorus (TP, mg/L), and total nitrogen (TN, mg/L) from January 2017 to June 2022. As the potential adverse trend of TN in the Danjiangkou Reservoir is particularly concerning (Liu et al., 2017), TN was considered as the main forecasting water quality parameter in this study. Additionally, three atmospheric parameters (i.e., nitrogen dioxide (NO_2, μg/m^3), nitrogen monoxide (NO, μg/m^3), and nitric acid (HNO_3^-, μg/m^3)) and precipitation (Pre, mm) were collected from the Copernicus Atmosphere Monitoring Service (CAMS) global reanalysis monthly averaged fields to establish the predictive framework for TN (https://ads.atmosphere.copernicus.eu/). A summary of the statistical characteristics of these parameters are shown in **Table 1**.

<Fig. 2>

<Table 1>

3.2 Model evaluation

To evaluate the predictive effects of our models, the Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE), and coefficient of determination (R^2) were used:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}}, \tag{15}$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}_i - y_i}{y_i}\right| \times 100\%, \tag{16}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(\bar{y} - y_i)^2}, \tag{17}$$

where n is the number of data points; $\hat{y}_i$ and y_i are the i-th prediction and observation, respectively; $\bar{y}$ is the mean of y_i.

In addition, the Coverage Rate (CR) and Average Relative Interval Length (ARIL) were used to assess the results of the uncertainty forecast:

$$CR = \frac{\sum_{i=1}^{n} I(\tilde{y}_{lo,i} < y_i < \tilde{y}_{up,i})}{n}, \tag{18}$$

$$ARIL = \frac{1}{n}\left(\sum_{i=1}^{n}\frac{\tilde{y}_{up,i} - \tilde{y}_{lo,i}}{y_i}\right), \tag{19}$$

where n is the number of data points; $\tilde{y}_{up,i}$ and $\tilde{y}_{lo,i}$ denote the upper and lower boundary of the forecast confidence interval for the i-th prediction, respectively; y_i is the i-th observation; $I(\cdot)$ is the indicator function.

4. Results

4.1 Prediction models with and without causal inference

The PCMCI was applied for feature screening in the prediction models, and the causal networks of indicators in different stations are shown in **Fig. 3**. The parameter τ_{max} was set as two days, indicating that a parent process earlier than two days would not be considered. For the predictand, the features that significantly impacted TN were investigated according to **Table S4**. The results revealed a strong autocorrelation of TN across all monitoring stations, meaning that the TN concentrations observed two days prior significantly affected the concentrations measured on the current day. Cond had a direct impact on TN in TC and QS stations, while DO had that on TN in TC and MD stations. NO_2 had a one-day delay effect on TN in the TC station and a direct impact on the QS station, respectively. The concentrations of TP showed a two-day delay effect on TN in the TC station. For the QS and MD stations, the Chl-a and WT showed different multi-day delay effects on TN, respectively. Based on the PCMCI, the features for predicting TN in different stations were selected (**Table 2**).

< Fig. 3>

<Table 2>

The performance of the LSTM and Informer models with PCMCI for water quality forecasting was compared with the models without PCMCI as shown in **Fig. 4**. More specifically, the LSTM and Informer models without PCMCI (i.e., NO_LSTM and NO_Informer in the figure) involved all parameters from two days ahead to the current day as inputs (3×11-1=32 features). In contrast, PCMCI_LSTM and PCMCI_Informer involved selected features as inputs. As shown in **Fig. 4**, the predictions versus observations across all monitoring stations were distributed around a 1:1 slope line in both Pre1 and Pre2 models. All the $R^2_{Pre1-Pre2}$ were higher than 0.85, indicating that reducing the number of inputs did not decrease forecasting performance. Furthermore, the model performance when using PCMCI was better than that without PCMCI in both models and three stations (**Table 3**), with the highest improvement rates of 22.88%, 24.79%, and 11.59% in terms of RMSE, MAPE, and R^2, respectively. These phenomena indicated a practical application of PCMCI for saving the indicator measurement cost and improving the prediction efficiency.

< Fig. 4>

<Table 3>

4.2 Prediction models with and without wavelet decomposition

Based on the results of Section 3.1, our following model simulations all took the features selected by PCMCI as inputs. In this section, the predictive effects of the LSTM and the Informer models with or without wavelet decomposition were compared for the single-step prediction task. The WLSTM and the WInformer approaches were developed and verified on the daily TN dynamics in each station. As shown in **Fig. S2** to **S4**, the TN concentrations in the Danjiangkou Reservoir presented a common fluctuation trend. Although the LSTM and the Informer models successfully captured the overall variations of TN in these non-stationary signal modes, they exhibited unsatisfactory performance at some local mutation points. For instance, several sharp changes occurred from the 290th to 350th day of TC and from the 90th to 180th day of QS, causing significant simulation errors to the LSTM and Informer model (**Fig. 5**). Besides, the forecasting performance of the LSTM and the Informer showed a minor difference in the single-step prediction for the full sequence in terms of R^2 statistic (0.8430 vs. 0.8463 in TC, 0.8568 vs. 0.8423 in QS, 0.8511 vs. 0.8120 in MD, respectively).

When coupled with the wavelet decomposition, the performance of the WLSTM and WInformer both improved with an increase of 0.17% to 10.37% compared to the original model for the entire sequence in terms of R^2 statistics. The daily original TN series (S) were decomposed to an approximation coefficient (A_3) and three levels of detailed coefficients (D_1 - D_3). The A_3 contains the low-frequency components of the signal and approximates the signal with reduced detail, while the D_1 - D_3 captures the high-frequency components of the signal at different scales and provides progressively finer details. Compared with the LSTM and the Informer, the apparent simulation errors of TN concentrations were smoothed and diminished by the WLSTM and WInformer. The wavelet decomposition coupled methods presented accurate predictions of the extreme situations, with around 24.39%, 32.68%, and 41.26% reduction at most on the average, standard deviation, and maximum of the prediction errors (**Table S5**). Moreover, further comparison proved the best forecasting performance of the WInformer at all the stations over the other three models, as shown in **Table 4** and **Fig. S5**. The highest accuracy of WInformer was reached at the evaluation phase of the MD station, shown by its smallest RMSE (0.0472 mg/L), lowest MAPE (2.85%), and highest R^2 (0.9400). In addition, the improvement rates of the Winformer model over the other three models in the evaluation stages are 14.83% to 27.38%, 15.37% to 24.39%, and 5.74% to 9.12% in terms of RMSE, MAPE, and R^2, respectively. All the results indicated that the developed hybrid WInformer method could reliably accomplish single-step prediction tasks based on historical data.

<Fig. 5>

<Table 4>

4.3 Uncertainty quantification for prediction

The uncertainty forecast is based on the selection of the best forecasting model. Following the process described in Section 2.4, we first fitted the marginal distributions of observations and predictions of TN in the calibration stages for all sites using Pearson III distribution (**Table S6**), a popular and important distribution in the field of water resources. Then, the joint distribution of the observations-predictions pair for each station was established based on the marginal distributions and the Copula theory (**Table S7**). Through the probability transformation of the predictions in the evaluation stages and calculations based on Eq. (13) and Eq. (14), we can obtain any quantiles of the probability prediction (uncertainty prediction). In this study, given the significance level $\alpha = 0.05$, the 2.5$^{\text{th}}$ percentile and 97.5$^{\text{th}}$ percentile of the posterior conditional probability distribution were calculated, corresponding to the lower and upper boundary of the 95% forecast confidence interval, respectively. Thus, each deterministic prediction result of the WInformer was associated with a corresponding forecast interval, achieving the uncertainty quantification. As shown in **Fig. 6**, the forecast interval covered almost all the observations at the evaluation phase, indicating that the probabilistic forecast is reliable. Besides, CR and ARIL were used to evaluate the results of the probabilistic forecast. The larger the CR, the higher the proportion of the observations covered by the forecast interval, while the smaller the ARIL, the narrower the average relative interval width of the forecast interval and the higher the accuracy. Studies have shown that as CR increases, ARIL also increases, meaning these two metrics are often contradictory. For a given confidence level, under the premise of ensuring a high coverage rate, the narrower the average relative width of the forecast interval, the better the prediction performance. It can be seen in **Fig.6** that CR remained above 90% at all stations, with the highest being 98.71% of the MD station. ARIL remained only around 20% across stations, with the smallest being 18.01% of the TC station. These results indicated that our uncertainty forecast is reliable and can provide more information for water resources management decisions.

<Fig. 6>

5. Discussion

5.1 Model improvement brought by causal inference and wavelet decomposition

Selecting the most relevant and informative features from all available features can improve data-driven models' predictive performance and explanatory power (Masmoudi et al., 2020). Driven by the need to establish more efficient, interpretable, and reliable models, causal inference was integrated into the forecasting framework in this study. It has advantages in enhancing forecasting accuracy, boosting computational efficiency, and providing insights into mechanisms Specifically, the causal inference can identify direct causal relationships between the features and the target variable while excluding indirect relationships caused by the presence of confounding variables; this facilitates the construction of more interpretable and reliable models (Pearl and Mackenzie, 2018), and has recently gained significant popularity across various fields (Kretschmer et al., 2018, Krich et al., 2022). As one of the advanced causal inference methods, the core technique of PCMCI is to infer causal relationships by evaluating conditional independences of variables, which do not need to rely on traditional path analysis of causality models or causal hypotheses. Because of this, this method can handle the linear relationship and capture the nonlinear causality to better adapt to the complexity and dynamics of the actual data (Runge et al., 2019a). In addition, high-dimensional and strongly autocorrelated data can be efficiently processed, and the lag-dependent temporal relationships can be found based on the PCMCI, which makes it very applicable for dealing with time-series-related problems (Krich et al., 2020). This study selected indicators with specific time lags as the input features based on PCMCI. It can be seen from the screening results (**Table 2**) that PCMCI not only selects the index set that meets the physical mechanism but also significantly reduces the dimensionality of the input data (from 32 features of the model without PCMCI to 5/6 features of the model with PCMCI). It has been verified that the complexity of the model increases with increasing input, potentially leading to the problem of low efficiency and overfitting (Wang et al., 2023). Our results have presented consistent conclusions: the models with selected features all showed better forecasting performance. These phenomena indicate a valuable application of PCMCI for saving indicator measurement costs and improving prediction efficiency.

Wavelet decomposition was also used to enhance the model in this study. Compared to the individual deep learning model, the forecasting performance of TN by the wavelet-coupled approaches was improved at all stations, with a maximum decrease of 24.75% and 23.25% in terms of RMSE and MAPE, respectively (**Fig. 5**). In the hybrid structures, the wavelet decomposition played a crucial role as an effective pre-processing tool. It extracted cyclic signals using dyadic decompositions, from which the extracted sub-series could exhibit distinct multi-timescale characteristics of the original series quasi-periodically and periodically (Nourani et al., 2014a). This feature greatly facilitated the utilization of deep learning algorithmic advantages in handling time series tasks. Furthermore, the wavelet-coupled approaches were also remarkably effective in simulating peak values with TN dynamics (**Fig. 5** and **Table S5**). Generally, it is quite difficult for data-driven models to accurately predict extreme situations, as they often treat extreme points as outliers before their normal prediction process (Song et al., 2021). However, by incorporating the robust resistance and smoothing capability of wavelet decomposition, the wavelet-coupled approaches effectively reduce the inclusion of extreme components in the input sub-series. The likelihood of models detecting original outliers is then reduced, while the fitting accuracy for well-transformed mutations is increased (Du et al., 2018). Danjiangkou reservoir basin has multiple and complex sources of pollution, resulting in sharp changes in TN dynamics

(Zhang et al., 2023). The accurate forecasting performance for mutations is absolutely useful for water quality management.

5.2 Necessity and potential of uncertainty prediction

In the past, it was common in most practical engineering management to make decisions based on the deterministic forecast values obtained from models. However, due to the inherent limitations and uncertainties present in real-world phenomena and data, the predictions made by the models are also uncertain (Krzysztofowicz, 1999). According to statistical decision theory, when making decisions without considering the uncertainty of the predictions, the value of the model forecasts in the decision-making process may not be non-negative in terms of expectation (Berger, 2013). In other words, the value of the model forecasts can remain positive only when the uncertainty of the predictions is considered in decision-making. The decision maker is responsible for deciding upon a reasonable water resources management course of action based on the forecaster, relying solely on a single-point estimate of the predictand may be insufficient (Kelly and Krzysztofowicz, 2000, Yang, 2020). Therefore, quantifying the uncertainty associated with the predictions regarding probability distribution and confidence level is necessary.

In this study, the Copula function was used to establish the joint distribution of observations and deterministic predictions to quantify the distribution of errors. Copula function is a statistical tool used to establish the structure of correlations between random variables (Dai et al., 2020). This approach can help us to better understand and model the dependencies between variables and provide more accurate results in uncertainty assessment, simulations, and predictions. It was widely used in finance, climatology, and risk management in the early years and has recently gained popularity in water resources (Sahoo et al., 2020, Zhi et al., 2022). The study of (Liu et al., 2018) analysed the effect of compound floods in Texas, USA, based on the Copula function with precipitation, surface runoff, El Nino-Southern Oscillation (ENSO) states, and rising temperatures as underlying conditions. Aiming at the potential abnormal algal proliferation in the MRSNWDPC, some scholars modelled dependency structures of water quality and hydrodynamic factors and conducted risk analysis based on Copula theory (Zhang et al., 2021). In addition, a Copula-based Bayesian network method was proposed and proved to be a powerful decision-support tool for the water quality management of Yuqiao Reservoir (Yu and Zhang, 2021). These studies reveal the power and flexibility of the Copula function, and the structure of Copula can well characterize the relationship between the variables. With the help of the Copula function and Bayesian theory, each deterministic prediction of our model can correspond to a range of possible outputs. The results also showed that the forecast interval covered almost all the observations, indicating that our method is reliable (**Fig. 6**). This range of possibilities reflects the inherent randomness and variability in the underlying processes and model establishment, which provides a measure of the reliability and robustness of the predictions. Such information is valuable in practical engineering management. By considering uncertainty, decision-makers can evaluate the level of uncertainty associated with different scenarios and adjust their strategies accordingly.

5.3 Contributions, challenges, and future work

Data-driven methods are being increasingly appreciated in the context of detailed real-world observations (Zhong et al., 2021). Various deep learning algorithms have been widely applied in time-series prediction research (Deng et al., 2021, Harris and Graham, 2017). This study involves two popular time-series deep learning algorithms, i.e., the LSTM and Informer. LSTM is known for its excellent long-term dependency modelling ability to capture temporal relationships in sequence data efficiently (Zheng et al., 2021). It has demonstrated capacity in the field of water resources. In contrast, as a newly proposed algorithm, the application of the Informer in this field is relatively limited. As an improvement of the Transformer, Informer is a model based on the self-attention mechanism that can effectively utilize the temporal and spatial correlation information within time-series data (Gong et al., 2022). In the study on short-term irrigation water use forecasting, (Zou et al., 2022) demonstrated the superiority of Informer over the other five data-driven methods. Based on long-term monitoring data and Informer, some researchers developed an effective prediction framework for water quality management (Yao et al., 2022). Our results also showed the best forecast performance of WInformer at all stations (**Fig. S5**), indicating the great potential of Informer in water quality prediction. These experiments enrich the application of Informer in the field of water resources. Besides, various advanced methods such as PCMCI, wavelet decomposition, and Copula function were used to improve the performance of deep learning algorithms in this research. We aimed to provide a more accurate and reliable framework to analyse and predict complex time-series data, providing strong support for applications in related fields and tasks.

There remains a substantial scope for future exploration and investigation in this domain. First, due to the funding constraints, the resolution of data monitoring in this study is only on a daily scale. Water resources management sometimes requires to be conducted on an hourly scale, so it is crucial to continue studying related models in the future. Second, although we selected the index set that meets the physical mechanism based on PCMCI, more detailed studies on the mechanism of water quality variation are still of concern. Considering that the DJKR will continue to operate for many years, specific research on models driven by physical-mathematical equations will be carried out in the future. Third, designing individual or ensemble deep learning models for multi-steps time-series prediction tasks has been an emerging area in recent years. Based on the sing-step forecasting framework we established, the results of multi-step ahead forecasting using alternative approaches, such as recursive- or batch- pattern model sets would be reported in our future work, aiming to develop more accurate and robust long-term forecasting models.

6. Conclusions

In this study, we developed a hybrid time-series forecasting framework integrating deep learning approach, causal inference, wavelet decomposition, and Copula function, which was used for TN prediction of the Danjiangkou Reservoir of China. The main conclusions are as follows:

(1) PCMCI is a powerful feature selection method based on causal inference. It can not only select the index set that meets the physical mechanism, but also significantly reduce the dimensionality of the input data. Our results demonstrated its ability to save indicator measurement costs and improve prediction efficiency.

(2) Compared to the individual models, the apparent forecasting errors of TN concentrations were well smoothed and diminished by the wavelet-coupled approaches, with 24.39%, 32.68%, and 41.26% reduction at most on the average, standard deviation, and maximum of the prediction errors. Furthermore, WInformer showed the best performance in all the experiments, indicating this new structure's valuable potential in water quality management.

(3) With the combinations of the Copula function and Bayesian theory, each deterministic prediction of our model can correspond to a range of possible outputs, which measure the reliability and robustness of the predictions. By considering uncertainty, decision-makers can evaluate the uncertainty associated with different scenarios and adjust their strategies accordingly.

This study provides insights for applying advanced data-driven methods in time-series forecasting tasks and a practical methodological framework for water resources management and similar projects. In future research, long-term series monitoring data, various mechanism models, and more in-situ/ computational experiments are still needed to be conducted.

Declaration of Interest Statement

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 research was funded by the National Natural Science Foundation of China (No. U21A20156), the Specific Research Project of Guangxi for Research Bases and Talents (No.AD22035185), the visiting scholars' fund at the WRHES (No.2021NSG02), the Belt and Road Special Foundation of the National Key Laboratory of Water Disaster Prevention (2022nkms06). Thanks are due to the reviewers and editors for their careful work and thoughtful suggestions which substantially improve the article, and we would also need to acknowledge the Water Source of the Middle-Route of the South-to-North Water Diversion Project of China Corporation Limited that supported the data collection.

References

Banerjee, A., Chakrabarty, M., Rakshit, N., Bhowmick, A.R. and Ray, S. (2019) Environmental factors as indicators of dissolved oxygen concentration and zooplankton abundance: Deep learning versus traditional regression approach. Ecological Indicators 100, 99-117.

Berger, J.O. (2013) Statistical decision theory and Bayesian analysis, Springer Science & Business Media.

Cai, C., Wang, J. and Li, Z. (2019) Assessment and modelling of uncertainty in precipitation forecasts from TIGGE using fuzzy probability and Bayesian theory. Journal of Hydrology 577.

Cai, K., Zhang, X., Zhang, M., Ge, Q., Li, S., Qiao, B. and Liu, Y. (2023) Improving air pollutant prediction in Henan Province, China, by enhancing the concentration prediction accuracy using autocorrelation errors and an Informer deep learning model. Sustainable Environment Research 33(1).

Cannas, B., Fanni, A., See, L. and Sias, G. (2006) Data preprocessing for river flow forecasting using neural networks: Wavelet transforms and data partitioning. Physics and Chemistry of the Earth 31(18), 1164-1171.

Challinor, A.J., Smith, M.S. and Thornton, P. (2013) Use of agro-climate ensembles for quantifying uncertainty and informing adaptation. Agricultural and Forest Meteorology 170, 2-7.

Colombo, D. and Maathuis, M.H. (2014) Order-Independent Constraint-Based Causal Structure Learning. Journal of Machine Learning Research 15, 3741-3782.

Dai, M., Huang, S., Huang, Q., Leng, G., Guo, Y., Wang, L., Fang, W., Li, P. and Zheng, X. (2020) Assessing agricultural drought risk and its dynamic evolution characteristics. Agricultural Water Management 231.

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

Du, Z., Qin, M., Zhang, F. and Liu, R. (2018) Multistep-ahead forecasting of chlorophyll a using a wavelet nonlinear autoregressive network. Knowledge-Based Systems 160, 61-70.

Glibert, P.M., Allen, J.I., Bouwman, A.F., Brown, C.W., Flynn, K.J., Lewitus, A.J. and Madden, C.J. (2010) Modeling of HABs and eutrophication Status, advances, challenges. Journal of Marine Systems 83(3-4), 262-275.

Gong, M., Zhao, Y., Sun, J., Han, C., Sun, G. and Yan, B. (2022) Load forecasting of district heating system based on Informer. Energy 253.

Größer, J. and Okhrin, O. (2021) Copulae: An overview and recent developments. WIREs Computational Statistics 14(3).

Hamed, M.G., Alligier, R. and Gianazza, D. (2016) High Confidence Intervals Applied to Aircraft Altitude Prediction. Ieee Transactions on Intelligent Transportation Systems 17(9), 2515-2527.

Harris, T.D. and Graham, J.L. (2017) Predicting cyanobacterial abundance, microcystin, and geosmin in a eutrophic drinking-water reservoir using a 14-year dataset. Lake and Reservoir Management 33(1), 32-48.

He, J., Chen, Y., Wu, J., Stow, D.A. and Christakos, G. (2020) Space-time chlorophyll-a retrieval in optically complex waters that accounts for remote sensing and modeling uncertainties and improves remote estimation accuracy. Water Research 171.

Huang, X. and Jiang, A. (2022) Wind Power Generation Forecast Based on Multi-Step Informer Network. Energies 15(18).

Kelly, K.S. and Krzysztofowicz, R. (2000) Precipitation uncertainty processor for probabilistic river stage forecasting. Water Resources Research 36(9), 2643-2653.

Kretschmer, M., Cohen, J., Matthias, V., Runge, J. and Coumou, D. (2018) The different stratospheric influence on cold-extremes in Eurasia and North America. Npj Climate and Atmospheric Science 1.

Krich, C., Mahecha, M.D., Migliavacca, M., De Kauwe, M.G., Griebel, A., Runge, J. and Miralles, D.G. (2022) Decoupling between ecosystem photosynthesis and transpiration: a last resort against overheating. Environmental Research Letters 17(4).

Krich, C., Runge, J., Miralles, D.G., Migliavacca, M., Perez-Priego, O., El-Madany, T., Carrara, A. and Mahecha, M.D. (2020) Estimating causal networks in biosphere-atmosphere interaction with the PCMCI approach. Biogeosciences 17(4), 1033-1061.

Krzysztofowicz, R. (1999) Bayesian theory of probabilistic forecasting via deterministic hydrologic model. Water Resources Research 35(9), 2739-2750.

Labat, D. (2005) Recent advances in wavelet analyses: Part I. A review of concepts. Journal of Hydrology 314(1-4), 275-288.

Li, S., Chen, X., Singh, V.P. and He, Y. (2018) Assumption-Simulation-Feedback-Adjustment (ASFA) Framework for Real-Time Correction of Water Resources Allocation: a Case Study of Longgang River Basin in Southern China. Water Resources Management 32(12), 3871-3886.

Liu, L., Peng, W., Wu, L., Liu, L. and Iop (2017) Water Quality Assessment of Danjiangkou Reservoir and its Tributaries in China, Thailand.

Liu, M., He, J., Huang, Y., Tang, T., Hu, J. and Xiao, X. (2022) Algal bloom forecasting with time-frequency analysis: A hybrid deep learning approach. Water Research 219.

Liu, Z., Cheng, L., Hao, Z., Li, J., Thorstensen, A. and Gao, H. (2018) A Framework for Exploring Joint Effects of Conditional Factors on Compound Floods. Water Resources Research 54(4), 2681-2696.

Masmoudi, S., Elghazel, H., Taieb, D., Yazar, O. and Kallel, A. (2020) A machine-learning framework for predicting multiple air pollutants' concentrations via multi-target regression and feature selection. Science of the Total Environment 715.

Nong, X., Lai, C., Chen, L., Shao, D., Zhang, C. and Liang, J. (2023) Prediction modelling framework comparative analysis of dissolved oxygen concentration variations using support vector regression coupled with multiple feature engineering and optimization methods: A case study in China. Ecological Indicators 146.

Nong, X., Shao, D., Zhong, H. and 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.

Nourani, V., Baghanam, A.H., Adamowski, J. and Kisi, O. (2014a) Applications of hybrid wavelet-Artificial Intelligence models in hydrology: A review. Journal of Hydrology 514, 358-377.

Nourani, V. and Behfar, N. (2021) Multi-station runoff-sediment modeling using seasonal LSTM models. Journal of Hydrology 601.

Nourani, V., Hosseini Baghanam, A., Adamowski, J. and Kisi, O. (2014b) Applications of hybrid wavelet–Artificial Intelligence models in hydrology: A review. Journal of Hydrology 514, 358-377.

Offie, I., Piadeh, F., Behzadian, K., Campos, L.C. and Yaman, R. (2023) Development of an artificial intelligence-based framework for biogas generation from a micro anaerobic digestion plant. Waste Management 158, 66-75.

Pearl, J. and Mackenzie, D. (2018) The book of why: the new science of cause and effect, Basic books.

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N. and Prabhat (2019) Deep learning and process understanding for data-driven Earth system science. Nature 566(7743), 195-204.

Runge, J., Bathiany, S., Bollt, E., Camps-Valls, G., Coumou, D., Deyle, E., Glymour, C., Kretschmer, M., Mahecha, M.D., Munoz-Mari, J., van Nes, E.H., Peters, J., Quax, R., Reichstein, M., Scheffer, M., Schoelkopf, B., Spirtes, P., Sugihara, G., Sun, J., Zhang, K. and Zscheischler, J. (2019a) Inferring causation from time series in Earth system sciences. Nature Communications 10.

Runge, J., Nowack, P., Kretschmer, M., Flaxman, S. and Sejdinovic, D. (2019b) Detecting and quantifying causal associations in large nonlinear time series datasets. Science Advances 5(11).

Runge, J., Petoukhov, V., Donges, J.F., Hlinka, J., Jajcay, N., Vejmelka, M., Hartman, D., Marwan, N., Paluš, M. and Kurths, J. (2015) Identifying causal gateways and mediators in complex spatio-temporal systems. Nature Communications 6(1), 8502.

Sahoo, B.B., Jha, R., Singh, A. and Kumar, D. (2020) Bivariate low flow return period analysis in the Mahanadi River basin, India using copula. International Journal of River Basin Management 18(1), 107-116.

Salimi, S. and Hammad, A. (2020) Sensitivity analysis of probabilistic occupancy prediction model using big data. Building and Environment 172.

Santy, S., Mujumdar, P. and Bala, G. (2020) Potential Impacts of Climate and Land Use Change on the Water Quality of Ganga River around the Industrialized Kanpur Region. Scientific Reports 10(1).

Sit, M.A., Koylu, C. and Demir, I. (2019) Identifying disaster-related tweets and their semantic, spatial and temporal context using deep learning, natural language processing and spatial analysis: a case study of Hurricane Irma. International Journal of Digital Earth 12(11), 1205-1229.

Sklar, M. (1959) Fonctions de répartition à n dimensions et leurs marges, pp. 229-231.

Song, C., Yao, L., Hua, C. and Ni, Q. (2021) A novel hybrid model for water quality prediction based on synchrosqueezed wavelet transform technique and improved long short-term memory. Journal of Hydrology 603.

Tibangayuka, N., Mulungu, D.M.M. and Izdori, F. (2022) Performance evaluation, sensitivity, and uncertainty analysis of HBV model in Wami Ruvu basin, Tanzania. Journal of Hydrology: Regional Studies 44, 101266.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L. and Polosukhin, I. (2017) Attention Is All You Need, Long Beach, CA.

Wang, Q., Yue, C., Li, X., Liao, P. and Li, X. (2023) Enhancing robustness of monthly streamflow forecasting model using embedded-feature selection algorithm based on improved gray wolf optimizer. Journal of Hydrology 617.

Wellen, C., Kamran-Disfani, A.-R. and Arhonditsis, G.B. (2015) Evaluation of the Current State of Distributed Watershed Nutrient Water Quality Modeling. Environmental Science & Technology 49(6), 3278-3290.

Xia, R., Wang, G., Zhang, Y., Yang, P., Yang, Z., Ding, S., Jia, X., Yang, C., Liu, C., Ma, S., Lin, J., Wang, X., Hou, X., Zhang, K., Gao, X., Duan, P. and Qian, C. (2020) River algal blooms are well predicted by antecedent environmental conditions. Water Research 185.

Xiao, X., He, J., Huang, H., Miller, T.R., Christakos, G., Reichwaldt, E.S., Ghadouani, A., Lin, S., Xu, X. and Shi, J. (2017) A novel single-parameter approach for forecasting algal blooms. Water Research 108, 222-231.

Yang, D. (2020) Reconciling solar forecasts: Probabilistic forecast reconciliation in a nonparametric framework. Solar Energy 210, 49-58.

Yao, S., Zhang, Y., Wang, P., Xu, Z., Wang, Y. and Zhang, Y. (2022) Long-Term Water Quality Prediction Using Integrated Water Quality Indices and Advanced Deep Learning Models: A Case Study of Chaohu Lake, China, 2019-2022. Applied Sciences-Basel 12(22).

Yu, R. and Zhang, C. (2021) Early warning of water quality degradation: A copula-based Bayesian network model for highly efficient water quality risk assessment. Journal of Environmental Management 292.

Zhang, C., Nong, X., Shao, D. and Chen, L. (2023) An integrated risk assessment framework using information theory-based coupling methods for basin-scale water quality management: A case study in the Danjiangkou Reservoir Basin, China. Science of the Total Environment 884.

Zhang, C., Nong, X., Shao, D., Zhong, H., Shang, Y. and Liang, J. (2021) Multivariate water environmental risk analysis in long-distance water supply project: A case study in China. Ecological Indicators 125.

Zheng, L., Wang, H., Liu, C., Zhang, S., Ding, A., Xie, E., Li, J. and Wang, S. (2021) Prediction of harmful algal blooms in large water bodies using the combined EFDC and LSTM models. Journal of Environmental Management 295.

Zhi, B., Wang, X. and Xu, F. (2022) Managing inventory financing in a volatile market: A novel data-driven copula model. Transportation Research Part E-Logistics and Transportation Review 165.

Zhong, S., Zhang, K., Bagheri, M., Burken, J.G., Gu, A., Li, B., Ma, X., Marrone, B.L., Ren, Z.J., Schrier, J., Shi, W., Tan, H., Wang, T., Wang, X., Wong, B.M., Xiao, X., Yu, X., Zhu, J.-J. and Zhang, H. (2021) Machine Learning: New Ideas and Tools in Environmental Science and Engineering. Environmental Science & Technology 55(19), 12741-12754.

Zhou, Y. (2020) Real-time probabilistic forecasting of river water quality under data missing situation: Deep learning plus post-processing techniques. Journal of Hydrology 589.

Zou, L., Zha, Y., Diao, Y., Tang, C., Gu, W. and Shao, D. (2022) Coupling the Causal

Inference and Informer Networks for Short-term Forecasting in Irrigation Water Usage. Water Resources Management 37(1), 427-449.

Figure Captions

Fig. 1. The framework of the proposed coupling predictive methods in this study.

Fig. 2. The location of the Danjiangkou Reservoir and three automatic water quality monitoring stations.

Fig. 3. Causal networks of all parameters in the three stations (Note: Based on the PCMCI method, the strength of causality is given by the link colour and the time lags are shown in the centre of each arrow).

Fig. 4. Comparisons of the predictive model performances with and without PCMCI in different stations.

Fig. 5. Observation and prediction series of TN using different models in three stations for one step ahead (Note: the inner plots represent the relative error (%)).

Fig. 6. Observations, predictions of the WInformer, and the 95% confidence interval for the TN of different stations in the evaluation stages (TC, QS, and MD are the names of stations; CR: Coverage Rate; ARIL: Average Relative Interval Length).

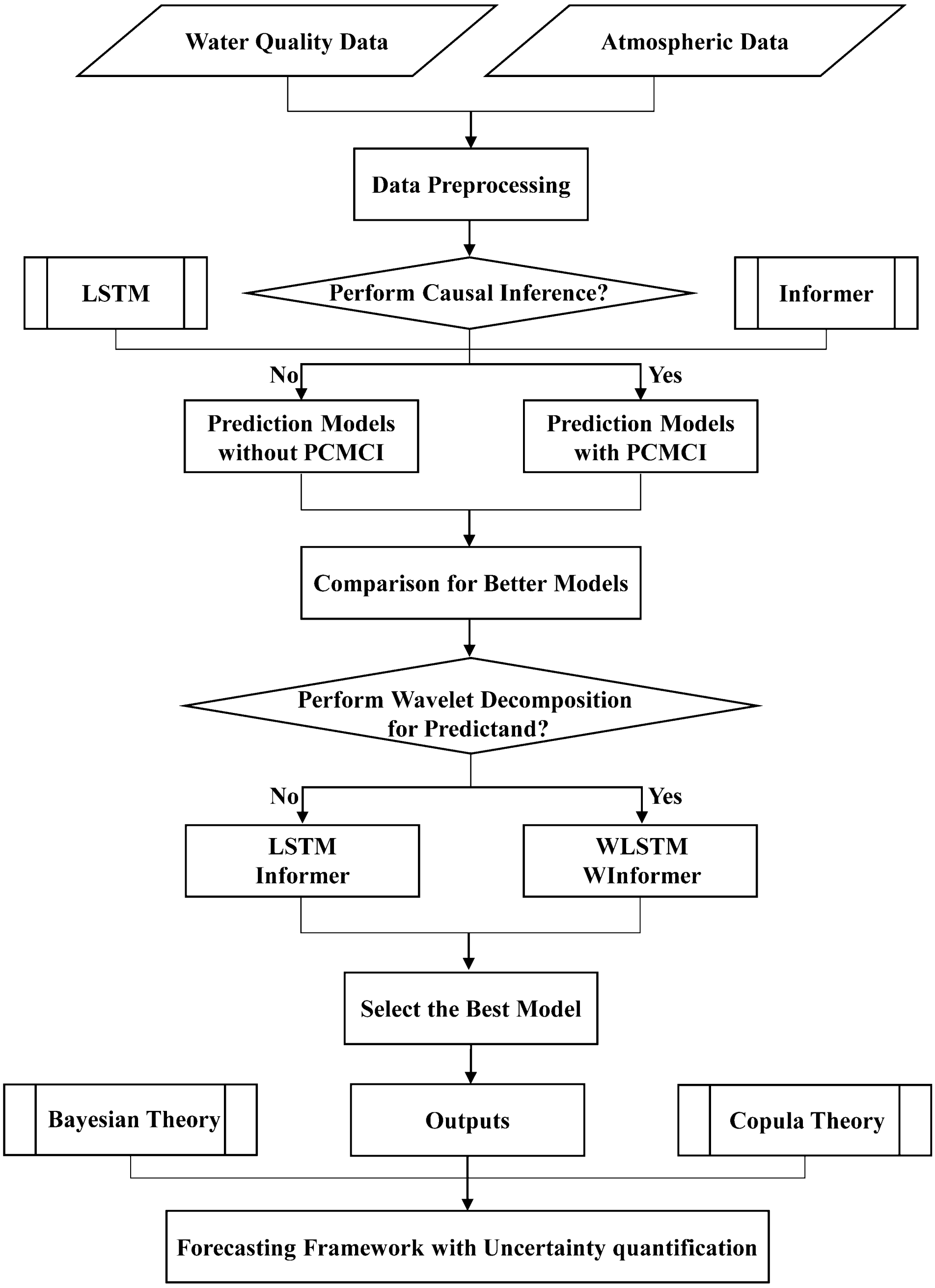


Fig. 1. The framework of the proposed coupling predictive methods in this study.

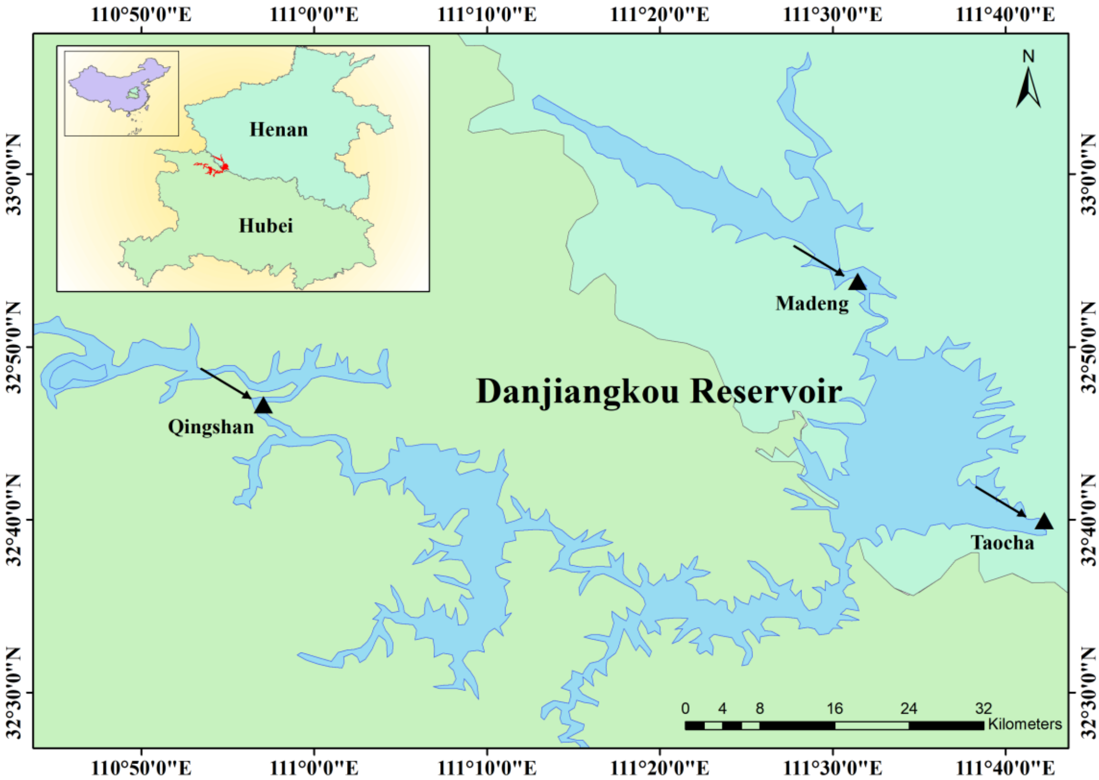


Fig. 2. The location of the Danjiangkou Reservoir and three automatic water quality monitoring stations (i.e., Taocha, Qingshan, and Madeng).

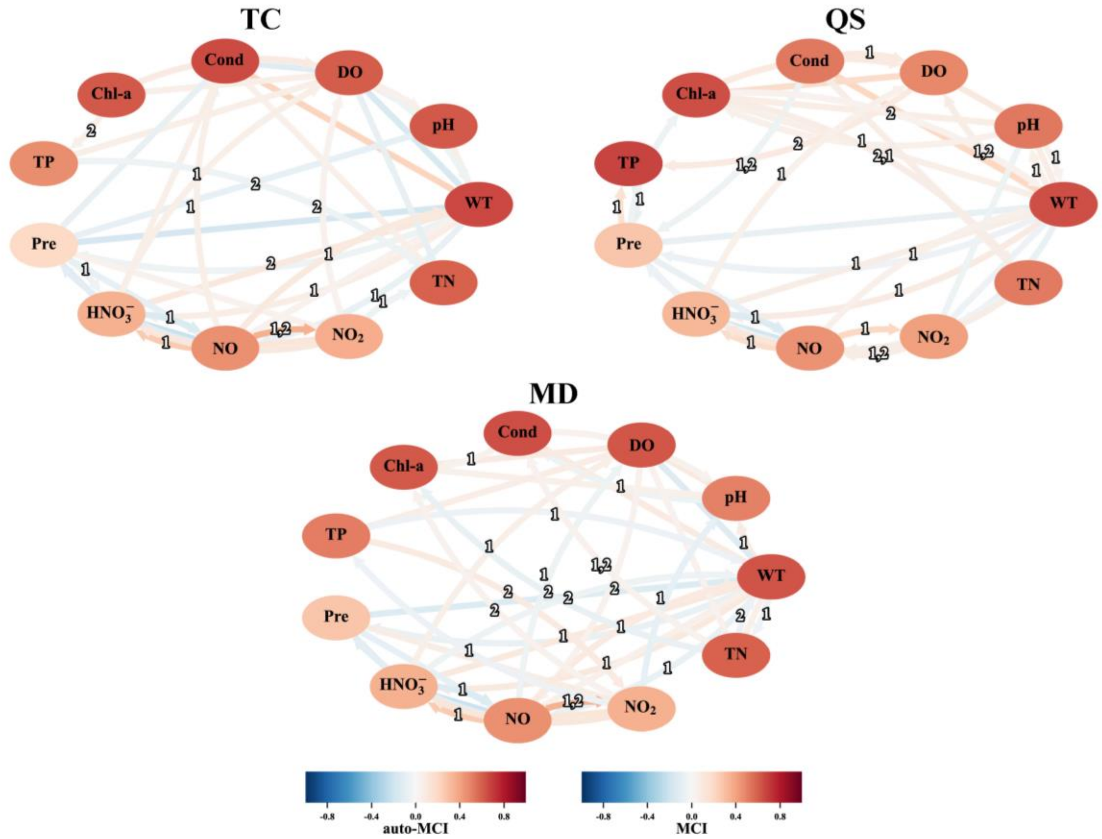


Fig. 3. Causal networks of all parameters in the three stations (Note: Based on the PCMCI method, the strength of causality is given by the link colour and the time lags are shown in the centre of each arrow).

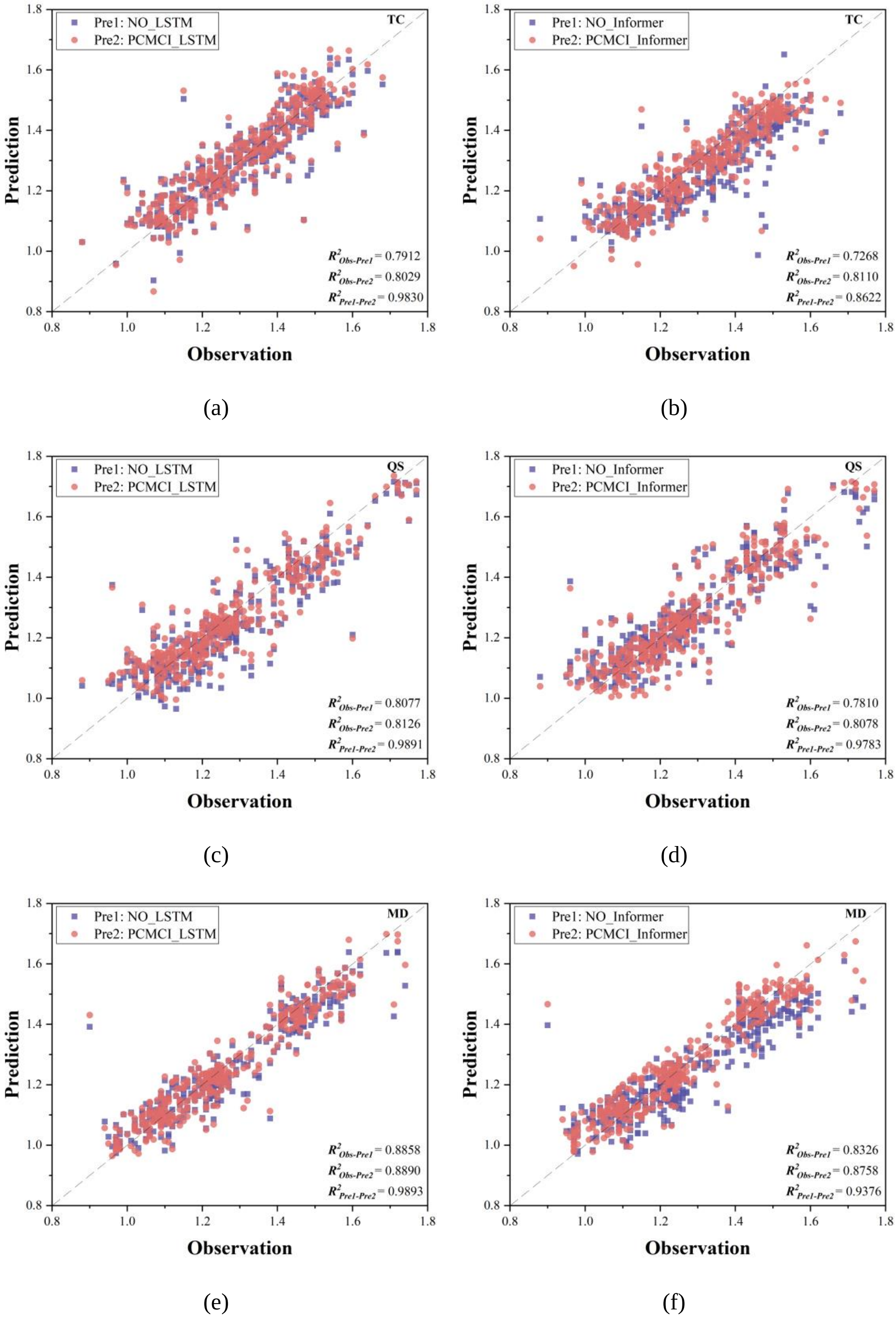
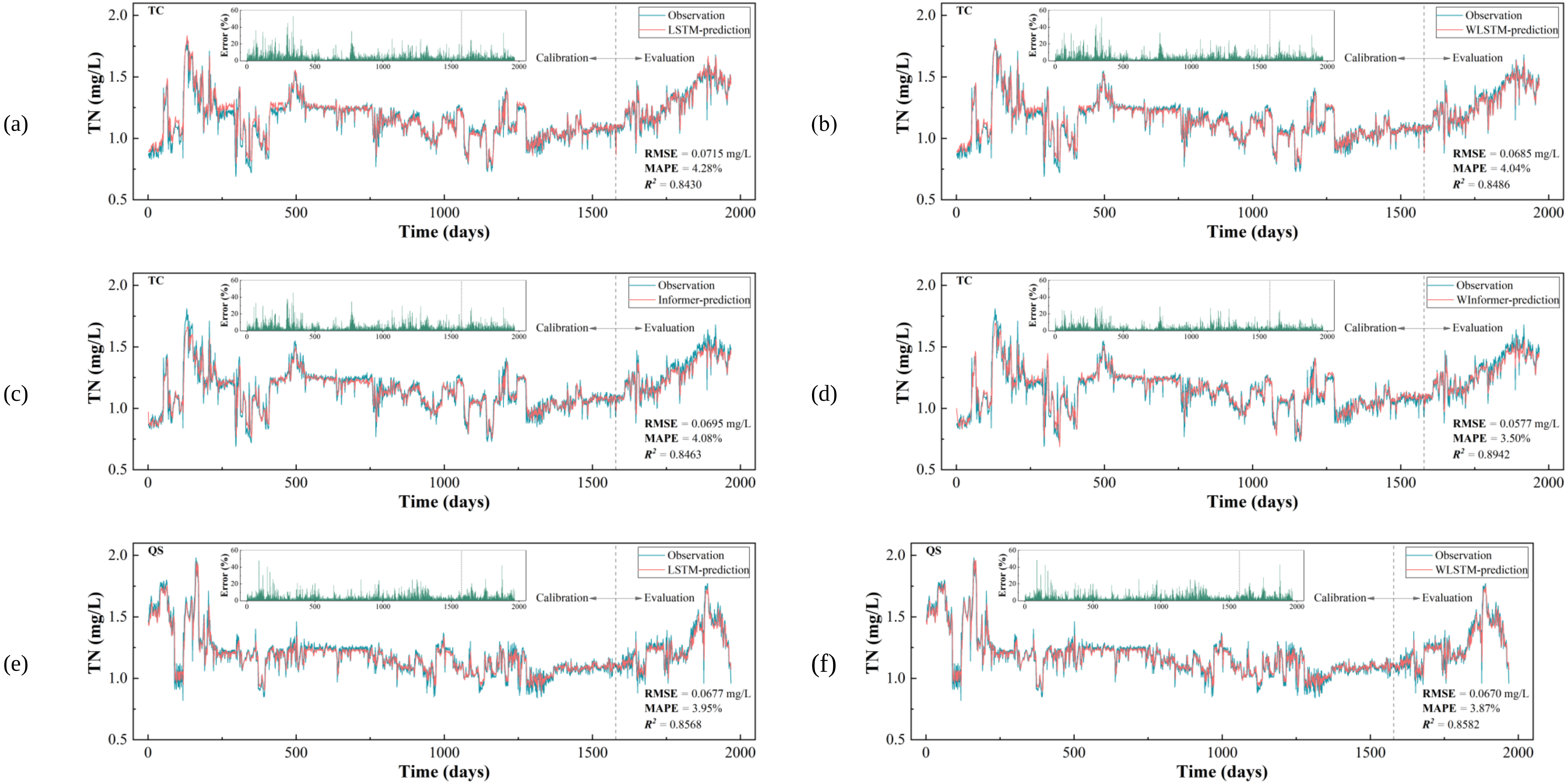


Fig. 4. Comparisons of the predictive model performances with and without PCMCI in different stations.


(a)
TC
Observation
LSTM-prediction
Calibration
Evaluation
RMSE = 0.0715 mg/L
MAPE = 4.28%
R² = 0.8430
TN (mg/L)
Time (days)
Error (%)
(b)
TC
Observation
WLSTM-prediction
Calibration
Evaluation
RMSE = 0.0685 mg/L
MAPE = 4.04%
R² = 0.8486
TN (mg/L)
Time (days)
Error (%)
(c)
TC
Observation
Informer-prediction
Calibration
Evaluation
RMSE = 0.0695 mg/L
MAPE = 4.08%
R² = 0.8463
TN (mg/L)
Time (days)
Error (%)
(d)
TC
Observation
WInformer-prediction
Calibration
Evaluation
RMSE = 0.0577 mg/L
MAPE = 3.50%
R² = 0.8942
TN (mg/L)
Time (days)
Error (%)
(e)
QS
Observation
LSTM-prediction
Calibration
Evaluation
RMSE = 0.0677 mg/L
MAPE = 3.95%
R² = 0.8568
TN (mg/L)
Time (days)
Error (%)
(f)
QS
Observation
WLSTM-prediction
Calibration
Evaluation
RMSE = 0.0670 mg/L
MAPE = 3.87%
R² = 0.8582
TN (mg/L)
Time (days)
Error (%)

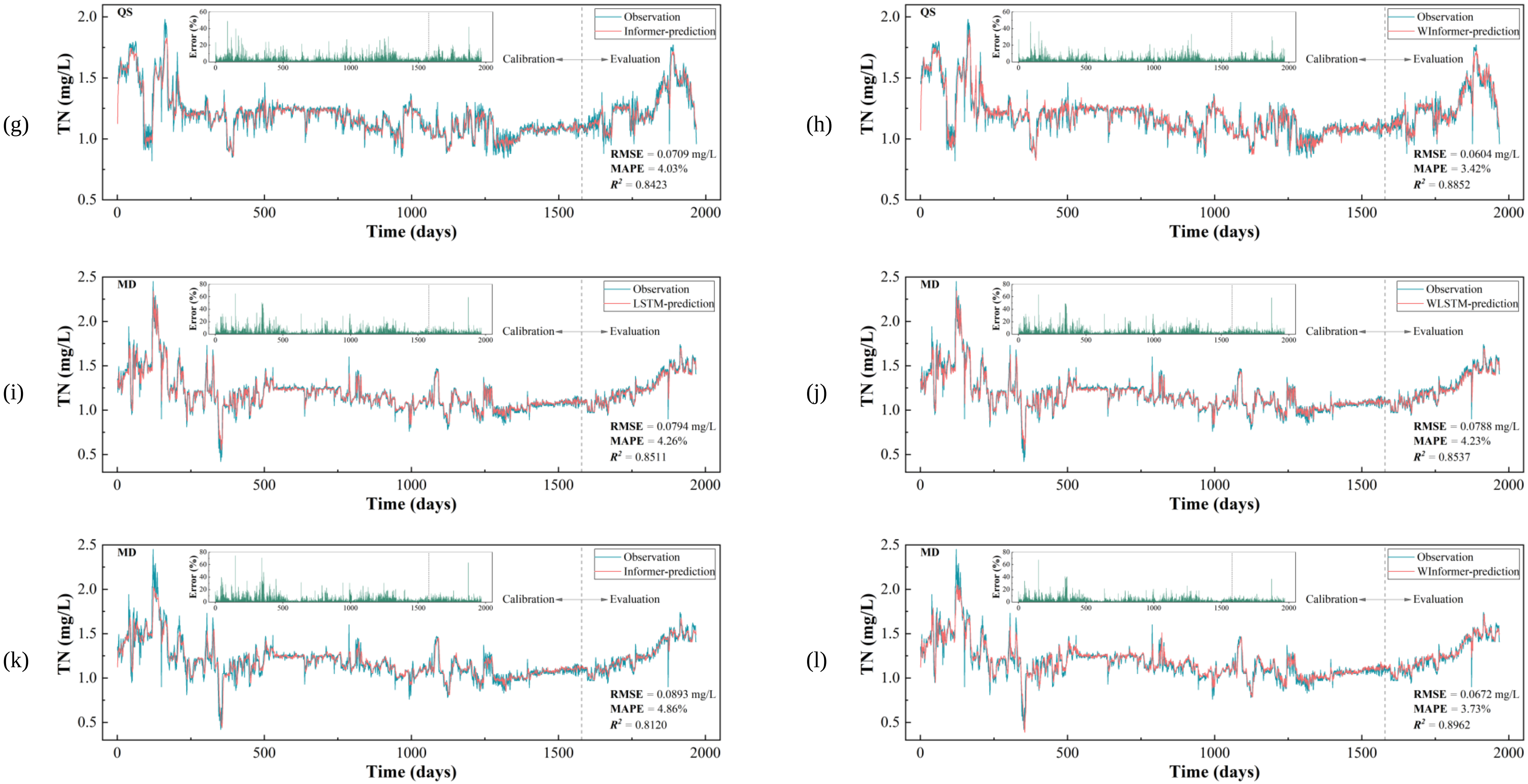


Fig. 5. Observation and prediction series of TN using different models in three stations for one step ahead (Note: the inner plots represent the relative error (%)).

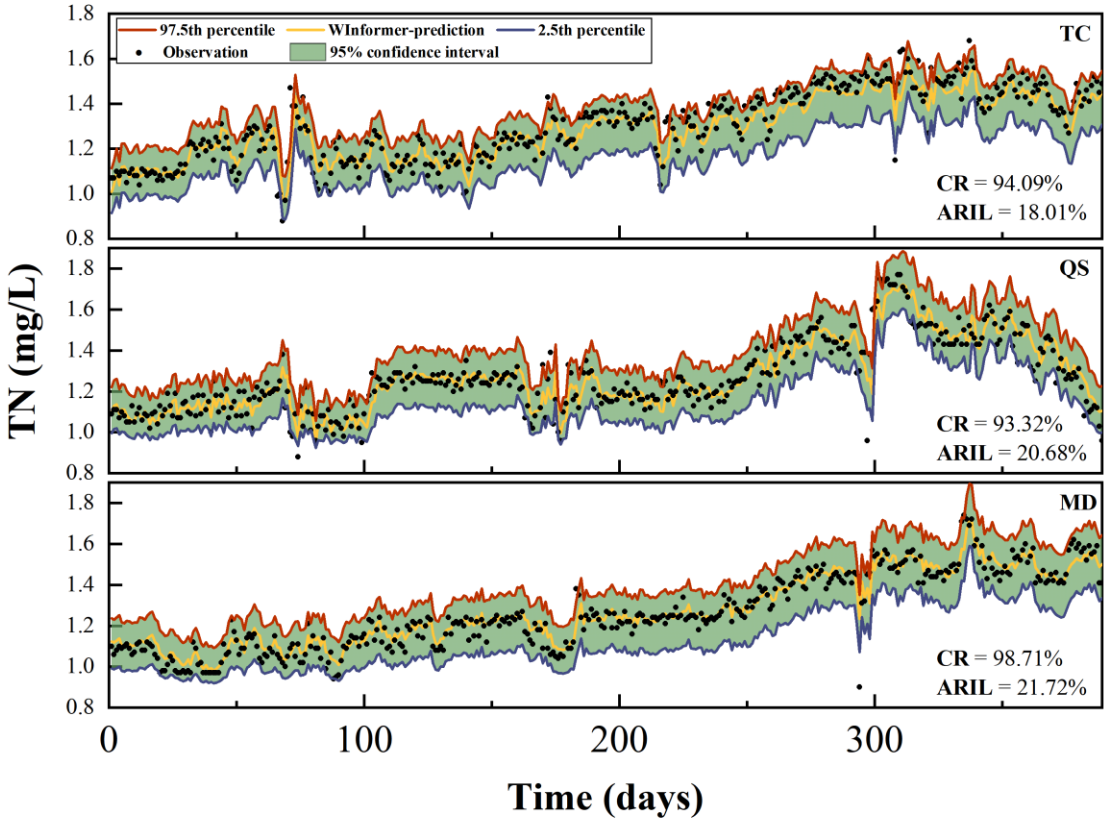


Fig. 6. Observations, predictions of the WInformer, and the 95% confidence interval for the TN of different stations in the evaluation stages (TC, QS, and MD are the names of stations; CR: Coverage Rate; ARIL: Average Relative Interval Length).

Tables

Table 1

Summary of all indicators in the three automatic monitoring stations from 2017 to 2022 (Avg.: Average; S.D.: Standard deviation).

Table 2

The selected features for different stations.

Table 3

Comparisons of the prediction models with and without causal inference in the evaluation stages.

Table 4

The forecasting performance of WInformer comparing to the other three models.

Table 1

Summary of all indicators in the three automatic monitoring stations from 2017 to 2022 (Avg.: Average; S.D.: Standard deviation).

Parameters	Taocha (TC)			Qingshan (QS)			Madeng (MD)		
	Avg. ± S.D.	Max	Min	Avg. ± S.D.	Max	Min	Avg. ± S.D.	Max	Min
WT (°C)	18.4 ± 6.9	32.7	5.9	18.0 ± 6.6	33	6.3	18.7 ± 6.6	32.5	4.9
pH	8.08 ± 0.35	9.10	6.50	8.15 ± 0.33	9.30	6.50	8.12 ± 0.32	9.10	6.00
DO (mg/L)	9.70 ± 1.30	12.70	6.10	9.90 ± 1.30	16.20	7.10	9.60 ± 1.30	16.20	6.59
Cond (μS/cm)	272.6 ± 46.6	550.8	175.0	256.7 ± 28.4	346.4	142.9	284.8 ± 60.5	1071.0	109.4
Chl-a (μg/L)	2.36 ± 3.19	98.50	0.20	2.41 ± 1.88	19.10	0.27	2.63 ± 1.89	16.40	0.20
TP (mg/L)	0.013 ± 0.004	0.041	0.002	0.017 ± 0.010	0.269	0.004	0.014 ± 0.005	0.051	0.001
Pre (mm)	6.9 ± 28.1	561.5	0	4.0 ± 20.5	361.0	0	7.3 ± 28.4	346.7	0
HNO_3 (μg/m^3)	6.76 ± 4.68	43.02	0.02	4.96 ± 4.01	39.16	0.04	6.76 ± 4.67	43.02	0.02
NO (μg/m^3)	30.04 ± 24.37	129.43	0.27	11.30 ± 11.30	59.28	0.07	30.03 ± 24.37	129.43	0.27
NO_2 (μg/m^3)	41.80 ± 13.06	173.16	13.95	27.93 ± 9.58	85.49	7.91	41.80 ± 13.06	173.16	13.95
TN (mg/L)	1.17 ± 0.18	1.81	0.69	1.20 ± 0.18	1.98	0.82	1.19 ± 0.21	2.45	0.42

Table 2

The selected features for different stations.

Station	Selected features
Taocha (TC)	TN(t-1), TN(t-2), DO(t), Cond(t), TP(t-2), NO_2(t-1)
Qingshan (QS)	TN(t-1), TN(t-2), Cond(t), Chl-a(t-2), Chl-a(t-1), NO_2(t)
Madeng (MD)	TN(t-1), TN(t-2), WT(t), DO(t), WT(t-1)

Table 3

Comparisons of the prediction models with and without causal inference in the evaluation stages.

Station	Model	RMSE	MAPE	R^2
TC	NO_LSTM	0.0716	4.06%	0.7912
	LSTM	0.0711	4.06%	0.8029
	NO_Informer	0.0924	5.31%	0.7268
	Informer	0.0713	4.12%	0.8110
QS	NO_LSTM	0.0800	4.71%	0.8077
	LSTM	0.0759	4.46%	0.8126
	NO_Informer	0.0811	4.71%	0.7810
	Informer	0.0764	4.47%	0.8078
MD	NO_LSTM	0.0642	3.71%	0.8858
	LSTM	0.0618	3.44%	0.8890
	NO_Informer	0.0824	5.01%	0.8326
	Informer	0.0649	3.77%	0.8758

Table 4

The forecasting performance of WInformer comparing to the other three models.

Calibration						Evaluation			
RMSE (mg/L)		**MAPE**		**R^2**		**RMSE (mg/L)**		**MAPE**	
0.0576		3.55%		0.8783		0.0581		3.30 %	
0.0690	(+Δ16.53%)	4.08%	(+Δ12.81%)	0.8234	(+Δ6.66%)	0.0713	(+Δ18.49%)	4.12%	(+Δ19.85
0.0684	(+Δ15.70%)	4.07%	(+Δ12.79%)	0.8274	(+Δ6.15%)	0.0692	(+Δ16.08%)	3.90%	(+Δ15.37
0.0715	(+Δ19.47%)	4.33%	(+Δ18.00%)	0.8208	(+Δ7.00%)	0.0711	(+Δ18.29%)	4.06%	(+Δ18.65
0.0594		3.35%		0.8848		0.0641		3.68%	
0.0694	(+Δ14.47%)	3.92%	(+Δ14.56%)	0.8433	(+Δ4.91%)	0.0764	(+Δ16.11%)	4.47%	(+Δ17.55
0.0648	(+Δ8.38%)	3.73%	(+Δ10.30%)	0.8627	(+Δ2.56%)	0.0753	(+Δ14.83%)	4.42%	(+Δ16.71
0.0656	(+Δ9.39%)	3.82%	(+Δ12.32%)	0.8607	(+Δ2.80%)	0.0759	(+Δ15.50%)	4.46%	(+Δ17.43
0.0712		3.95%		0.8842		0.0472		2.85%	
0.0943	(+Δ24.51%)	5.12%	(+Δ22.85%)	0.7928	(+Δ11.54%)	0.0649	(+Δ27.38%)	3.77%	(+Δ24.39
0.0824	(+Δ13.56%)	4.41%	(+Δ10.32%)	0.8419	(+Δ5.02%)	0.0624	(+Δ24.46%)	3.48%	(+Δ18.21
0.0832	(+Δ14.40%)	4.46%	(+Δ11.44%)	0.8387	(+Δ5.43%)	0.0618	(+Δ23.70%)	3.44%	(+Δ17.17

Note: The values in parentheses represent the improvement rates of the WInformer model over the other three models in terms of the corresponding metrics.

Supplementary Materials

A new framework for water quality forecasting coupling causal inference, time-frequency analysis, and uncertainty quantification

Chi Zhang [a], Xizhi Nong [a, b, c, *], Kourosh Behzadian [d, e], Luiza C. Campos [d], Lihua Chen [b], Dongguo Shao [a, *]

[a] State Key Laboratory of Water Resources Engineering and Management, Wuhan University, Wuhan 430072, China

[b] College of Civil Engineering and Architecture, Guangxi University, Nanning 530004, China

[c] The National Key Laboratory of Water Disaster Prevention, Nanjing Hydraulic Research Institute, Nanjing 210029, China

[d] Centre for Urban Sustainability and Resilience, Department of Civil, Environmental and Geomatic Engineering, University College London, London WC1E 6BT, United Kingdom

[e] School of Computing and Engineering, University of West London, London W5 5RF, United Kindom

* Corresponding author: Dr. Xizhi Nong; Professor Dongguo Shao

E-mail addresses: nongxizhi@gxu.edu.cn, dongguoshao@163.com

Table S1

Hyperparameter selections of the LSTM for different data series in the three stations in this study.

Station	Series	num_layers	num_neurons	Epoch	batch_size	dropout_rate
TC	S_NO	1	128	80	64	0.1
	S	1	32	60	32	0.2
	A_3	1	32	40	32	0.2
	D_1	1	32	80	16	0.5
	D_2	1	256	100	16	0.4
	D_3	1	256	80	64	0.4
QS	S_NO	1	256	80	64	0.3
	S	1	32	80	64	0.3
	A_3	1	32	80	64	0.1
	D_1	1	32	100	16	0.4
	D_2	1	256	80	128	0.1
	D_3	1	256	100	64	0.1
MD	S_NO	1	64	40	64	0.4
	S	1	256	80	128	0.4
	A_3	1	128	100	128	0.1
	D_1	1	128	100	128	0.2
	D_2	1	128	60	32	0.5
	D_3	1	128	60	64	0.5

Table S2

Hyperparameter selections of the Informer for different data series in the three stations in this study.

Station	Series	n_heads	e_layers	d_layers	seq_len	label_len	pred_len	epoch	batch_size	dropout_rate
TC	S_NO	8	3	1	30	14	1	20	16	0.05
	S	8	2	1	14	7	1	20	16	0.05
	A_3	4	3	1	30	7	1	20	16	0.05
	D_1	4	3	1	7	3	1	20	16	0.05
	D_2	4	3	1	14	3	1	20	16	0.05
	D_3	4	3	1	7	3	1	20	16	0.05
QS	S_NO	8	3	1	14	3	1	20	16	0.05
	S	8	3	1	7	3	1	20	16	0.05
	A_3	8	3	1	7	3	1	20	32	0.05
	D_1	8	3	1	14	7	1	20	16	0.05
	D_2	8	2	1	7	3	1	20	16	0.05
	D_3	4	2	1	14	3	1	20	16	0.05
MD	S_NO	8	3	1	14	3	1	20	16	0.05
	S	8	2	1	14	3	1	20	16	0.05
	A_3	8	2	1	14	3	1	20	16	0.05
	D_1	8	2	1	14	3	1	20	16	0.05
	D_2	4	2	1	14	3	1	20	16	0.05
	D_3	4	2	1	7	3	1	20	16	0.05

Table S3

The selected bivariate copula functions and their mathematical expressions in this study.

Copula functions	Abbreviation	Mathematical expressions	Parameters
Gaussian	N	$\int_{-\infty}^{\emptyset^{-1}(u)}\int_{-\infty}^{\emptyset^{-1}(v)}\frac{1}{2\pi\sqrt{1-\theta^2}}exp\,(\frac{2\theta xy-x^2-y^2}{2(1-\theta^2)})dxdy$	$\theta\in[-1,\ 1]$
Student-t	t	$\int_{-\infty}^{t_{\theta_2}^{-1}(u)} z\int_{-\infty}^{t_{\theta_2}^{-1}(v)}\frac{\Gamma(\frac{\theta_2+2}{2})}{\Gamma(\frac{\theta_2}{2})\pi\theta_2\sqrt{1-\theta_1{}^2}}\,(1+\frac{x^2+y^2-2\theta_1 xy}{\theta_2})^{\frac{\theta_2+2}{2}}dxdy$	$\theta_1\in[-1,\ 1]\&\theta_2\in(0,\ \infty)$
Gumbel	G	$exp\left\{-\left[(-\ln u)^{\theta}+(-\ln v)^{\theta}\right]^{1/\theta}\right\}$	$\theta\in[1,\ \infty)$
Clayton	C	$\left(u^{-\theta}+v^{-\theta}-1\right)^{-1/\theta}$	$\theta\in(0,\ \infty)$
Frank	F	$-\frac{1}{\theta}ln\left[1+\frac{(e^{-\theta u}-1)(e^{-\theta v}-1)}{e^{-\theta}-1}\right]$	$\theta\in\mathrm{R}\backslash\{0\}$

Table S4

Numerical results for PCMCI with parents, corresponding lags, and dependency coefficients (link strength) in this study.

TC					QS					MD				
Variable	**Parents**	**Lag**	**p-value**	**Dep. Coef.**	**Variable**	**Parents**	**Lag**	**p-value**	**Dep. Coef.**	**Variable**	**Parents**	**Lag**	**p-value**	**Dep. Coef.**
TN	TN	-1	0.0000	0.594	TN	TN	-1	0.0000	0.522	TN	TN	-1	0.0000	0.591
	TN	-2	0.0001	0.089		TN	-2	0.0000	0.194		TN	-2	0.0000	0.100
	DO	0	0.0016	-0.071		Cond	0	0.0033	0.066		WT	0	0.0243	0.051
	Cond	0	0.0233	0.051		Chl-a	-2	0.0045	0.064		DO	0	0.0379	0.047
	TP	-2	0.0264	-0.050		Chl-a	-1	0.0110	-0.057		WT	-1	0.0494	-0.044
	NO_2	-1	0.0378	-0.047		NO_2	0	0.0399	-0.046	WT	WT	-1	0.0000	0.632
WT	WT	-1	0.0000	0.661	WT	WT	-1	0.0000	0.644		Pre	0	0.0000	-0.133
	Cond	0	0.0000	0.223		Cond	0	0.0000	0.207		DO	0	0.0000	-0.128
	Pre	0	0.0000	-0.177		pH	0	0.0007	0.077		HNO_3^-	0	0.0000	0.098
	DO	0	0.0000	-0.158		WT	-2	0.0012	0.073		Cond	0	0.0000	0.093
	HNO_3^-	0	0.0000	0.103		Pre	0	0.0015	-0.072		WT	-2	0.0001	0.091
	NO	-1	0.0000	0.097		HNO_3^-	0	0.0031	0.067		NO	-1	0.0001	0.087
	pH	0	0.0015	0.072		Chl-a	0	0.0126	0.056		TP	-1	0.0081	-0.060
	DO	-1	0.0187	-0.053		NO_2	0	0.0154	-0.055		TN	-2	0.0090	-0.059
	NO	0	0.0383	0.047		NO	-1	0.0191	0.053		TN	0	0.0243	0.051
pH	pH	-1	0.0000	0.616		pH	-1	0.0381	0.047		NO	0	0.0283	0.050
	DO	0	0.0001	0.089		Chl-a	-2	0.0409	0.046		HNO_3^-	-2	0.0364	-0.047
	Pre	0	0.0002	-0.083	pH	pH	-1	0.0000	0.529	pH	pH	-1	0.0000	0.504
	WT	0	0.0015	0.072		pH	-2	0.0000	0.133		pH	-2	0.0000	0.116
	Cond	-1	0.0371	0.047		DO	0	0.0000	0.096		DO	0	0.0000	0.098

DO	DO	-1	0.0000	0.602
	WT	0	0.0000	-0.158
	Cond	0	0.0000	-0.124
	pH	0	0.0001	0.089
	Chl-a	-1	0.0002	0.085
	DO	-2	0.0012	0.073
	TN	0	0.0016	-0.071
	TP	0	0.0068	0.061
	Chl-a	0	0.0091	0.059
	HNO_3^-	-1	0.0109	0.057
	NO_2	-2	0.0417	0.046
Cond	Cond	-1	0.0000	0.656
	WT	0	0.0000	0.223
	DO	0	0.0000	-0.124
	Pre	0	0.0007	-0.077
	NO	-1	0.0080	0.060
	HNO_3^-	0	0.0115	0.057
	TN	0	0.0233	0.051
Chl-a	Chl-a	-1	0.0000	0.614
	Chl-a	-2	0.0000	-0.352
	DO	0	0.0091	0.059
TP	TP	-1	0.0000	0.466
	TP	-2	0.0000	0.187
	DO	0	0.0068	0.061
	Chl-a	-2	0.0229	0.051
Pre	Pre	-1	0.0000	0.202

	WT	0	0.0007	0.077
	Chl-a	0	0.0016	0.071
	NO_2	0	0.0230	-0.051
	WT	-1	0.0336	0.048
DO	DO	-1	0.0000	0.484
	Chl-a	0	0.0000	0.183
	DO	-2	0.0000	0.173
	pH	0	0.0000	0.096
	Chl-a	-1	0.0000	0.093
	Cond	-1	0.0001	0.086
	WT	-1	0.0052	0.063
	HNO_3^-	-1	0.0328	0.048
	WT	-2	0.0381	0.047
Cond	Cond	-1	0.0000	0.528
	WT	0	0.0000	0.207
	Cond	-2	0.0000	0.122
	TN	0	0.0033	0.066
Chl-a	Chl-a	-1	0.0000	0.647
	DO	0	0.0000	0.183
	pH	0	0.0016	0.071
	WT	0	0.0126	0.056
	Pre	-1	0.0227	-0.051
	pH	-1	0.0314	0.049
TP	TP	-1	0.0000	0.673
	TP	-2	0.0000	-0.280
	Pre	-1	0.0000	0.141

	WT	-1	0.0002	0.084
	NO_2	-1	0.0003	-0.082
	Chl-a	0	0.0278	0.050
	Cond	-1	0.0386	0.047
DO	DO	-1	0.0000	0.618
	WT	0	0.0000	-0.128
	pH	0	0.0000	0.098
	Chl-a	0	0.0002	0.084
	TP	0	0.0008	0.075
	HNO_3^-	-1	0.0065	0.061
	NO	-1	0.0187	-0.053
	TN	0	0.0379	0.047
Cond	Cond	-1	0.0000	0.641
	Cond	-2	0.0000	-0.141
	WT	0	0.0000	0.093
	pH	-1	0.0186	-0.053
	TN	-1	0.0381	0.047
	TN	-2	0.0414	-0.046
Chl-a	Chl-a	-1	0.0000	0.613
	DO	0	0.0002	0.084
	Chl-a	-2	0.0015	-0.072
	TN	-2	0.0031	-0.067
	pH	0	0.0278	0.050
	Cond	-1	0.0408	0.046
	NO_2	-2	0.0415	0.046
TP	TP	-1	0.0000	0.509

	WT	0	0.0000	-0.177		DO	-2	0.0001	0.092		TP	-2	0.0000	0.129
	HNO_3^-	0	0.0000	-0.119	Pre	Pre	-1	0.0000	0.275		DO	0	0.0008	0.075
	NO	-1	0.0000	-0.105		HNO_3^-	0	0.0002	-0.083		NO_2	-1	0.0152	-0.055
	pH	0	0.0002	-0.083		TP	-1	0.0006	-0.077	Pre	Pre	-1	0.0000	0.278
	NO	0	0.0005	-0.079		WT	0	0.0015	-0.072		WT	0	0.0000	-0.133
	Cond	0	0.0007	-0.077		NO	0	0.0042	-0.065		HNO_3^-	0	0.0000	-0.119
	WT	-2	0.0183	-0.053		NO	-1	0.0063	-0.062		NO	-1	0.0001	-0.101
	NO_2	0	0.0459	0.045		WT	-1	0.0093	-0.059		NO	0	0.0002	-0.085
HNO_3^-	HNO_3^-	-1	0.0000	0.355		Cond	-1	0.0238	-0.051		NO_2	0	0.0039	0.065
	NO	-1	0.0000	0.262		Cond	-2	0.0434	0.046		WT	-1	0.0055	-0.063
	NO	0	0.0000	-0.220	HNO_3^-	HNO_3^-	-1	0.0000	0.334		Pre	-2	0.0336	0.048
	Pre	0	0.0000	-0.119		NO	0	0.0000	-0.187	HNO_3^-	HNO_3^-	-1	0.0000	0.353
	NO_2	-1	0.0000	0.109		NO	-1	0.0000	0.182		NO	-1	0.0000	0.262
	WT	0	0.0000	0.103		NO_2	-1	0.0000	0.117		NO	0	0.0000	-0.214
	WT	-1	0.0007	0.076		WT	-1	0.0002	0.085		Pre	0	0.0000	-0.119
	Cond	0	0.0115	0.057		Pre	0	0.0002	-0.083		NO_2	-1	0.0000	0.115
	Pre	-1	0.0480	-0.045		WT	0	0.0031	0.067		WT	-1	0.0000	0.099
NO	NO	-1	0.0000	0.459	NO	NO	-1	0.0000	0.445		WT	0	0.0000	0.098
	HNO_3^-	0	0.0000	-0.220		HNO_3^-	0	0.0000	-0.187	NO	NO	-1	0.0000	0.460
	HNO_3^-	-1	0.0000	-0.142		HNO_3^-	-1	0.0000	-0.130		HNO_3^-	0	0.0000	-0.214
	NO_2	0	0.0000	0.124		NO	-2	0.0000	-0.095		HNO_3^-	-1	0.0000	-0.135
	Pre	0	0.0005	-0.079		TN	-1	0.0041	0.065		NO_2	0	0.0000	0.120
	WT	-1	0.0042	-0.065		Pre	0	0.0042	-0.065		Pre	0	0.0002	-0.085
	NO	-2	0.0162	-0.054		NO_2	-1	0.0055	0.063		WT	-1	0.0046	-0.064
	WT	0	0.0383	0.047		NO_2	-2	0.0286	0.050		NO	-2	0.0103	-0.058
NO_2	NO_2	-1	0.0000	0.365	NO_2	NO_2	-1	0.0000	0.409		WT	0	0.0283	0.050

NO	-1	0.0000	0.353	NO	-1	0.0000	0.206		DO	-2	0.0446	0.045
NO	0	0.0000	0.124	WT	-1	0.0036	-0.066	NO_2	NO_2	-1	0.0000	0.357
NO_2	-2	0.0002	0.083	WT	0	0.0154	-0.055		NO	-1	0.0000	0.356
NO	-2	0.0202	-0.053	pH	0	0.0230	-0.051		NO	0	0.0000	0.120
Pre	0	0.0459	0.045	TN	0	0.0399	-0.046		TP	-2	0.0001	0.087
									NO_2	-2	0.0013	0.073
									Pre	0	0.0039	0.065
									NO	-2	0.0192	-0.053

Table S5

Statistic characteristics of prediction errors (%) for different models in this study (Avg.: average, S.D.: Standard deviation).

Station	Model	Calibration			Evaluation		
		Avg.	S.D.	Max	Avg.	S.D.	Max
TC	LSTM	4.33	5.15	53.31	4.06	3.98	33.16
	Informer	4.08	4.90	45.26	4.12	3.72	27.81
	WLSTM	4.07	4.94	51.89	3.90	3.80	30.51
	Winformer	3.55	4.02	29.23	3.30	3.05	24.57
QS	LSTM	3.82	4.27	47.62	4.46	4.38	42.37
	Informer	3.92	4.56	48.99	4.47	4.44	42.06
	WLSTM	3.73	4.30	48.39	4.42	4.45	43.03
	Winformer	3.35	3.92	48.38	3.68	3.60	30.43
MD	LSTM	4.46	5.83	65.13	3.44	4.01	58.99
	Informer	5.12	6.42	74.60	3.77	4.22	62.99
	WLSTM	4.41	5.74	63.54	3.48	4.00	58.35
	Winformer	3.95	4.85	67.80	2.85	2.84	37.01

Table S6

Fitting results of the marginal distribution of TN for different data sets in this study.

Station	Set	$\bar{x}$	Cv	Cs	α	$1/\beta$	a_0
TC	Cali-Obs	1.13	0.14	0.35	32.65	0.028	0.226
	Cali-Pre	1.14	0.13	0.27	54.87	0.020	0.042
QS	Cali-Obs	1.18	0.15	1.3	2.37	0.115	0.908
	Cali-Pre	1.18	0.14	1.2	2.78	0.099	0.905
MD	Cali-Obs	1.17	0.17	1.4	2.04	0.139	0.886
	Cali-Pre	1.18	0.16	1.1	3.31	0.104	0.837

Note: $\bar{x}$: Mean; Cs: Coefficient of Skewness; Cv: Coefficient of Variation.The parameters of Pearson III distribution are α, β and a_0, respectively.

$$\alpha=\frac{4}{C_s^2} \quad \beta=\frac{2}{\bar{x}C_sC_v} \quad a_0=\bar{x}-\frac{2C_v\bar{x}}{C_s}$$

Table S7

Fitting results of the Copula function for TN observations-predictions pair in different stations.

Station	family	Par.1	Par.2	tau	AIC	BIC
TC	t	0.944	3.127	0.7851	-3526.544	-3515.815
QS	Gumbel	4.024		0.7515	-3338.221	-3332.856
MD	Gumbel	4.329		0.7690	-3599.480	-3594.116

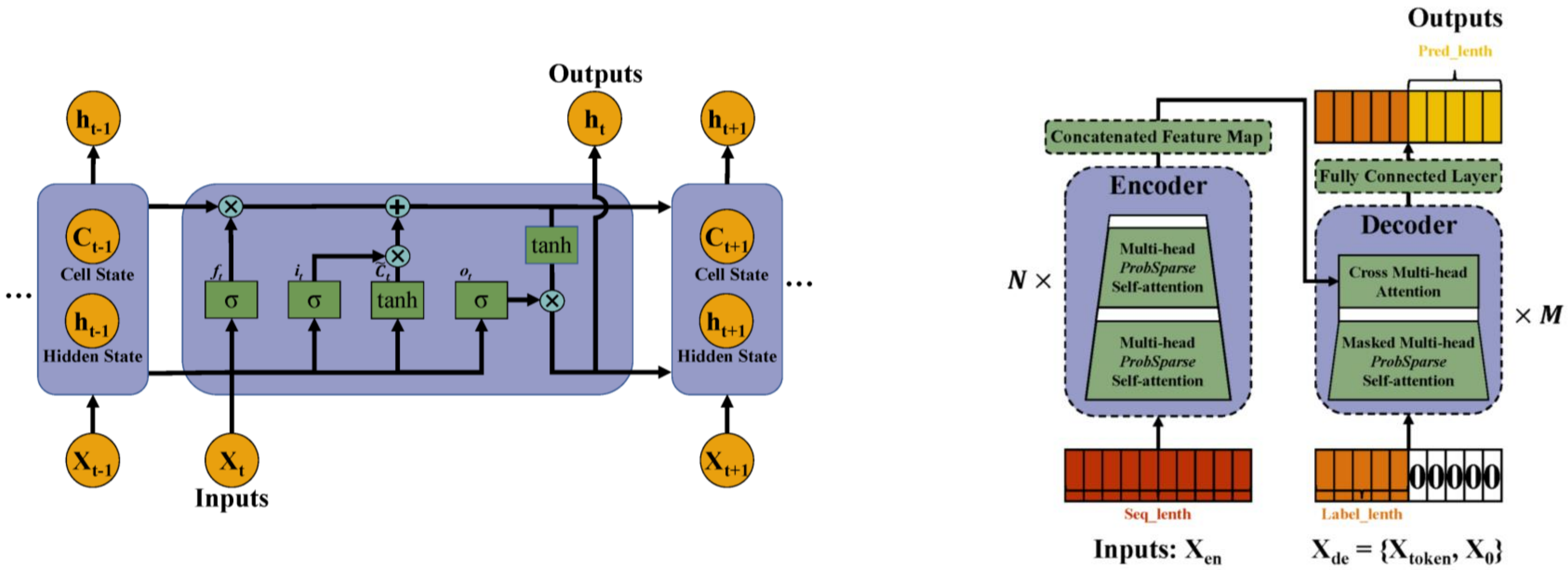


Fig. S1. Structures of the LSTM (a) and Informer (b) models.

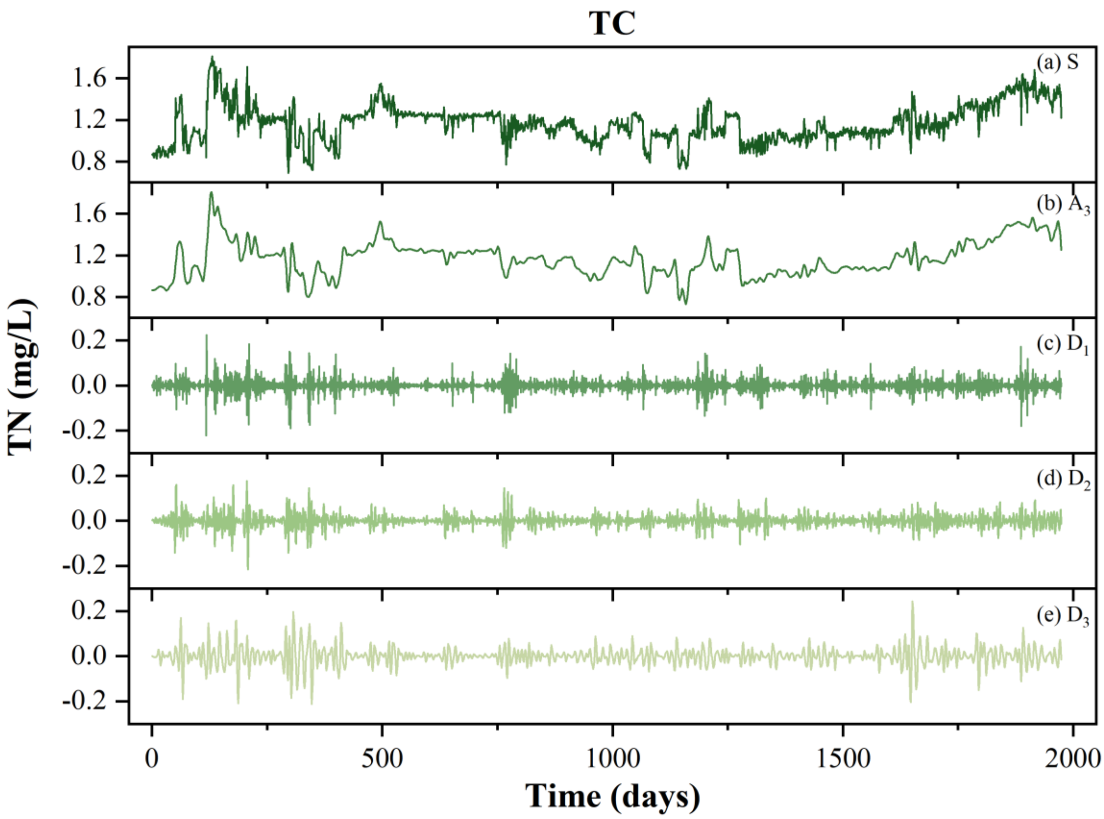


Fig. S2. Wavelet decomposition of the TN dynamics in the TC station (S: original series), using the db4 mother wavelet with approximation sub-series (A_3) and three levels of detailed sub-series (D_1 - D_3).

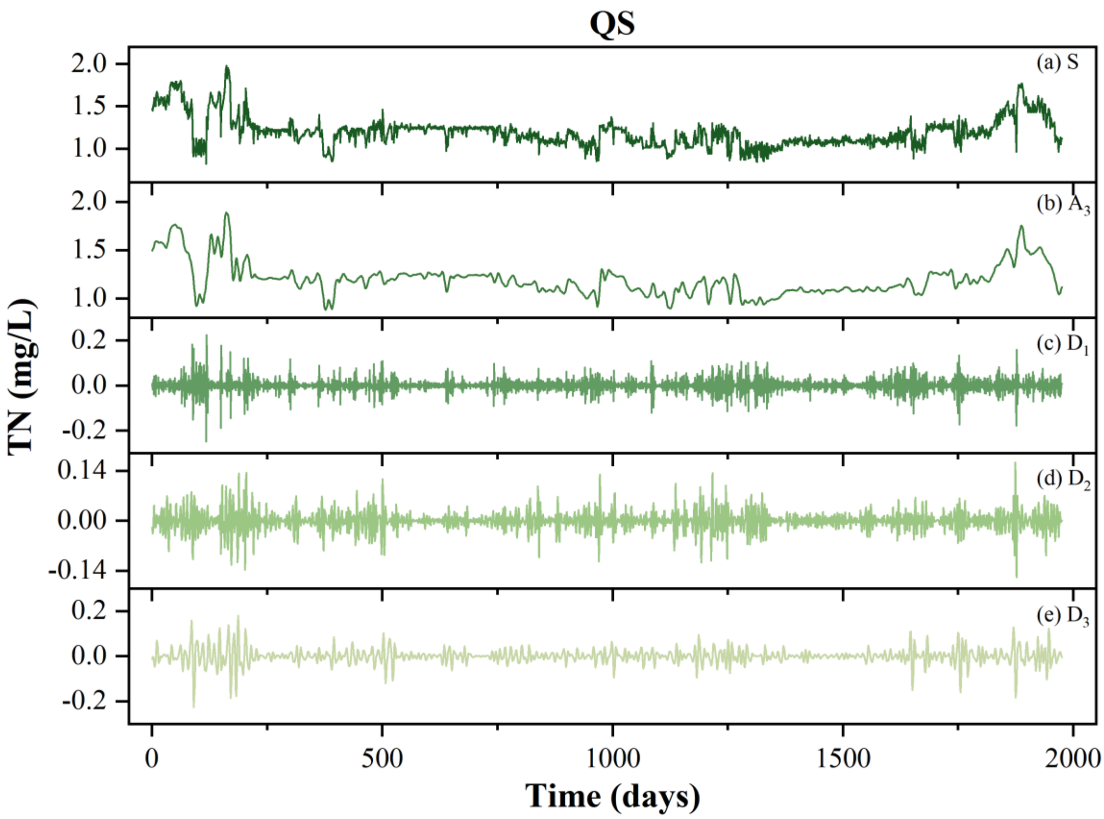


Fig. S3. Wavelet decomposition of the TN dynamics in the QS station (S: original series), using the db4 mother wavelet with approximation sub-series (A_3) and three levels of detailed sub-series (D_1 - D_3).

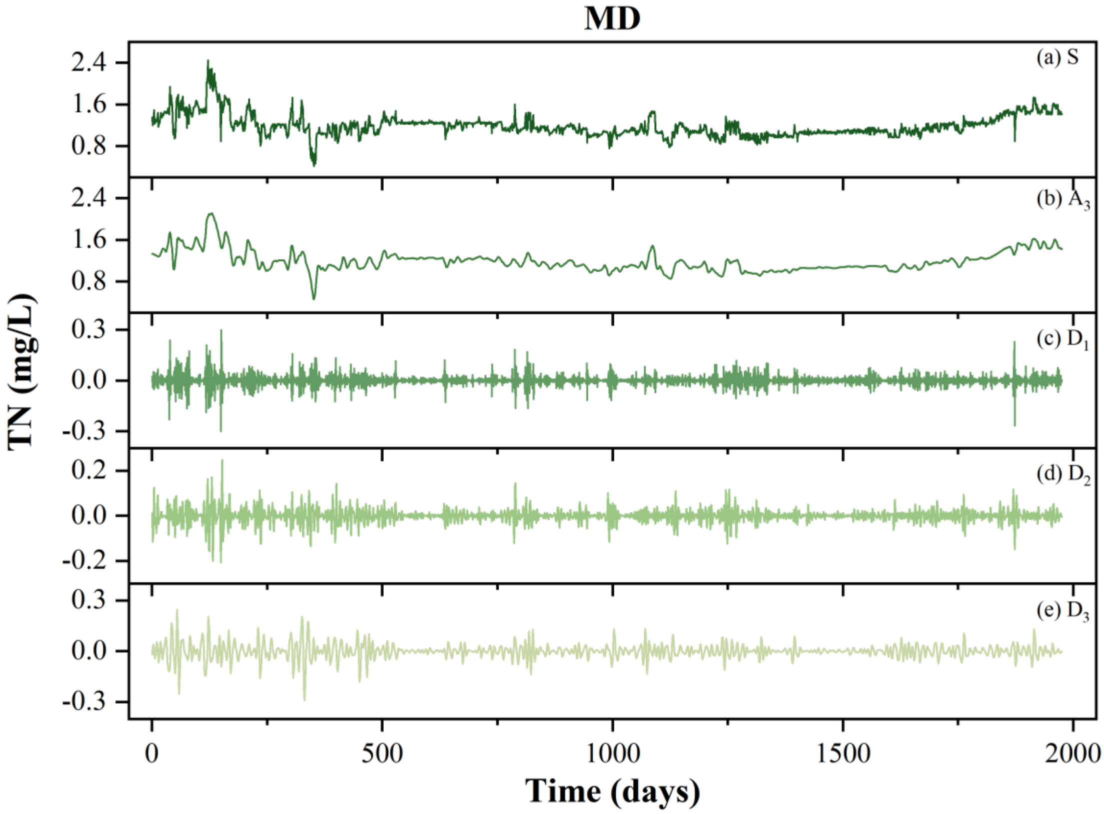


Fig. S4. Wavelet decomposition of the TN dynamics in the MD station (S: original series), using the db4 mother wavelet with approximation sub-series (A_3) and three levels of detailed sub-series (D_1 - D_3).

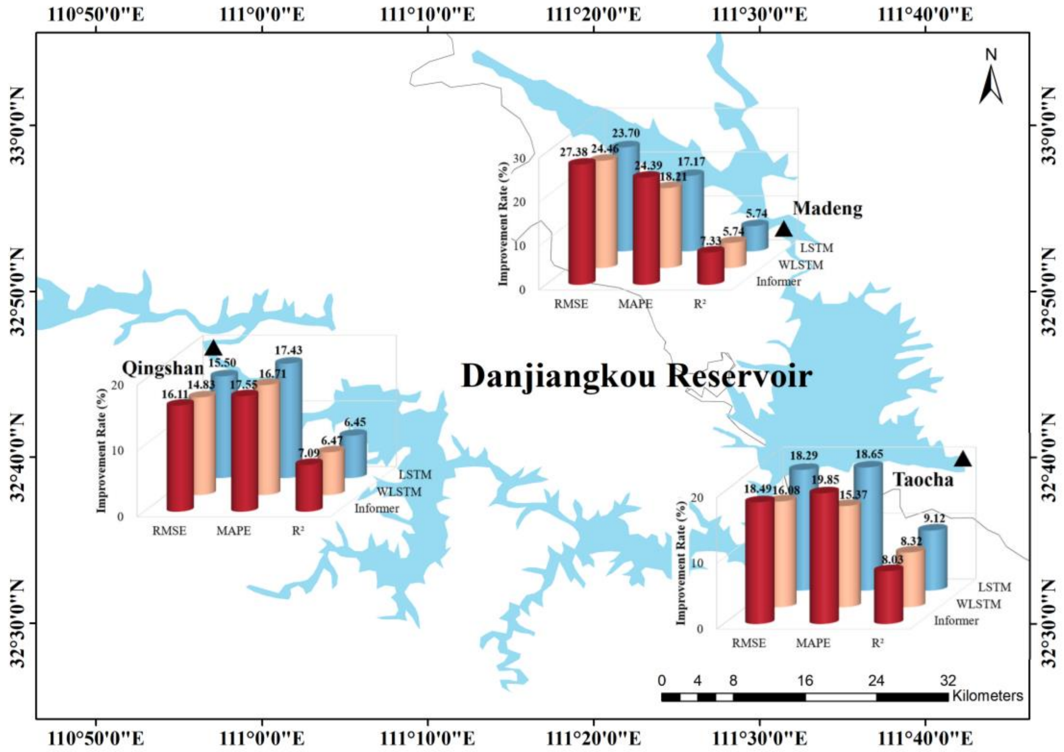


Fig. S5. Improvement rates of the WInformer model over other models in the evaluation stages for TN predictions (Note: Improvement rate = $\frac{|\text{Criterion (Winformer)-Criterion (another model)}|}{\text{Criterion (another model)}} \times 100\%$