Real-Time Probabilistic Forecasting of River Water Quality under Data Missing Situation: Deep Learning plus Post-Processing Techniques

Yanlai Zhou[1,2]

[1] State Key Laboratory of Water Resources and Hydropower Engineering Science, Wuhan University, Wuhan, 430072, China.

[2] Department of Geosciences, University of Oslo, P.O. Box 1047 Blindern, N-0316 Oslo, Norway.

**Correspondence to*: Yanlai Zhou (yanlai.zhou@whu.edu.cn).

Abstract

Quantifying the uncertainty of probabilistic water quality forecasting induced by missing input data is fundamentally challenging. This study introduced a novel methodology for probabilistic water quality forecasting conditional on point forecasts. A Multivariate Bayesian Uncertainty Processor (MBUP) was adopted to probabilistically model the relationship between the point forecasts made by a deep learning artificial neural network (ANN) and their corresponding observed water quality. The methodology was tested using hourly water quality series at an island of Shanghai City in China. The novelties relied upon: firstly, the use of a transfer learning algorithm to overcome flatten- and under-prediction bottlenecks of river water quality raised in artificial neural networks, and secondly, the use of the MBUP to capture the dependence structure between observations and forecasts. Two deep learning ANNs were used to make the point forecasts. Then the MBUP approach driven by the point forecasts demonstrated its competency in improving the accuracy of probabilistic water quality forecasts significantly, where predictive distributions encountered in multi-step-ahead water quality forecasts were effectively reduced to small ranges. The results demonstrated that the deep learning plus the post-processing approach suitably extracted the complex dependence structure between the model's output and observed water quality so that model reliability (Containing Ratio > 85% and average Relative Band-width < 0.25) as well as forecast accuracy (Nash-Sutcliffe Efficiency coefficient > 0.8 and Root-Mean-Square-Error < 0.4 mg/l) for future horizons from 1 hour up to 10 hours were significantly improved, even if the input data missing rate reaches 50%.

Keywords: Probabilistic forecast; River water quality; Missing data; Artificial intelligence; Deep learning

Nomenclature

Abbreviations

ANFIS	adaptive neural fuzzy inference system
ANN	artificial neural network
BASINS	better assessment science integrating point and nonpoint Sources
BPNN	back propagation neural networks
BMA	Bayesian model averaging
BUP	Bayesian uncertainty processor
COD_{Cr}	chemical oxygen demand using the chromium test
CNN	convolutional neural networks
CR	containing ratio
DO	dissolved oxygen
FC	fuzzy clustering
GLUE	generalized likelihood uncertainty estimation
HSPF	hydrological simulation program fortran
LSTM	long-short term memory
MBUP	multivariate Bayesian uncertainty processor
MLR	multiple linear regression
NARX	non-linear auto-regressive with exogenous inputs neural network
NH_3-N	ammonium nitrogen under the NH4/NO3/NO2 environment
NSE	Nash-Sutcliffe efficiency coefficient
PH	pondus hydrogenii
PLOAD	pollutant load
PMI	partial mutual information
QQ	quantile-quantile
QRNN	quantile regression neural networks
RB	relative band-width
RBF	radial basis function
RF	random forest
RMSE	root-mean-square-error
RTS	reference temporal sequence
TL-LSTM	transfer learning-based LSTM
TTS	target temporal sequence
SOM	self-organizing map
SVM	support vector machine
SWAT	soil and water assessment tool
USEPA	United States Environmental Protection Agency
WT	wavelet transform

Indices

i	index of monitoring station, from 1 to K
t	index of time step, from 1 to N
m	index of forecast horizon, from 1 to M

Parameters

N	number of time step

K	number of monitoring station
M	number of forecast horizon

Variables

$\mathrm{S^T}$	incomplete target temporal sequence (=$[\mathrm{S_1^T}, S_2^M, S_3^T]$)
$\mathrm{S_2^M}$	missing segment in $\mathrm{S^T}$
$\mathrm{S_1^T}$	first complete segment in $\mathrm{S^T}$
$\mathrm{S_3^T}$	last complete segment in $\mathrm{S^T}$
$\mathrm{S^R}$	highest correlation complete sequence of $\mathrm{S^T}$
$\mathrm{S^{RR}}$	highest correlation complete sequence of $\mathrm{S^R}$
$\hat{\mathrm{Y}}(t)$	forecasted data (i.e. model output) at the *t*th time
$\mathrm{Y}(t)$	observed data at the *t*th time
$\overline{\mathrm{Y}}(t)$	average of observed data at the *t*th time
$q_l(t)$	lower limitation of model forecasts at the *t* time
$q_u(t)$	upper limitation of model forecasts at the *t* time
S^i	complete sequence at the *i*th monitoring station (=$[\mathrm{S}_1^i, \mathrm{S}_2^i, \mathrm{S}_3^i]$)
S_1^i	fist segment of complete sequence S^i
S_2^i	second segment of complete sequence S^i
S_3^i	third segment of complete sequence S^i
$\mathrm{N_C}$	number of concordant pairs in two datasets
$\mathrm{N_D}$	number of discordant pairs in two datasets

1. Introduction

Water quality monitoring and forecasting became crucial problems since plenty of contaminants were discharged into the marine environment every year (Mian et al., 2018). Point sources (e.g. municipal and industrial sewage discharges, etc.) and nonpoint sources (e.g. farmland and livestock, aquaculture operations, etc.) are two common categories of water pollution sources (Perelman et al., 2012). It is imperative to make accurate and reliable water quality forecasts in advance to mitigate health risks and govern water pollution sources. A lot of studies were dedicated to building various models to forecast water quality (Fu et al., 2018; Newhart et al., 2019). Two fundamental challenging themes have occurred in water quality prediction for fulfilling the increasing public consciousness of human health. Firstly, missing input data not only would increase the difficulty in water quality forecasting but also would limit the discoveries in impact assessment. Secondly, real-time water quality forecasting is gradually shifting from traditional deterministic forecasting to probabilistic forecasting.

Water quality datasets were collected using automated machine sensors located at different sites. Due to facility malfunction, routine maintenance, changes of sensors setting, insufficient sampling and other reasons, data collection usually contained a large number of missing data (Ekeu-wei et al., 2018). Missing data situation is not a unique problem for water quality prediction but a ubiquitous concern in many scientific fields (Gao et al., 2017; Tencaliec et al., 2015), such as hydro-meteorology, air quality and traffic load, etc. Data imputation (Yang et al., 2017) and transfer learning (Che et al., 2018) algorithms are two common methods used to mitigate the impacts of missing values on forecasting (Lepot et al., 2017). The data imputation algorithm is direct to fill the missing data from the perspective of data spatial-temporal scale while the transfer learning algorithm is indirect to estimate the missing data from the perspective of model and parameters transferring. Although the combination of data imputation algorithm and forecast model was widely used, previous studies suggested that this combination was easy to create systematical flatten-prediction and under-prediction results due to inducing a substantial bias in multi-step-ahead forecasts (Ding et al., 2018). Accordingly, the topic of integrating transfer learning algorithm and forecast model for multi-step-ahead water quality forecasts is interesting, as it is becoming a challenge for water quality forecasting under high data missing rate.

In recent years, two main categories of forecasting models, physically-based (or chemical transport) (Krapu et al., 2019) and data-driven (or artificial intelligence) (Regina and Stefan 2019; García-Alba et al., 2019) ones, were introduced for water quality forecasting. The United States Environmental Protection Agency (USEPA) developed the Better Assessment Science Integrating Point and Nonpoint Sources (BASINS) software system, which integrated several powerful hydrological and water quality simulation packages of the Hydrological Simulation Program Fortran (HSPF), the Soil and Water Assessment Tool (SWAT), Pollutant Load (PLOAD) and the enhanced stream water quality module (https://www.epa.gov/). The advantage of physically-based models is their capability to adequately simulate the chemical mechanisms of the water pollution process, whereas their disadvantages are that they become invalid for imitating the water pollution process if the data missing and changing environment occurred (Krapu et al., 2019). The data-driven models can handle nonlinear and highly stochastic predictions through dynamically and adaptively correcting model elements (e.g. structure, algorithms and parameters) (Isiyaka et al., 2019; Yaseen et al., 2019). Additionally, deep learning is classified as one of machine learning algorithms based on Artificial Neural Networks (ANNs) that employs multiple hidden processing layers between the input and output layers to progressively extract higher-level (whatever it be linear or complex nonlinear) features from the raw datasets (Yann et al., 2015). The core theoretic principles of deep learning are three-fold: Firstly, deep learning is a learning algorithm based on ANNs. Secondly, artificial neural networks have multiple (≥ 2) hidden layers between the input and output layers. Thirdly, deep learning is commonly used to discover intricate structures in relation to large data sets (Schmidhuber 2015). In the past decades, ANNs were successfully utilized for water quality and environmental prediction, classification and pattern recognition (Aguilera et al., 2001; Peleato et al., 2018). For instance, the Random Forest (RF), the Quantile Regression Neural Networks (QRNN), the Back Propagation Neural Networks (BPNN), the Radial Basis Function (RBF), the Self-Organizing Map (SOM), the Support Vector Machine (SVM), the Non-linear Auto-Regressive with eXogenous inputs neural network (NARX), the Adaptive Neural Fuzzy Inference System (ANFIS), the Convolutional Neural Networks (CNN) and the Long-Short Term Memory (LSTM) were widely introduced for water quality (Pearce et al., 2013; Jiang et al., 2016; Zhang et al., 2018; Gerhard and Gunsch, 2019; Helbich et al., 2019) and hydro-meteorological forecasting (Cannon 2011; Chang and Tsai, 2016; Zhou et al., 2019a,b). Owing to the powerful learning capability for time-sequential data, the LSTM was successfully applied in speech recognition, image segmentation, traffic volume prediction, and meteorological prediction (e.g., Akbari et al., 2018; Yi et al., 2018; Zhao et al., 2018; Gallego et al., 2019; Kao et al., 2020), etc. However, the available literature on utilizing LSTM for multi-step-ahead water quality forecasts under the missing data conditions is limited in number (Liang et al., 2019; Tiyasha and Yaseen, 2020). The LSTM was introduced for predicting traffic flow with the missing data (Tian et al., 2018), whereas it was incline to produce flatten values if the data missing rate was high (≥ 0.30). In other words, when plenty of input datasets were missed, the LSTM model was easier to trigger flatten prediction and/or under-prediction problems. Hence, under the high missing data conditions, it is essential to conduct the hybrid of the transfer learning algorithm and deep learning LSTM model for improving the reliability and accuracy of data-driven water quality forecasting models.

The uncertain and inaccurate meteorological forcing, initial condition (i.e. natural and anthropogenic sources), and model structure and parameters have a significant impact on the reliability and accuracy of water quality forecasts (Moreno-Rodenas et al., 2019). Several techniques were commonly used to quantify the uncertainty of water quality forecasts, for instance, (1) pre-processing techniques: the Fuzzy Clustering (FC) method (Kim and Pachepsky, 2010), the Wavelet Transform (WT) (Barzegar et al., 2018) and the bias-correction method (Libera and Sankarasubramanian, 2018) and (2) post-processing techniques: the Multiple Linear Regression (MLR) (Wallace et al., 2016), the Kalman filtering (Rajakumar et al., 2019; Zhou et al., 2020), the Generalized Likelihood Uncertainty Estimation (GLUE) (Zhang et al., 2015), the Bayesian Model Averaging (BMA) (Mok et al., 2018) and the Bayesian Uncertainty Processor (BUP) (Borsuk et al., 2002; Arhonditsis et al., 2019). The creation of probabilistic forecast intervals could be taken as one of the effective approaches to quantify the impact of different uncertainties on water quality forecasting (Krapu et al., 2019). The deterministic forecast model plus the probabilistic post-processing techniques were widely employed to complement the predictive information of point-value predictions (Camacho et al., 2018). The BUP was a vital component of probabilistic post-processing techniques used to measure the predictive uncertainties (Herr and Krzysztofowicz, 2015). Follow up on the BUP framework developed by Krzysztofowicz (1999), two probabilistic post-processing approaches were developed and effectively adopted to predict water quality time series (Liang et al., 2016; Yang et al., 2016). The univariate BUP (UBUP) (Krzysztofowicz, 2002) approach was employed to extract the nonlinear bivariate correlation between forecasts and observations, whereas the multivariate BUP (MBUP) (Krzysztofowicz and Maranzano, 2004) approach was used to quantify the nonlinear multivariate ($\geq$ 3) correlation between forecasts and observations (Krzysztofowicz and Maranzano, 2004). Bayesian multivariate probabilistic post-processing (i.e. MBUP) not only puts forward challenges but also brings about various opportunities for probabilistic water quality forecasting. Hence, it is interesting to implement in-depth research on the MBUP for characterizing and decreasing the uncertainty associated with multi-step-ahead water quality forecasting by extracting the nonlinear multivariate correlation between forecasts and observations.

This study proposed an MBUP-based approach hybriding deep learning ANN and MBUP to reduce the prediction intervals of multi-step-ahead water quality forecasts under the data missing situation. There existed two main contributions in this work: First, seamless integration of transfer learning and deep learning ANN was conducted to overcome flatten/under-predictions of deterministic river water quality forecasts induced by missing input data. Second, the multivariate uncertainty processor (i.e., MBUP) was further employed as the post-processing technology to increase the reliability of probabilistic river water quality forecasts.

In the beginning two ANNs, a Transfer Learning-based LSTM (i.e. TL-LSTM) and a standard LSTM, were utilized to construct water quality forecast models under the data missing situation, and the model that created more reliable and accurate point forecasts was employed to carry out probabilistic forecasting. Next, the MBUP probabilistic post-processing approach was implemented to transform point water quality forecasts into probabilistic water quality forecasts. Finally, the meteorological and water quality series at an island of Shanghai City in China were utilized as a study case to demonstrate the reliability and applicability of the deep learning ANN plus the MBUP post-processing approach.

2. Methods

Figure 1 illustrated the probabilistic forecast architecture that integrated the multi-output deep learning LSTM model with h (≥ 2) hidden layers (Figure 1 (a), descried in Section 2.1 and Appendix A), the transfer learning algorithm (Figure 1 (b), described in Section 2.2 and Appendix B) and the MBUP probabilistic forecast approach (Figure 1 (c), described in Section 2.3). The TL-LSTM model was employed to create deterministic point forecasts under the data missing condition, where the LSTM model was taken as the benchmark. The deterministic forecast model was established and evaluated to provide inputs for the following probabilistic forecasts. And then, the MBUP approach was used to create probabilistic forecasts. The related methods were briefly described below.

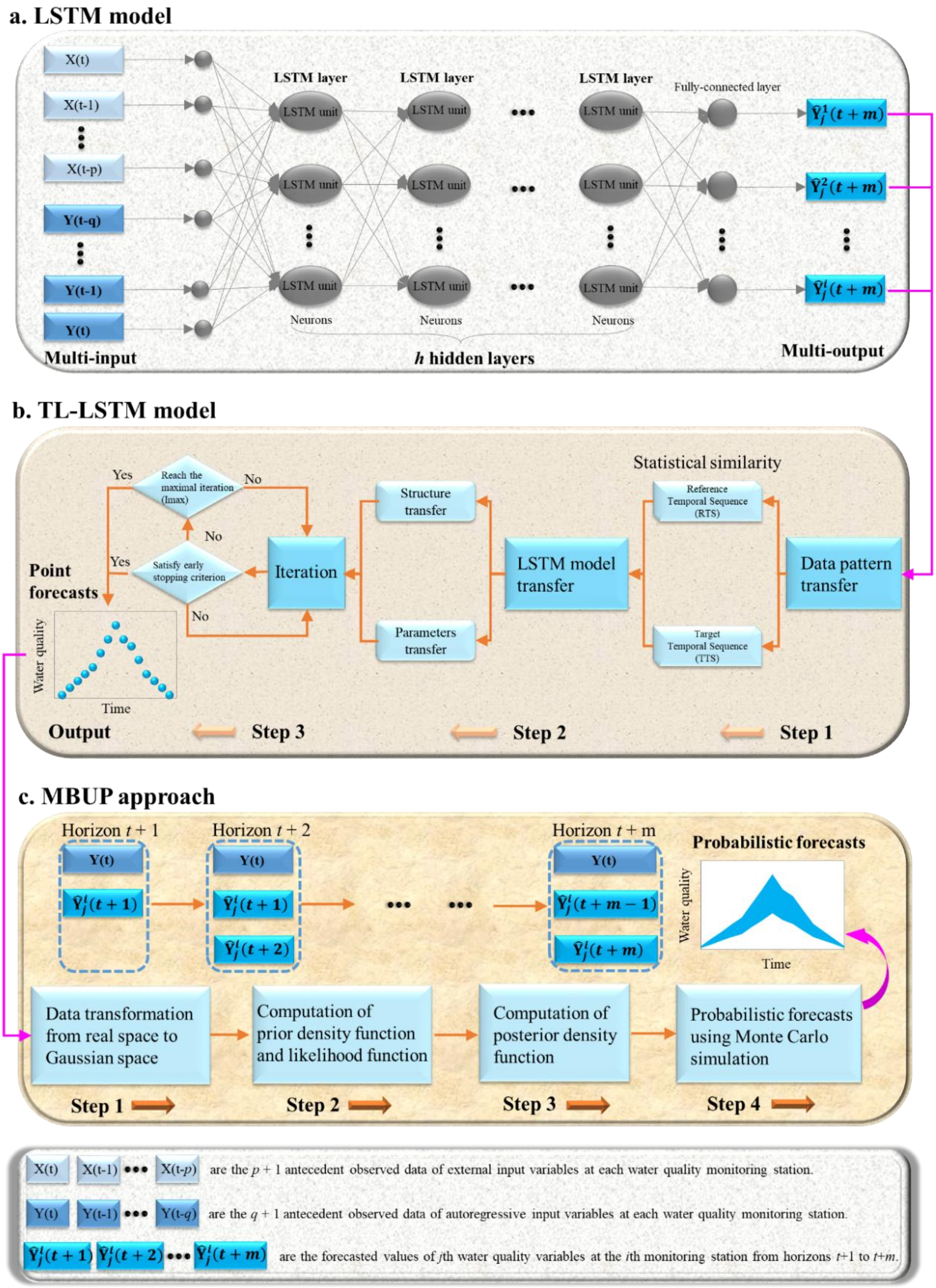

Fig. 1. Probabilistic forecast architecture. (a) LSTM neural network model. (b) Hybrid of Transfer Learning and LSTM model (TL-LSTM). (c) MBUP approach.

2.1 *Long Short-Term Memory (LSTM) neural network*

The ANN models usually considered forecasts of water quality as a mathematical function of water quality as well as hydro-meteorological variables (Olsen et al., 2012; Guo et al., 2019). The LSTM model adopted in this study is a special architecture of recurrent neural network proposed by Hochreiter and Schmidhuber (1997). The LSTM model is capable of learning 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. The difference between LSTM and other ANNs is that the hidden layer in LSTM is constituted of an internal self-looped unit. Moreover, the common ANN models (e.g. BPNN, ANFIS, NARX) need to construct multiple independent models to make water quality forecast at various monitoring stations whereas the multi-output deep learning LSTM model h (≥ 2) hidden layers demands only one model to achieve regional water quality multi-outputs (Figure 1(a)). The detailed description concerning the LSTM structure, the readers could find it from Appendix A.

2.2 *Hybrid of Transfer Learning and LSTM model (TL-LSTM)*

The transfer learning algorithm can transfer the learned knowledge from one similar domain (Reference) to another related domain (Target). The transfer learning algorithm was commonly used in cases that the forecast model for the target domain is too complicated or the target domain has long-interval data missing condition (Gupta et al., 2019). In this study, the transfer learning algorithm was introduced to learn and transfer the knowledge from the Reference Temporal Sequence (RTS) which has complete data to the Target Temporal Sequence (TTS) which has data missing situation. Transfer learning mechanisms are classified to three different settings, i.e. data pattern transfer (e.g. trend and statistical characteristics), model transfer (e.g. model structure and parameters) and task transfer (e.g. multi-task learning about classification and clustering) (Pan and Yang, 2009). Since it needs to learn and model the pattern from an RTS, both data pattern transfer (statistic characteristics) and model transfer (model structure and parameters) would be adopted in this study. Figure 1 (b) showed the seamless integration of the transfer learning algorithm and the LSTM model (TL-LSTM model). The general implementation procedure of the TL-LSTM model was described in Appendix B.

For comparison analysis, two deterministic forecast models (TL-LSTM & LSTM) were established and evaluated to provide inputs for the following probabilistic forecasts. The differences between TL-LSTM and LSTM models consist of: (1) the former uses the transfer learning algorithm to process the data missing situation whereas the latter does not use it; (2) the input data of two models in the training and validating stages are significantly different, as shown in Table 1.

Table 1. Input data of deterministic forecast models under missing data conditions

Stage	**TL-LSTM model**	**LSTM model**
Training	$\{[\mathrm{S}_{t-q}^{\mathrm{R}}, \mathrm{S}_{t-q+1}^{\mathrm{R}}, \cdots, \mathrm{S}_{t-1}^{\mathrm{R}}, \mathrm{S}_{t}^{\mathrm{RR}}] \to [\mathrm{S}_{t+m}^{\mathrm{R}}]\}$	
Validating	$\{[\mathrm{S}_{t-p}^{\mathrm{T}}, \mathrm{S}_{t-p+1}^{\mathrm{T}}, \cdots, \mathrm{S}_{t-1}^{\mathrm{T}}, \mathrm{S}_{t}^{\mathrm{R}}] \to [\mathrm{S}_{t+m}^{\mathrm{T}}]\}$	$\{[\mathrm{S}_{t-p}^{\mathrm{T}}, \mathrm{S}_{t-p+1}^{\mathrm{T}}, \cdots, \mathrm{S}_{t-1}^{\mathrm{T}}, \mathrm{S}_{t}^{\mathrm{R}}] \to [\mathrm{S}_{t+m}^{\mathrm{T}}]\}$
Testing	$\{[\mathrm{S}_{t-p}^{\mathrm{T}}, \mathrm{S}_{t-p+1}^{\mathrm{T}}, \cdots, \mathrm{S}_{t-1}^{\mathrm{T}}, \mathrm{S}_{t}^{\mathrm{R}}] \to [\mathrm{S}_{t+m}^{\mathrm{T}}]\}$	

Notes: Each stage (training, validating and testing) of the dataset was erased with one percentage (e.g. 50%) during the establishment and application of the LSTM models. S^{R} and S^{RR} were the selected RTSs. S^{T} was the incomplete TTS. Take the incomplete TTS with one missing segment $S^{T} = [S_1^{T}, S_2^{M}, S_3^{T}]$ for example, S_2^{M} was the missing segment, S_1^{T} and S_3^{T} were the complete segments. If S_2^{M} was at the beginning or the end of S^{T}, S_1^{T} or S_3^{T} was empty dataset. S^{R} was the highest correlation complete sequence of S^{T} while S^{RR} was the highest correlation complete sequence of S^{R}.

2.3 Multivariate Bayesian Uncertainty Processor (MBUP)

Four basic steps in Figure 1(c) constituted the general implementation procedures of the MBUP and were briefly described as follows (Krzysztofowicz and Maranzano, 2004).

Step 1: Data conversion. Both observed and forecasted datasets with real space were transformed to the Gaussian data by using the meta-Gaussian strategy (Krzysztofowicz, 2002).

Step 2: Determination of prior density and likelihood functions. The meta-Gaussian strategy was also employed to compute the prior density function and the likelihood function.

Step 3: Determination of posterior density function. After the prior density and likelihood functions were determined, the posterior density function was calculated accordingly.

Step 4: Probabilistic forecasts. A Monte Carlo simulation was conducted to create probabilistic forecasts. A realization of observation at the horizon *m* was simulated according to the posterior density function and the Monte Carlo simulation was repeated for K times. K was the number of Monte Carlo simulation and was set as 1000 in this study. 90 % confidence intervals were employed to reveal the uncertainty of water quality probabilistic forecasts. And then, both observed and forecasted datasets (e.g. DO, NH_3-N, COD) with the Gaussian space were transformed to the real space for evaluating the performance of MBUP probabilistic forecasts.

The general implementation programming of the machine learning model (e.g., LSTM) and the transfer learning algorithm can be obtained from the Statistics and Machine Learning Toolbox of the Matlab software (website: https://ww2.mathworks.cn/products/statistics.html#machine-learning) while the Bayesian model can be acquired from the Econometrics Toolbox of the Matlab software (website: https://ww2.mathworks.cn/help/econ/index.html).

2.4 Evaluation criteria

For comparison purpose, the Root-Mean-Square-Error (RMSE) as well as the Nash-Sutcliffe Efficiency coefficient (NSE) were introduced to evaluate the performance of deterministic forecast models. The indicators of RMSE and NSE were presented as follows.

$$\mathrm{RMSE} = \sqrt{\frac{1}{\mathrm{N}}\sum_{t=1}^{\mathrm{N}}\left(\hat{\mathrm{Y}}(t) - \mathrm{Y}(t)\right)^2}, \quad \mathrm{RMSE} \geq 0 \tag{1}$$

$$\mathrm{NSE} = 1 - \frac{\sum_{t=1}^{\mathrm{N}}\left(\hat{\mathrm{Y}}(t) - \mathrm{Y}(t)\right)^2}{\sum_{t=1}^{\mathrm{N}}\left(\mathrm{Y}(t) - \bar{\mathrm{Y}}(t)\right)^2}, \quad \mathrm{NSE} \leq 1 \tag{2}$$

where $\hat{\mathrm{Y}}(t)$, $\mathrm{Y}(t)$and $\bar{\mathrm{Y}}(t)$ is the forecasted data (i.e. model output), observed data and the average of observed data at the *t*th time, respectively. N is the number of time step.

The average Relative Band-width (RB) as well as the Containing Ratio (CR) were adopted to evaluate the performance of probabilistic forecast models (Gneiting, 2008; Xiong et al., 2008). Their mathematical formulas were described below.

$$\mathrm{RB} = \frac{1}{\mathrm{N}}\sum_{t=1}^{\mathrm{N}}\left(\frac{q_u(t)-q_l(t)}{\mathrm{Y}(t)}\right) \tag{3}$$

$$\mathrm{N}(t) = \begin{cases} 1, & if\left(q_l(t) \le \widehat{\mathrm{Y}}(t) \le q_u(t)\right) \\ 0, & else \end{cases} \tag{4a}$$

$$\mathrm{CR} = \frac{\sum_{t=1}^{\mathrm{N}}\mathrm{N}(t)}{\mathrm{N}} \times 100\% \tag{4b}$$

where $q_l(t)$ and $q_u(t)$ are the lower and upper limitations of the model forecasts with respect to a confidence level at the *t* time. If the NSE and CR values are higher and the RMSE and RB values are lower, the models would achieve better performance.

3. Study area and background discussion

3.1 Study area

The study area (Figure 2) is briefly introduced as follows. The island in Shanghai City of China has 52 km^2 administrative area and is located at the estuary of Yangtze River Delta. Annual precipitation ranged between 600 mm and 1400 mm as well as mean annual temperature is 15 ℃. In 2018, the land uses in this island are as follows: 1.95 % urbanization, 65.68 % agriculture, 1.45 % industry, 12.32 % forest, 18.37 % surface water and 0.23 % others while the total population of the island was about 34 thousand (source: https://sthj.sh.gov.cn/, in Chinese). With the economy and population fast boosting, one of the hot topics in Shanghai City concentrates on water quality deterioration. People in the island are compelled to handle a high-level intervention of water pollution. In recent years, water pollution got a serious focus in Shanghai City of China (Liu et al., 2015; Zhao et al., 2015). Water pollution not just induced cancer, stone and cardiovascular sclerosis diseases but also caused a matter of life or death. Hence, it must make accurate and reliable water quality forecasts to adequately process the health risk caused by regional water pollution.

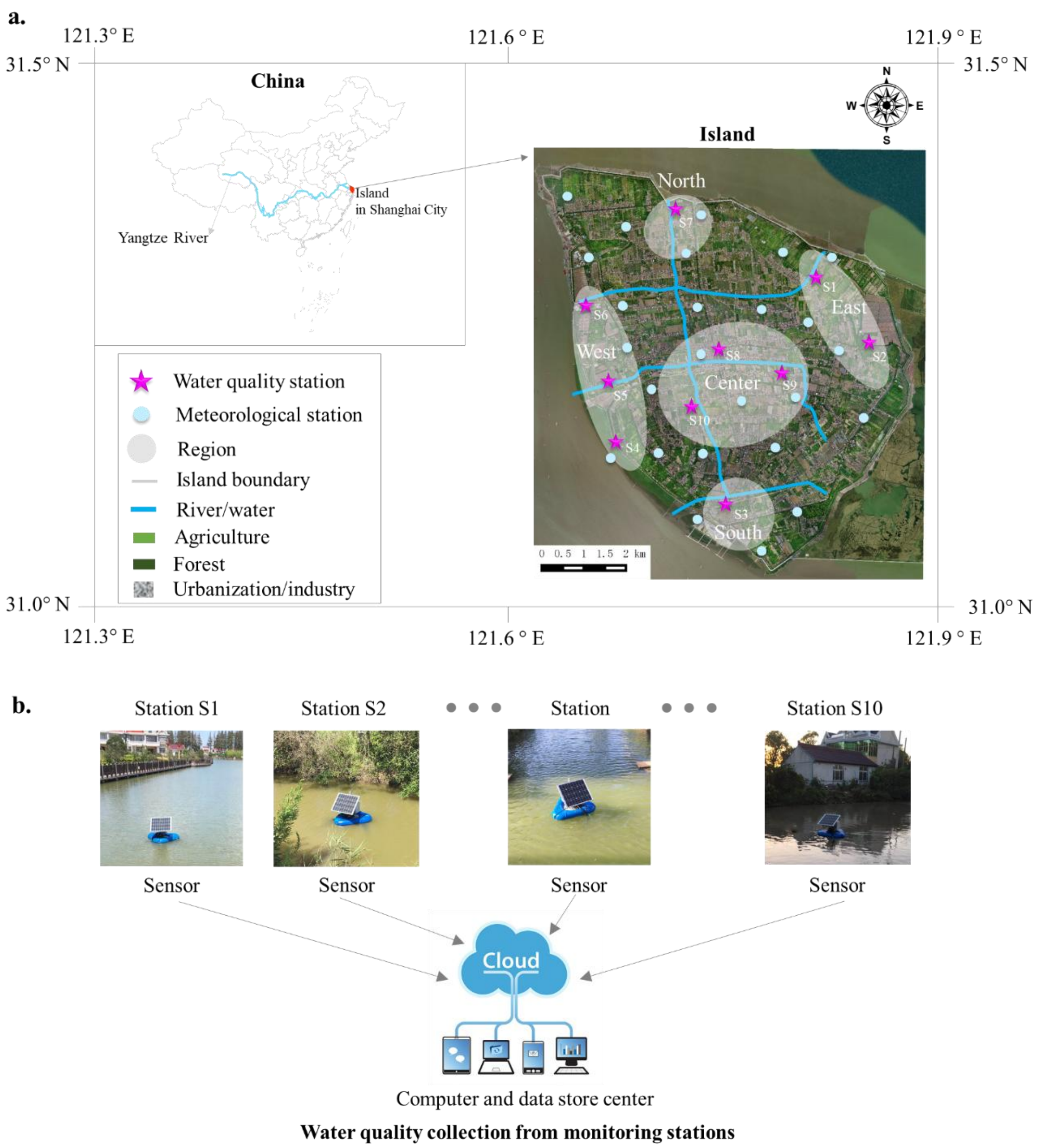

Fig. 2. Study area and water quality data collection. (a) Meteorological and river water quality monitoring stations in the island of Shanghai City. (b) Water quality data collection from monitoring stations.

The positions of the island, 25 meteorological as well as 10 water quality monitoring stations monitoring stations were presented in Figure 2(a), while water quality datasets are collected from monitoring stations as depicted in Figure 2(b). The basic information on ten monitoring stations in five regions were summarized in Table 2. Hourly data of water quality factors (nine variables: Dissolved Oxygen (DO), Ammonium Nitrogen (NH_3-N) under the $NH_4/NO_3/NO_2$ environment, Chemical Oxygen Demand (COD_{Cr}) using the chromium test, Pondus Hydrogenii (pH), oxidation-reduction potential (ORP), electrical conductivity (EC), turbidity, water level and water temperature) and meteorological factors (three variables: precipitation, wind speed and light intensity) over a span of four years (31/08/2015-31/08/2019) are available.

Table 2. Basic information on ten monitoring stations in five regions

Region	Station	Type of pollution	Source
East	S1 & S2	Nonpoint source	Aquaculture or natural area
South	S3	Point source	Industry
West	S4-S6	Point source	Industry
North	S7	Nonpoint source	Farmland and livestock
Center	S8-S10	Point source	Urban domestic sewage

The data calibration procedure was executed in the phase of the measurement prior to model construction and validation. The Oxidation-Reduction Potential (ORP) values were calibrated to potential redox (Eh) and pH using Quinhydrone, where a typical Quinhydrone calibration (to the standard hydrogen electrode), using an ORP meter was undertaken at pH = 4, and 7 (an example calibration is, Eh (mV) = −65.667 pH + 744.67 + ORP (mV)). The data calibration procedure is similar to what described in Jardim (2014). For more information about the field measurement of ORP, the interested reader is pointed to the operating procedure provided by the U.S. Environmental Protection Agency and some international examples of the quinhydrone calibration procedure (http://www.pulseinstrument.com/ and http://www.astisensor.com/). The procedures of data calibration and data quality control were also applied to the datasets of EC, DO, COD and Nitrogen (e.g., Fofonoff and Millard, 1983).

Similar to Shrestha and Kazama (2007), the statistical analysis was performed by using the principal component analysis in this study as to what water quality factors and meteorological factors were the most important in explaining the variability of river water quality concentrations. The twelve water quality and meteorological factors afforded more than 94% contribution to river water quality concentrations, where the eight factors (precipitation, water level, water temperature, DO, COD_{Cr}, EC, NH_3-N, ORP) afforded more than 87% contribution as well as the other factors (pH, turbidity, wind speed and light intensity) afforded more than 7% contribution. Besides, the multivariate statistical analysis by Shrestha and Kazama (2007) clearly pointed out that the factors contribution to water quality concentrations are closely associated with the streamflow (or water level) and water temperature in natural regions; organic pollution (point source: domestic wastewater) in less pollution regions; organic pollution (point source: domestic wastewater) and nutrients (non-point sources: farmland and livestock) in medium pollution regions; and both organic pollution and nutrients (point sources: domestic wastewater, wastewater treatment plants and industries) in high pollution regions.

The correlation analysis of input variables using the Kendall tau coefficient further revealed that the input variables (water level, DO, COD_{Cr}, EC, NH_3-N, ORP, turbidity, wind speed and light intensity) would be regarded as the independent factors, meanwhile, partial dependencies between the input variables (link water temperature to EC, COD, or DO, and link precipitation to water level) were identified. The reasons for taking both 9 water quality factors (DO, NH_3-N, COD_{Cr}, pH, ORP, EC, turbidity, water level and water temperature) and 3 meteorological factors (precipitation, wind speed and light intensity) as input variables simultaneously consist of: (1) various pollution sources with natural, organic pollution or nutrients appeared in the ten monitoring stations of five regions (Table 2) as well as modeling various pollution sources implied demand for different model inputs; (2) the multi-input and multi-output LSTM model (described in section 2.1) adopted in this study not only could grant the input variables to have independent features and partial dependencies, but also could adaptively adjust the model weight parameters (varied in the interval [0, 1]) for different input variables according to the pollution sources in various monitoring stations. Consequently, the forecasts for water quality (e.g. DO, NH_3-N, COD_{Cr}) are considered as a math function of water quality (9 factors) as well as meteorological (3 factors) variables. Each forecast model could output the forecast results of water quality (e.g. DO, NH_3-N and COD_{Cr}) at 10 stations.

In this study, the Partial Mutual Information (PMI) (Sharma, 2000) and Kendall tau coefficient methods were used to select input variable combinations. In accordance with the highest values of the PMI ($\geq$ 0.5) (Galelli et al., 2014) as well as the Kendall tau coefficient ($\geq$ 0.6) (Zhou et al., 2019a), the results of selected time lags were identical. In brief, the time lags of 1h-7h were identified for water quality factors as well as the time lags of 1h-5h were identified for meteorological factors.

A total of 3,157,920 (= 4 (years) × 365 (or 366 days) × 24 (hours) × 10 (stations) × 9 (factors)) hourly water quality datasets and a total of 2,631,600 (= 4 (years) × 365 (or 366 days) × 24 (hours) × 25 (stations) × 3 (factors)) hourly meteorological datasets were used in this study, where 40 % datasets (31/08/2015-06/04/2017) were employed for ANN model training while the remaining 30 % datasets (07/04/2017-18/06/2018) and 30 % datasets (19/06/2018-31/08/2019) were employed for validating and testing ANN model respectively.

3.2 Background discussion

Figure 3 presented the statistic indexes of DO, NH_3-N and COD_{Cr} concentrations at five regions while Table 3 summarized the statistic indexes of the other 9 input factors at five regions. Since the higher value of DO and the lower values of COD_{Cr} and NH_3-N usually indicated better water quality, the three water quality factors (i.e. DO, NH_3-N and COD_{Cr}) were specified to discuss the research background. It indicated that the values of the maximum, average as well as quartiles of COD_{Cr} and NH_3-N (DO) concentrations at the North region were the highest (lowest) whereas those in the Center region were the lowest (highest), which would be owing to the primary source of water pollution of a region. The nonpoint source pollution from farmland and livestock was the primary source of water pollution at the North region while the point source pollution from urban domestic sewage was the primary source of water pollution at the Center region. In other words, the nonpoint source pollution (agriculture) was stronger driving force of water pollution than the point source pollution (industry and urban domestic sewage) in this island. The five regions do represent three situations (agriculture (East & North), industry (South & West) and urban domestic sewage (Center)) and significant differences in the statistical indexes of the monitoring data (Figure 3).

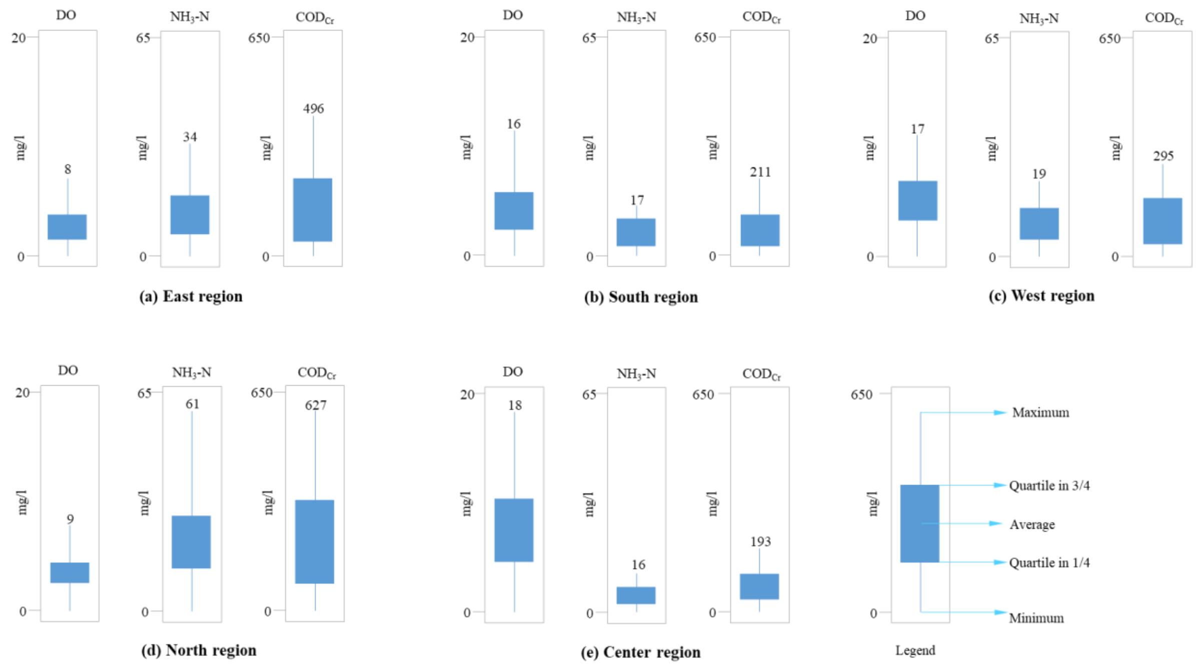

Fig. 3. Statistic indexes of DO, NH_3-N and COD_{Cr} concentrations at five regions (a–e) in the island. The abbreviations (max, ave, min, std) denote the maximum, average, minimum and standard deviation respectively. The time period of statistic covers four years (31/08/2015-31/08/2019).

Table 3. Statistic indexes of the other 9 input factors at five regions

Region	Index	Factor								
		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
East	Max.	13.6	1380.6	1524.8	1087.3	14.9	39.8	53.2	9.2	20.7
	Ave.	7.5	271.0	351.3	241.6	12.9	17.6	22.7	3.4	16.5
	Min.	6.9	200.0	5.8	7.1	4.3	0.0	0.0	0.0	0.0
South	Max.	11.5	1460.7	1337.0	948.5	14.7	35.8	47.8	4.4	24.5
	Ave.	7.5	219.7	334.4	229.4	12.0	18.1	19.7	1.8	17.8
	Min.	6.2	153.5	4.2	6.4	3.6	0.0	0.0	0.0	0.0
West	Max.	13.7	1388.6	1358.2	980.3	15.2	39.8	43.6	4.3	21.2
	Ave.	7.5	217.1	462.6	315.3	12.0	17.9	14.6	1.6	16.9
	Min.	6.3	128.1	7.6	5.8	2.1	0.0	0.0	0.0	0.0
North	Max.	19.3	1402.1	1679.2	1191.5	14.3	39.8	51.7	6.7	19.7
	Ave.	7.5	234.5	453.9	307.4	11.3	17.8	20.5	2.4	15.3
	Min.	6.9	180.1	8.4	10.3	1.8	0.0	0.0	0.0	0.0

	Max.	11.5	1400.9	1113.8	751.8	14.6	32.2	45.2	3.6	22.3
Center	Ave.	7.5	229.9	439.4	298.7	11.3	17.6	15.9	0.7	17.2
	Min.	6.8	180.3	7.9	4.7	0.9	1.3	0.0	0.0	0.0

The abbreviations of Max, Ave and Min denoted the maximum, average and minimum. The factors in columns No. (1)-(9) were pondus hydrogenii (/), oxidation-reduction potential (mV), conductivity (S/m), turbidity (mg/l), water level (m), water temperature (℃), precipitation (mm/h), wind speed (m/s) and light intensity (mega-joule/m^2) respectively.

4. Results

The LSTM and TL-LSTM models were used to make deterministic forecasts of river water quality independently, and then the MBUP approach was used to make probabilistic forecasts of river water quality. The results and findings were displayed in the order of the deterministic water quality forecasts (Section 4.1) and the probabilistic water quality forecasts and summarization (Section 4.2), shown as follows.

4.1 Deterministic water quality forecasts

Lead times up to 10 hours (t+1 - t+10) at a temporal scale of one hour were employed to evaluate the validity of the two deterministic water quality forecast models (LSTM and TL-LSTM). Take the horizon t+10 and data missing rate 0.5 (all input factors) in the training and validating stages for instance, the optimal parameters of the LSTM and the TL-LSTM models were presented in Table 4.

The results pointed out that: the optimal number of neurons was 30 owing to the maximum NSE of 0.72 and the minimum RMSE of 0.43, while the optimal number of hidden layers was 3 owing to the maximal NSE value of 0.75 and the minimal RMSE value of 0.31 in the training stage as well as better indicator values in the validating stage regarding the LSTM model. Moreover, under the same data missing rate (= 0.5), the TL-LSTM model produced the smallest RMSE value and the largest NSE value as compared with other LSTM models. Hence, in the following comparison analysis, the parameters of each LSTM model and each TL-LSTM model included the maximal generation (G_{max}), number of neurons, number of hidden layers, learning rate and dropout probability, which were set as 1000, 30, 3, 0.001 and 0.5 respectively.

Table 4. Parameters of the LSTM and TL-LSTM models at horizon t+10 in the training and validating stages

Model	Data missing rate	Parameters					Training		Validating	
		G_{max}	Neurons	Hidden layer	Learning rate	Dropout probability	RMSE	NSE	RMSE	NSE
LSTM [a]	0.5	1000	20	1	0.001	0.5	0.65	0.64	0.68	0.62
			30				0.43	0.72	0.42	0.73
			40				0.58	0.67	0.61	0.64
			50				0.71	0.61	0.71	0.61
LSTM	0.5	1000	30	2	0.001	0.5	0.37	0.71	0.39	0.70
				3			0.31	0.75	0.29	0.76
				4			0.49	0.68	0.51	0.67
TL-LSTM [b]	0.5	1000	30	3	0.001	0.5	0.24	0.88	0.23	0.89

A value in bold indicated the optimal parameter. The data missing rate (= 0.5) denoted that all DO, NH_3-N and COD_{Cr} time series at 10 stations missed 50% of datasets and each stage (training, validating and testing) of the dataset was erased with the same percentage (i.e. 50%) during the establishment and application of the LSTM models. The computation result was the average result of 10 runs of each model. The value of RMSE was the average RMSE of water quality forecasts (DO, NH_3-N and COD_{Cr} values with standardization) while the value of NSE was the average NSE of water quality forecasts (DO, NH_3-N and COD_{Cr} values with standardization).

[a] LSTM denoted the long-short term memory model.

[b] TL-LSTM denoted the hybrid of transfer learning and long-short term memory model.

To further assess the impacts of different data missing rates (0 − 0.9) at different water quality stations (S1 – S10) on model performance, four sets of comparison experiments were designed to evaluate the accuracy of the two deterministic forecasting models.

Firstly, to investigate the performance of TL-LSTM model for different missing rates, the experiment scheme was set as the data missing rate (0 − 0.9, step = 0.1) and the incomplete target temporal sequence (S^{T}) from the Station S10. The reference temporal sequence (S^{R}) was identified as the sequence from the Station S8 while the highest correlation complete sequence to S^{R} was identified as the sequence (S^{RR}) from the Station S9. Take the horizons t+2 (2 hours), t+6 (6 hours) and t+10 (10 hours) for example, Figure 4 displayed the model performance of deterministic forecasts concerning water quality under different data missing rates in the testing stage.

The results revealed that: 1) the LSTM model produced an inferior performance for water quality forecasting under each data missing rate at each horizon; 2) the TL-LSTM model acquired the best performance not only in individual data missing rate but also at each horizon. It was easy to find that the TL-LSTM model created much higher values of NSE indicator but much smaller values of RMSE indicator under all data missing rates in the testing stages, in comparison to the LSTM model. For horizon t+10 and data missing rate (= 0.9), the improvement rates of RMSE and NSE indicators reached 24.7 % and 23.3 % respectively.

Previous researches (e.g. Lepot et al., 2017; Yang et al., 2017; Che et al., 2018; Tian et al., 2018) reported the maximum data missing rate that most of the methods could withstand was less than 0.3. The performance of forecast models became unsatisfied when the missing rate was large. The maximum data missing rate can be further extended according to forecast horizons, while its reliability and accuracy would be further decreased. The maximum data missing rate (= 0.5) that the proposed technique (TL-LSTM) could withstand was determined based on the forecast accuracy requirement (NSE > 0.75 and RMSE < 0.4) corresponding to the maximum horizon t+10, where this forecast accuracy could meet the practical needs of the users, decision-makers and stakeholders. Therefore, the data missing rate (= 0.5) was specified to assess the reliability and accuracy of the proposed approach in the following results.

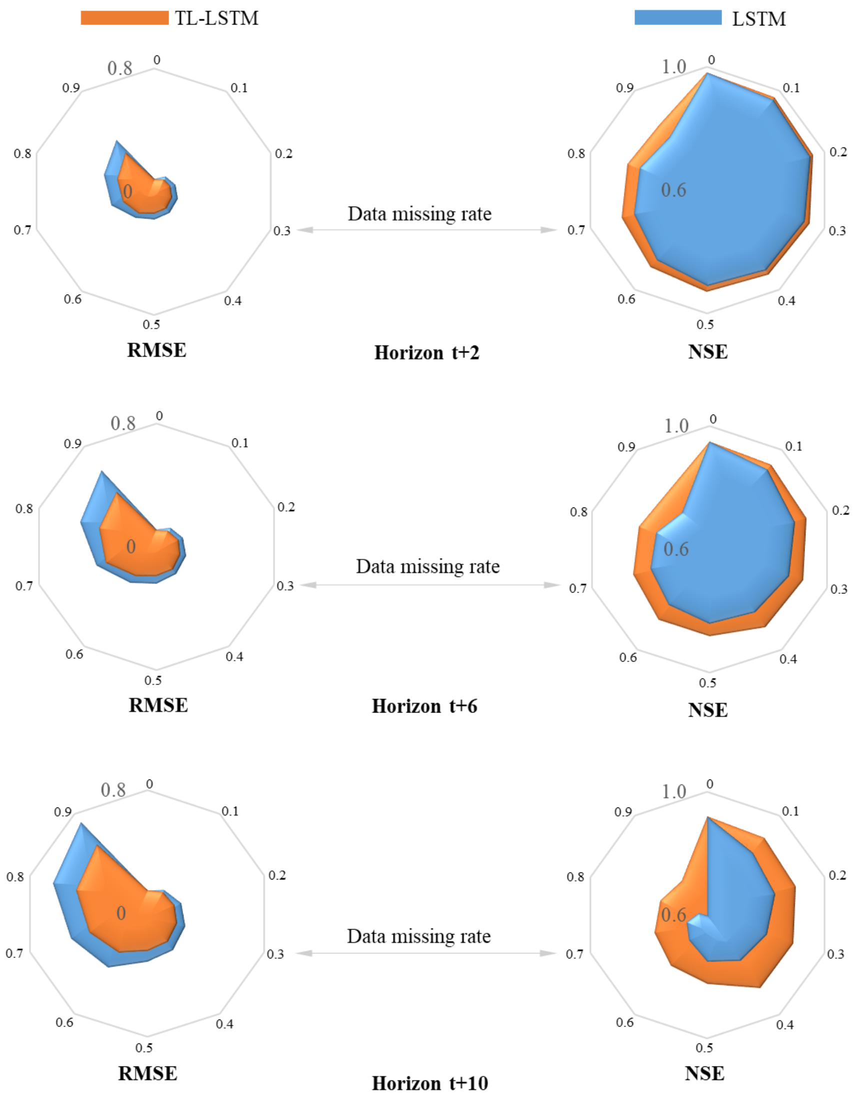

Fig. 4. Model performance of deterministic forecasts concerning water quality under different data missing rates (0 − 0.9, step = 0.1) at horizons t+2, t+6, t+10 at the Station S10 in the testing stage. In comparison analysis between TL-LSTM and LSTM models, the position of data missing in the initial data input always kept consistent in both models. That was to say, the position of data missing was randomly generated for the TL-LSTM model while the LSTM model had the same position of data missing with the TL-LSTM model. The computation result was the average result

of 10 runs of each model. The value of RMSE was the average RMSE of water quality forecasts (DO, NH_3-N and COD_{Cr} values with standardization) while the value of NSE was the average NSE of water quality forecasts (DO, NH_3-N and COD_{Cr} values with standardization).

Secondly, to investigate the performance of TL-LSTM model for different water quality stations, the experiment scheme was set as the incomplete target temporal sequence (S^{T}) varying from the Station S1 to Station S10 and the data constant missing rate (= 0.5). Take the horizons t+2, t+6 and t+10 for example, Figure 5 displayed the model performance of deterministic forecasts concerning water quality at different stations in the testing stages. The results indicated that the TL-LSTM model created much higher values of NSE indicator but much smaller values of RMSE indicator at all monitoring stations in the testing stages, as compared with the LSTM model. Take horizon t+10 and Station S7 for instance, the improvement rates of RMSE and NSE indicators achieved as much as 22.2 % and 12.5 % respectively. The results of Figure 5 demonstrated the technique had universally applicable to the data missing referring to different types of pollutions.

Thirdly, to investigate the impact of data missing in meteorological factors (e.g. precipitation and wind speed) and water quality factors (e.g. NH_3-N and COD_{Cr}) on the performance of LSTM models, the experiment scheme was set as the incomplete target temporal sequence (S^{T}) occurred at the Station S7 under the data constant missing rate (= 0.5). Take the horizons t+2, t+6 and t+10 for example, Table 5 presented the model performance of deterministic forecasts concerning water quality in the testing stages. The results pointed out that both LSTM models under the water quality data missing situation (Scenarios No. 3 and No. 4) produced much higher

RMSE values but much smaller NSE values than these under the meteorological data missing situation (Scenarios No. 1 and No. 2). In other words, the data missing in water quality factors had a more significant impact on the performance of LSTM models, as compared with the data missing in meteorological ones. The reason for causing such results was: if the forecasts for water quality (e.g. DO, NH_3-N, COD_{Cr}) were considered as the math function of water quality (9 factors) as well as meteorological (3 factors) variables, the autoregressive variables (e.g. NH_3-N and COD_{Cr}) had a more significant impact on the performance of forecast model, in comparison with the implicit exogenous variables (e.g. precipitation and wind speed). In other words, the modeler and forecaster should pay more attention to the raw data quality control and TL-LSTM model application when the data missing situation appeared in the autoregressive factors.

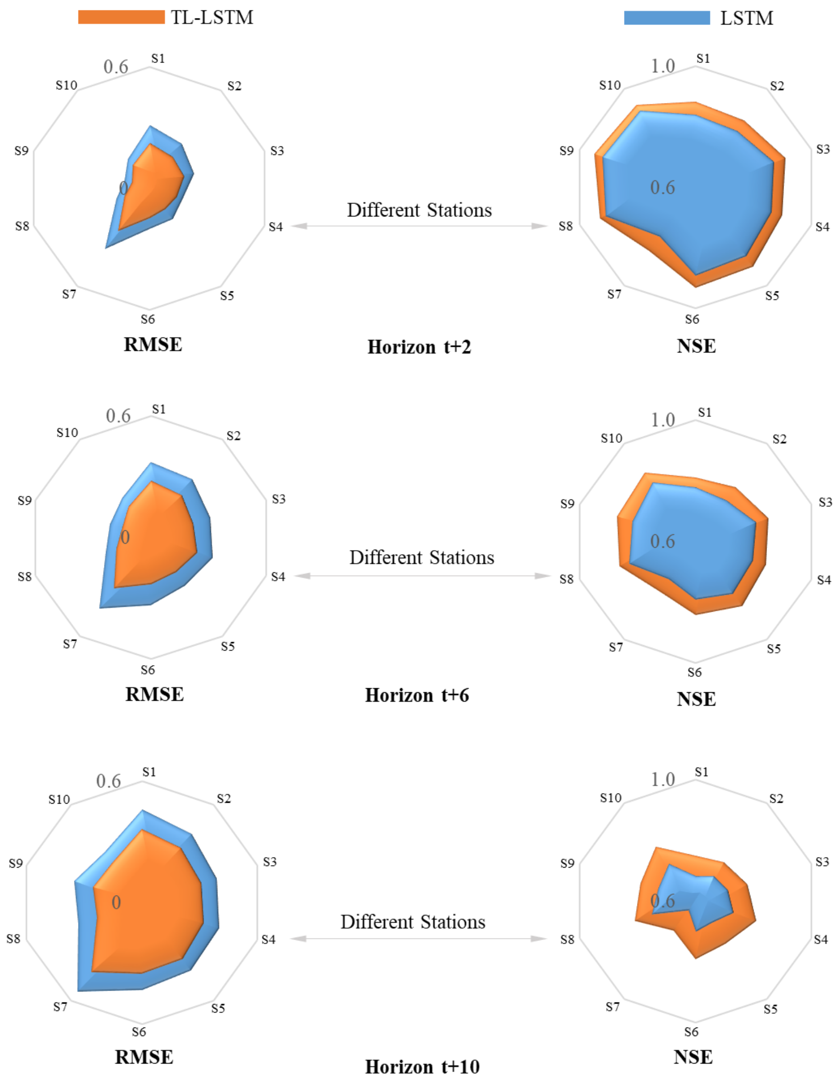

Fig. 5. Model performance of deterministic forecasts concerning water quality (DO, NH_3-N, and COD_{Cr}) under the data missing rate (= 0.5) at horizons t+2, t+6, t+10 at different stations (S1 – S10) in the testing stages. In comparison analysis between TL-LSTM and LSTM models, the position of data missing in the initial data input always kept consistent in both models. The computation result was the average result of 10 runs of each model. The value of RMSE was the average RMSE of water quality forecasts (DO, NH_3-N and COD_{Cr} values with standardization) while the value of NSE was the average NSE of water quality forecasts (DO, NH_3-N and COD_{Cr} values with standardization).

Table 5. Impact of data missing in meteorological and water quality factors on the performance of LSTM models at the Station S7 in the testing stage.

Scenario: missing factor	Model	Indicator	Horizon		
			t+2	*t*+6	*t*+10
No.1: Precipitation	TL-LSTM	RMSE	0.19	0.24	0.29
		NSE	0.87	0.82	0.77
	LSTM	RMSE	0.22	0.28	0.33

		NSE	0.83	0.78	0.73
No.2: Wind speed	TL-LSTM	RMSE	0.16	0.21	0.27
		NSE	0.92	0.86	0.81
	LSTM	RMSE	0.19	0.26	0.32
		NSE	0.89	0.84	0.79
No.3: NH_3-N	TL-LSTM	RMSE	0.22	0.27	0.32
		NSE	0.88	0.83	0.79
	LSTM	RMSE	0.31	0.38	0.48
		NSE	0.83	0.78	0.69
No.4: COD_{Cr}	TL-LSTM	RMSE	0.21	0.28	0.30
		NSE	0.90	0.85	0.81
	LSTM	RMSE	0.32	0.36	0.46
		NSE	0.84	0.80	0.71
No.5: All meteorological and water quality factors	TL-LSTM	RMSE	0.26	0.31	0.39
		NSE	0.86	0.81	0.76
	LSTM	RMSE	0.37	0.43	0.54
		NSE	0.80	0.75	0.64

The value of RMSE was the average RMSE of water quality forecasts (DO, NH_3-N and COD_{Cr} values with standardization) while the value of NSE was the average NSE of water quality forecasts (DO, NH_3-N and COD_{Cr} values with standardization).

Fourthly, the incomplete target temporal sequence (S^T) occurred at the Station S7 under the data constant missing rate (= 0.5) was specified to investigate the impact of data missing positions on the performance of LSTM models. Take the horizons t+2, t+6 and t+10 for example, Table 6 summarized the model performance of deterministic forecasts concerning water quality at the Station S7 in the testing stages. It was easy to find that both LSTM models under the peak data missing situation (Scenario No. 1) created the largest values of RMSE indicators but the smallest values of NSE indicators. Moreover, the loss of the trough data (Scenario No. 2) had the smallest impact on the performance of LSTM models. That was to say, the loss of the peak/trough data in the data sequence and the loss of the non-peak/non-trough data resulted in different forecast impacts on the performance of LSTM models (Scenarios: No.1 > No. 3 > No. 4 > No. 2). The results revealed that the modeler and forecaster should pay more attention to the raw data quality control and TL-LSTM model application if the data missing situation occurred in the peak datasets.

Table 6. Impact of data missing positions on the performance of LSTM models at the Station S7 in the testing stage

Scenario: data missing position	**Model**	**Indicator**	**Horizon**		
			***t*+2**	***t*+6**	***t*+10**
No.1: Peak data possessing the missing rate (0.5)	TL-LSTM	RMSE	0.22	0.29	0.37
		NSE	0.84	0.80	0.75
	LSTM	RMSE	0.26	0.38	0.49
		NSE	0.81	0.76	0.69
No.2: Trough data possessing the missing rate (0.5)	TL-LSTM	RMSE	0.17	0.22	0.27
		NSE	0.93	0.87	0.83
	LSTM	RMSE	0.21	0.25	0.30
		NSE	0.90	0.84	0.80
No.3: Peak and trough data possessing the missing rate (0.25) respectively	TL-LSTM	RMSE	0.20	0.25	0.31
		NSE	0.89	0.84	0.78
	LSTM	RMSE	0.24	0.28	0.35
		NSE	0.85	0.80	0.73
No.4: Non-peak and non-trough data possessing the missing rate (0.5)	TL-LSTM	RMSE	0.19	0.24	0.29
		NSE	0.91	0.85	0.80
	LSTM	RMSE	0.22	0.27	0.33
		NSE	0.88	0.82	0.77

The value of RMSE was the average RMSE of water quality forecasts (DO, NH_3-N and COD_{Cr} values with standardization) while the value of NSE was the average NSE of water quality forecasts (DO, NH_3-N and COD_{Cr} values with standardization).

In short, the TL-LSTM model created the best forecasting performance not only at different data missing situations (e.g. different water quality monitoring stations, data missing in meteorological and water quality factors, data missing positions) but also at each horizon. Furthermore, it was interesting to find that the TL-LSTM model could improve forecast accuracy and reliability (NSE values > 0.75 and RMSE values < 0.4) even under the most adverse data missing scenario.

To differentiate the capabilities of the LSTM and TL-LSTM models, three water pollution events at three monitoring stations (S1, S7, S10) were specified to validate both models by evaluating the goodness-of-fit of observed and forecasted datasets under the data missing rate (= 0.5) at horizon t+10 in the testing stages, as shown in Figure 6. It can be seen from Figure 6 that the TL-LSTM model was capable of forecasting well at horizon t+10 whereas the LSTM model had an apparent flatten prediction phenomenon as well as induced significantly large gaps between observed and forecasted data. It clearly revealed that the TL-LSTM model adequately followed the trails of water pollution events, effectively conquered the technical bottleneck of the flatten prediction, and created reliable as well as accurate multi-step-ahead forecasts of river water quality.

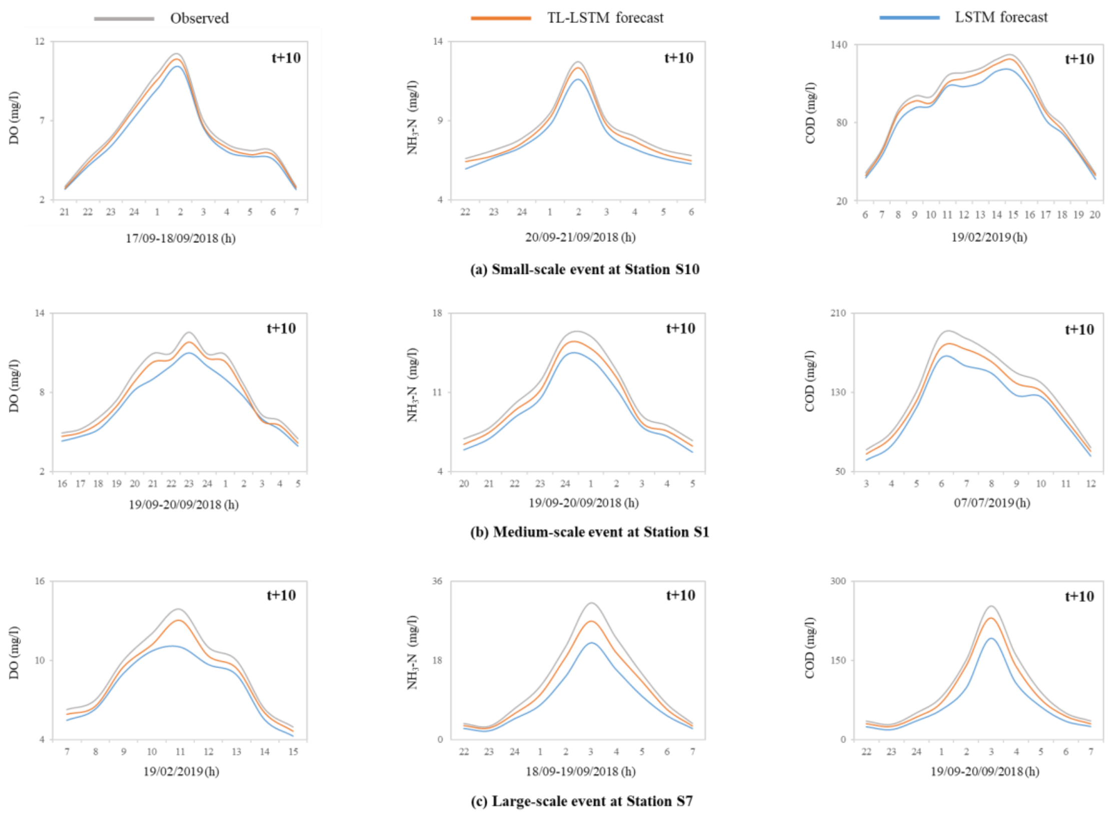

Fig. 6. Deterministic water quality forecast results (DO, NH_3-N and COD_{Cr}) of LSTM and TL-LSTM models under the data missing rate (= 0.5) at horizon t+10 in the testing stages at the Station S1 (East region), the Station S7 (North region) and the Station S10 (Center region) respectively. In comparison analysis between TL-LSTM and LSTM models, the position of data missing in the initial data input always kept consistent in both models. The computation result was the average result of 10 runs of

each model. The test event with small-scale (a) occurred at the Station S10. The test event with medium-scale (b) occurred at the Station S1. The test event with high-scale (c) occurred at the Station S7.

The results revealed that forecasts of the TL-LSTM model at horizons higher than t+2 were more excellent by using transfer learning algorithm under the input data missing circumstances. In other words, the transfer learning algorithm significantly improved water quality forecasts with different data missing rates by transferring model structure and parameters.

Although the forecast results of the TL-LSTM model exhibited well-off evidence of superior model performance as well as attained high confidence in deterministic forecasts, the values of water quality forecasting, regrettably, were easy to fall into systematic under-prediction for extreme water pollution events (Figure 6). Furthermore, apart from input data missing in meteorological and water quality factors, the uncertainties of parameters and the structure of LSTM models were the main reasons for inducing time-lag and flatten prediction phenomena that appeared in multi-step-ahead forecasts. Accordingly, the post-processing technique (MBUP) was further adopted for quantifying the predictive uncertainty of probabilistic water quality forecasts. The below subsection concentrated on the comparison analyzing between LSTM plus MBUP and TL-LSTM plus MBUP approaches for probabilistic water quality forecasting.

4.2 Probabilistic water quality forecasts

Several horizons (e.g. t+2, t+6, t+10) and water quality monitoring Stations (e.g. S1, S7, S10) were specified for validating the performance of probabilistic forecast techniques. The values of CR and RB corresponding to deterministic forecast model (LSTM or TL-LSTM) plus the post-processing technique MBUP with the data missing rate (= 0.5) were summarized in Table 7.

The results demonstrated that the TL-LSTM plus MBUP approach made better forecasting accuracy at all horizons and all stations whereas the LSTM plus MBUP approach performed inadequately at horizons larger than t+6 (the value of CR was lower than 89% and the value of RB was higher than 0.15). Take the Station S7 and horizon t+10 for instance, the TL-LSTM plus MBUP approach obtained the improvement rate of 7.4% for the CR indicator and the improvement rate of 21.1% for the RB indicator in the testing stage, in comparison to the LSTM plus MBUP approach. In other words, the TL-LSTM plus MBUP technique not only improved probabilistic forecast accuracy in a significant extent according to the high CR values denoting a narrow prediction but also mitigated the influence of the magnitude of pollutant concentration for the band-width of the prediction bounds according to the small RB values at the same time.

Table 7. Results of probabilistic water quality forecasting under the data missing rate (= 0.5) at horizons t+2, t+6, t+10 in the testing stages

Station	Model	Indicator	Horizon		
			***t*+2**	***t*+6**	***t*+10**
S1	TL-LSTM plus MBUP	CR(%)	96.17	92.39	88.62
		RB	0.09	0.18	0.25
	LSTM plus MBUP	CR(%)	95.22	90.04	83.56
		RB	0.12	0.22	0.30
S7	TL-LSTM plus MBUP	CR(%)	95.07	91.43	85.96
		RB	0.13	0.21	0.30
	LSTM plus MBUP	CR(%)	94.24	89.25	80.07
		RB	0.15	0.27	0.38

S10	TL-LSTM plus MBUP	CR(%)	98.63	93.17	89.66
		RB	0.08	0.15	0.22
	LSTM plus MBUP	CR(%)	97.48	91.24	84.39
		RB	0.10	0.21	0.26

The value of CR was the average CR of water quality forecasts (DO, NH_3-N and COD_{Cr}) while the value of RB was the average RB of water quality forecasts (DO, NH_3-N and COD_{Cr}).

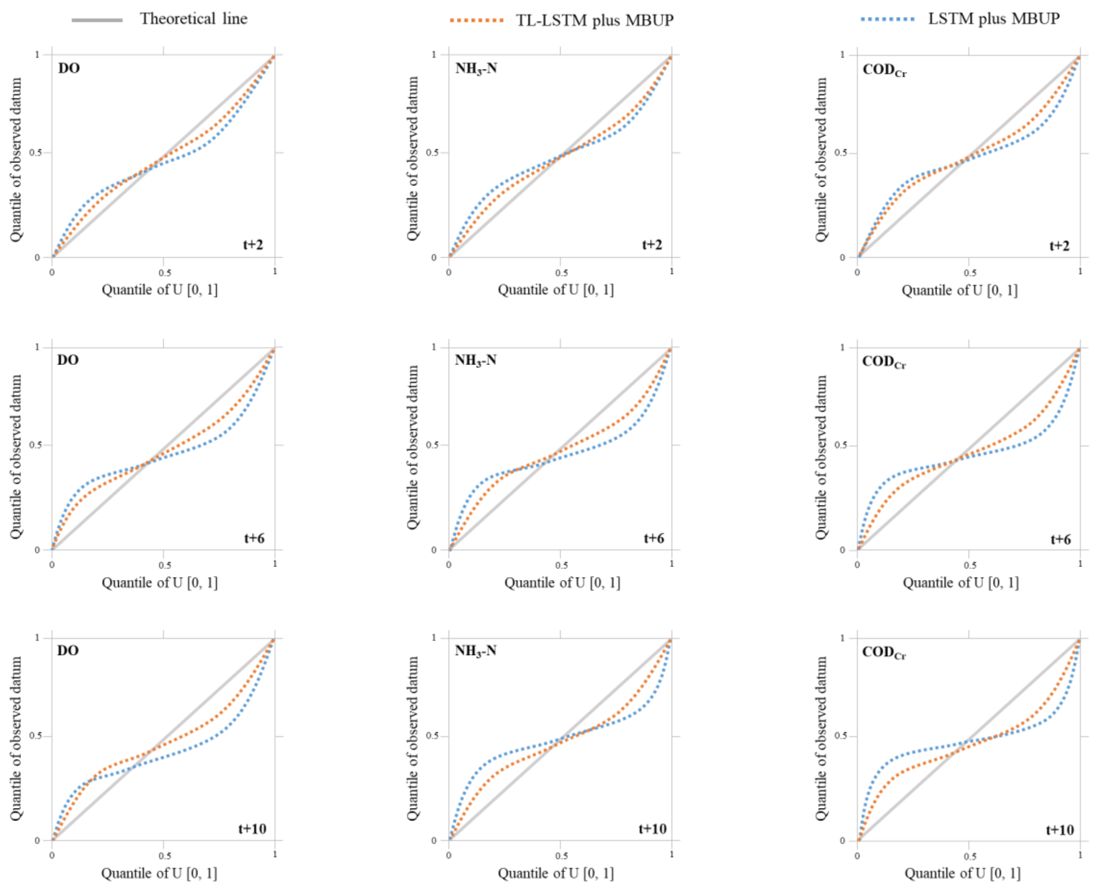

Fig. 7. Quantile-Quantile (QQ) plots of probabilistic water quality (DO, NH_3-N and COD_{Cr}) forecasts at the Station S7 under the data missing rate (= 0.5) at horizons t+2, t+6, t+10 in the testing stages.

Moreover, QQ plots were employed for evaluating the probabilistic forecasting reliability (LSTM plus MBUP & TL-LSTM plus MBUP). Figure 7 displayed the QQ plots for probabilistic water quality forecasting (e.g. Station S7) under the data missing rate (= 0.5) at horizons t+2, t+6, t+10 in the testing stages. It revealed that the QQ plot points created by the TL-LSTM plus MBUP approach were prone to be closer to the 1:1 line, as compared to that of the LSTM plus MBUP approach. In other words, the former (i.e. the TL-LSTM plus MBUP approach) acquired smaller bias as well as higher reliability than the latter (i.e. the LSTM plus MBUP approach).

The results pointed out that the TL-LSTM plus MBUP approach could provide effective support for quantifying predictive uncertainty because of the better goodness-of-fit between the predicted and the observed datasets. This finding demonstrated that the TL-LSTM plus MBUP approach executed better in terms of reliability assessment.

To distinctly distinguish the capabilities of probabilistic forecast models (LSTM plus MBUP & TL-LSTM plus MBUP) in the testing stages, the water pollution events (DO, NH_3-N and COD_{Cr}) at Station S7 were selected to test both models under the data missing rate (= 0.5) through evaluating if the water quality observations dropped within the interval of 90% prediction at horizon t+10 (Figure 8).

The results indicated that: (1) the 90% prediction intervals generated by TL-LSTM plus MBUP approach could cover the observed pollutant concentration peaks whereas the 90% prediction intervals generated by LSTM plus MBUP approach were still prone to systematically under-predictions, and (2) the TL-LSTM plus MBUP approach produced a narrower distribution of predictive water quality than that of the LSTM plus MBUP approach. The aim of probabilistic forecasting was to output the maximal sharpness for river water quality predictions, where the sharpness denoted the density of the predictive distributions. Hence, the hybrid of the TL-LSTM and MBUP approach was superior to the hybrid of the LSTM and MBUP approach for probabilistic river water quality forecasting. It was noticed that in Figure 8 the fluctuation range of water quality prediction became wider with the increase of corresponding water quality value. All fluctuation ranges in TL-LSTM plus MBUP (e.g. -0.9 mg/l ≤ Range of DO ≤ +1.8 mg/l) were significantly smaller than LSTM plus MBUP ones (e.g. -2.5 mg/l ≤ Range of DO ≤ +1.3 mg/l). Though the fluctuation ranges of COD_{Cr} values (-32.2 mg/l ≤ Range ≤ +19 mg/l) in the TL-LSTM plus MBUP approach were still wide, it was able to meet the needs of the practical application (forecast horizon up to 10 hours) of the model from the standpoint of relative error values (-12% ≤ Relative error ≤ +8%).

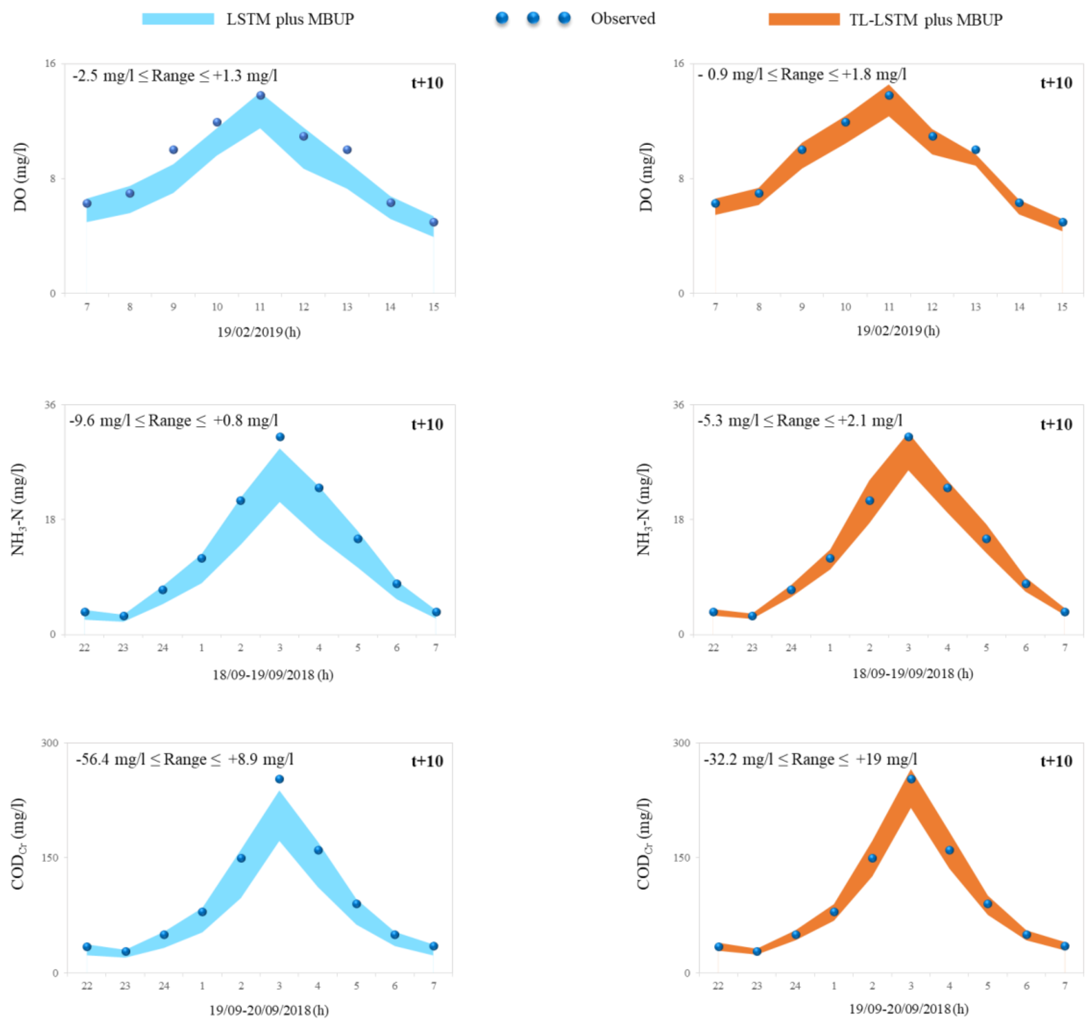

Fig. 8. Probabilistic water quality (DO, NH_3-N and COD_{Cr}) forecasts for Station S7

under the data missing rate (= 0.5) at horizon t+10 in the testing stages. The range is equal to the forecast minus the observation.

5. Conclusions and discussion

5.1 Conclusions

This study explored deep learning ANNs with MBUP approach for modelling probabilistic water quality forecasts. How to enhance the forecasting accuracy and reliability at water quality monitoring stations with plenty of missing data was fundamentally challenging. Moreover, the need for the probabilistic forecast instead of the deterministic forecast approach was attributed to the requirement of real-world operational forecasting and decreasing the stochasticity of water quality forecasts. Firstly, two deep learning ANNs (TL-LSTM and LSTM) were deployed to construct deterministic forecasting models for the local water quality values of the island in Shanghai City. The comparison of TL-LSTM as well as LSTM models was to demonstrate the contributions of the transfer learning algorithm on more accurate deterministic forecasts. Then, the exploration of the post-processing technique (MBUP) was implemented for transforming the deterministic forecasting (i.e. LSTM models) into the probabilistic forecasting. The contribution of the MBUP approach relied upon extracting the complex nonlinear multivariate (≥3) correlation between observations and forecasts as well as upon decreasing the predictive uncertainty of river water quality forecasts.

Both two deterministic models utilized for forecasting the regional water quality (DO, NH_3-N and COD_{Cr}) series of the island in Shanghai City illustrated that the

TL-LSTM model remarkably performed better than the relative LSTM model for the three cases (i.e. training, validation and testing) at various horizons as well as different monitoring stations. It indicated that the TL-LSTM model could make highly more accurate forecasts for the river water quality series at long lead times (future 10 hours) and could effectively overcome flatten prediction bottlenecks in comparison to the LSTM model. However, the TL-LSTM model still undergone the technical difficulty of flatten-predicting the peaks of river water quality.

The MBUP would explicitly extract the complex nonlinear multivariate correlation between observations and forecasts as well as would alleviate the stochasticity of probabilistic river water quality forecasting. The comparison analysis demonstrated that the TL-LSTM plus MBUP approach was substantially preferable to the LSTM plus MBUP one, according to the values of CR and RB indicators as well as the 90% prediction intervals. The hybrid of TL-LSTM plus MBUP technique succeeded in obtaining excellent results of probabilistic river water quality forecasting would be attributed to the first key strategy: the incorporation of the transfer learning algorithm into ANNs for reinforcing the model structure and parameters transferring to overcome input data missing drawback, and the second core strategy: the adequate extraction of the nonlinear multivariate correlation information between model forecasts and observations for lowering the predictive uncertainty through the multivariate Bayesian uncertainty processing technique.

5.2 Discussion

From the standpoint of water pollution mechanisms, the point source pollution processes associated with industry and urban domestic sewage conditions (e.g. Station S10) made a too slight difference in forecasting accuracy between the LSTM and the TL-LSTM models, whereas the nonpoint source pollution processes associated with agricultural activities (e.g. Station S7) made a significant difference in forecasting accuracy between the LSTM and the TL-LSTM models. The island in Shanghai City has experienced rapid development, and the local water quality of the island has constantly undergone interactions with intensive industrial sewages, urban and agricultural activities. A high water pollution event was commonly driven by the processes of nonpoint source pollutions either associating with the local transformation of the aged fertilizer/aquatic feed or associating with the secondary transportation of eutrophication pollutants. A water pollution event corresponding to the point source pollution processes was prone to associate with the primary sewage discharges as well as regional weather conditions. The LSTM model made a better forecasting accuracy at the Station S10 than at the Station S1 and Station S7. Nevertheless, the TL-LSTM model gained better improvement rates of RMSE and NSE at Station S1 and Station S7 than at the Station S10. The TL-LSTM model not only attained higher improvement rates for forecasting accuracy at water quality Station S1 and Station S7 (nonpoint source pollution processes) but also executed as good as the performance of the LSTM model at the water quality Station S10 (point source pollution processes).

The imitation of real-time evolution in water quality was attributed to twofold: First, the data collection from water quality stations was based on real-time processing (hourly collecting). Second, it was worth noting that the computational time (less than 2 minutes) of the proposed approach was extremely short and therefore it could be applied with success to real-time water quality forecasting. From the standpoint of science forward, this study no only initiated effective research on probabilistic water quality forecasts under data missing situation that was beneficial to water quality warning and prediction but also contributed to innovating artificial intelligence-based solutions to river environmental management in the interest of green economy development. Following this study that constructed a framework to conquer the under-prediction phenomena and quantify the uncertainty of probabilistic water quality forecasting induced by input data missing, several subsequent studies can be conducted, for instance, incorporating extreme learning mechanisms into this framework to predict water quality better once an extreme phenomenon happens. Additionally, future research would explore the hybrid of deep learning and probabilistic post-processing techniques from the small and medium spatial scale (a local or regional city) of time series to large spatial scale (country or global) ones.

Appendix A

LSTM model structure

The LSTM model structure consists of six components: input block, three gates, self-looped cell and output block. The following steps illustrate how the LSTM model is updated at every time step *t*.

Step 1: The input block is employed to create memory information ($\hat{C}_t$) at the current time t by jointing the output state (h_{t-1}) at the previous time t-1 with the model input (x_t) at the current time t.

$$\hat{C}_t = \tanh(W_c x_t + U_c h_{t-1} + b_c) \quad (1)$$

where $\tanh(\cdot)$ is a hyperbolic tangent function. W_c is the weight for the input of the current state in the input block. U_c is the weight for the output of the previous state in the input block. b_c is the bias in the input block at the current state.

Step 2: The input gate (i_t) is conducted to calculate how much information to allocate to the current cell state through learning from the output state (h_{t-1}) at the previous time t-1 and the model input (x_t) at the current time t.

$$i_t = \theta(W_i x_t + U_i h_{t-1} + b_i) \quad (2)$$

where $\theta(\cdot)$ is a sigmoid transfer function. W_i is the weight for the input of the current state in the input gate. U_i is the weight for the output of the previous state in the input gate. b_i is the bias in the input gate at the current state.

Step 3: The forget gate (f_t) is conducted to quantify how much information to delete from the current cell state through learning from the output state (h_{t-1}) at the previous time t-1 and the model input (x_t) at the current time t.

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

where W_f is the weight for the input of the current state in the forget gate. U_f is the weight for the output of the previous state in the forget gate. b_f is the bias in the forget gate at the current state.

Step 4: The self-looped cell (C_t) is used to update the previous self-looped cell state (C_{t-1}) through integrating the information of the input and forget gates with the current input block ($\hat{C}_t$).

$$C_t = i_t \cdot \hat{C}_t + f_t \cdot C_{t-1} \quad (3)$$

Step 5: The output gate (o_t) is conducted to quantify the output of the self-recurrent cell. The tanh function is also adopted to transform the self-looped cell state () to confirm that the value lies in the interval of [-1, 1] and the transformed results would be multiplied by the value of the output gate, which creates the current output state (h_t).

$$o_t = \theta(W_o x_t + U_o h_{t-1} + V_0 C_t + b_0) \quad (5a)$$

$$h_t = o_t \cdot \tanh(C_t) \quad (5b)$$

where W_o is the weight for the input of the current state in the output gate. U_o is the weight for the output of the previous state in the output gate. V_o is the weight for the self-recurrent cell state in the output gate. b_o is the bias in the output gate at the current state.

Step 6: The output block is employed to calculate the output of the LSTM model, which is regarded as the algebraic sum of the output gate.

$$\hat{y}_t = W_y h_t + b_y \quad (6)$$

where $\hat{y}_t$ is the output of the LSTM model. W_y is the weight for the current output state. b_y is the bias in the output block at the current state.

Appendix B

General implementation procedure of transfer learning-based LSTM model

Step 1: Data pattern transfer. After implementations of data collection, cleaning and normalization, a RTS is selected according to the most statistic similarity between

TTS and potential RTSs. The transfer learning algorithm is employed owing to the statistical similarity between RTS and TTS. The Kendall tau coefficient (Maidment et al., 1993) is used to identify the highest correlation between TTS and RTS. The computation equations for selecting RTS are described as follows.

$$\mathrm{S^R} = \max\left|Tau\left(\mathrm{S^T}, \mathrm{S}^i\right)\right|,\ \ 1 \leq i \leq \mathrm{K},\ \ i \neq \mathrm{T} \tag{7}$$

$$Tau\left(\mathrm{S^T}, \mathrm{S}^i\right) = Tau\left([\mathrm{S_1^T}, \mathrm{S_3^T}], \left[\mathrm{S}_1^i, \mathrm{S}_3^i\right]\right) \tag{8}$$

$$Tau = \frac{\mathrm{n_c - n_d}}{\frac{1}{2}\mathrm{n(n-1)}},\ \ -1 \leq Tau \leq 1 \tag{9}$$

where $\mathrm{S^R}$ is the selected RTS. $\mathrm{S^T}$ is the incomplete TTS, and take the incomplete TTS with one missing segment $\mathrm{S^T} = [\mathrm{S_1^T}, \mathrm{S_2^M}, \mathrm{S_3^T}]$ for example, $\mathrm{S_2^M}$ is the missing segment, $\mathrm{S_1^T}$ and $\mathrm{S_3^T}$ are the complete segments. If $\mathrm{S_2^M}$ is at the beginning or the end of $\mathrm{S^T}$, $\mathrm{S_1^T}$ or $\mathrm{S_3^T}$ would be empty dataset. S^i is the complete sequence (i.e. potential RTS) at the *i*th monitoring station, $\mathrm{S}^i = \left[\mathrm{S}_1^i, \mathrm{S}_2^i, \mathrm{S}_3^i\right]$, $1 \leq i \leq \mathrm{N}$ and $i \neq \mathrm{T}$, S_1^i, S_2^i and S_3^i are three segments of complete sequence S^i corresponding to three segments in $\mathrm{S^T}$, where K is the number of monitoring stations. *n* is the number of dataset. $\mathrm{N_C}$ and $\mathrm{N_D}$ are the number of concordant pairs and discordant pairs in two datasets (TTS & RTS) respectively. In this step, two RTSs ($\mathrm{S^R}$ & $\mathrm{S^{RR}}$) would be selected for training TL-LSTM model. $\mathrm{S^R}$ is the highest correlation complete sequence of $\mathrm{S^T}$ while $\mathrm{S^{RR}}$ is the highest correlation complete sequence of $\mathrm{S^R}$.

Step 2: Model structure and parameters transfer. A reference TL-LSTM model ($\mathrm{Model_R}$) would be trained using the RTS while validates and tests the model (structure and parameters) using the TTS. In the training stage, the input data of $\mathrm{Model_R}$ is $\left\{\left[\mathrm{S}_{t-q}^{\mathrm{R}}, \mathrm{S}_{t-q+1}^{\mathrm{R}}, \cdots, \mathrm{S}_{t-1}^{\mathrm{R}}, \mathrm{S}_t^{\mathrm{RR}}\right] \rightarrow [\mathrm{S}_{t+m}^{\mathrm{R}}]\right\}$ instead of

$\{[S^R_{t-q}, S^R_{t-q+1}, \cdots, S^R_{t-1}, S^R_t] \rightarrow [S^R_{t+m}]\}$, where q is the time-lags of input variables, m is the forecast horizon and m = 1, …, M. After the $\mathrm{Model_R}$ is given, the model structure is frozen while in the validating stage the model ($\mathrm{Model_R}$) parameters are fine-tuned using the input data $\{[S^T_{t-p}, S^T_{t-p+1}, \cdots, S^T_{t-1}, S^R_t] \rightarrow [S^T_{t+m}]\}$ to create the target TL-LSTM model ($\mathrm{Model_T}$) in the validating stage. The $\mathrm{Model_T}$ can maintain the model structure $\mathrm{Model_R}$ (structure transfer), fine-tune the model parameters (parameters transfer) based on the data pattern transfer $\{S^{RR} \rightarrow S^R\}$ and $\{S^R \rightarrow S^T\}$ so as to reduce the flatten forecasts and improve the model transferability.

Step 3: Iteration: the stopping rules are employed to terminate the computation process. If the value of the objective function would not decline in the next 100 consecutive iterations, the accuracy of the ANN model would no longer be increased, which causes the calculation to stop. Once the maximum of iterations is attained, the training and validating processes stop. Otherwise, update the iteration, and repeat Step 2. The given $\mathrm{Model_T}$ can be used for multi-step-ahead forecasts under missing data conditions in the testing stage.

Output: the optimized structure (multi-output and number of hidden layers) and parameters (the learning rate, the weight vector and the bias vector) of the TL-LSTM model would be saved and the TL-LSTM model would create the deterministic water quality forecasts for different monitoring stations.

Acknowledgments

This work was supported by the National Key Research and Development Program of

China (2018YFC0407904) and the Research Council of Norway (FRINATEK Project 274310). The authors would like to thank the Editors and anonymous Reviewers for their constructive comments that greatly contributed to improving the manuscript.

References

Aguilera, P. A., Frenich, A. G., Torres, J. A., Castro, H., Vidal, J. M., & Canton, M., 2001. Application of the Kohonen neural network in coastal water management: methodological development for the assessment and prediction of water quality. Water Res., 35(17), 4053-4062.

Akbari Asanjan, A., Yang, T., Hsu, K., Sorooshian, S., Lin, J., & Peng, Q., 2018. Short‐term precipitation forecast based on the PERSIANN system and LSTM recurrent neural networks. J. Geophys. Res.: Atmos., 123(22), 12-543.

Arhonditsis, G. B., Neumann, A., Shimoda, Y., Javed, A., Blukacz-Richards, A., & Mugalingam, S., 2019. When can we declare a success? A Bayesian framework to assess the recovery rate of impaired freshwater ecosystems. Environ. Int., 130, 104821.

Barzegar, R., Moghaddam, A. A., Adamowski, J., & Ozga-Zielinski, B., 2018. Multi-step water quality forecasting using a boosting ensemble multi-wavelet extreme learning machine model. Stoc. Environ. Res. Risk Assess., 32(3), 799-813.

Borsuk, M. E., Stow, C. A., & Reckhow, K. H., 2002. Predicting the Frequency of Water Quality Standard Violations: A Probabilistic Approach for TMDL Development. Environ. Sci. Technol., 36(10), 2109-2115.

Cannon, A. J., 2011. Quantile regression neural networks: implementation in R and application to precipitation downscaling. Comput. Geosci., 37(9), 1277-1284.

Camacho, R. A., Martin, J. L., Wool, T., & Singh, V. P., 2018. A framework for uncertainty and risk analysis in total maximum daily load applications. Environ. Modell. Software, 101, 218-235.

Chang, F. J., & Tsai, M. J., 2016. A nonlinear spatio-temporal lumping of radar rainfall for modeling multi-step-ahead inflow forecasts by data-driven techniques. J. Hydrol., 535, 256-269.

Che, Z., Purushotham, S., Cho, K., Sontag, D., & Liu, Y., 2018. Recurrent neural networks for multivariate time series with missing values. Sci. Rep., 8(1), 6085.

Ding, Z., Mei, G., Cuomo, S., Li, Y., & Xu, N., 2018. Comparison of estimating missing values in IoT time series data using different interpolation algorithms. Int. Journal Parallel Program., 1-15.

Ekeu-wei, I., Blackburn, G., & Pedruco, P., 2018. Infilling missing data in hydrology: solutions using satellite radar altimetry and multiple imputation for data-sparse regions. Water, 10(10), 1483.

Gerhard, W. A., & Gunsch, C. K., 2019. Metabarcoding and machine learning analysis of environmental DNA in ballast water arriving to hub ports. Environ. Int., 124, 312–319.
Fofonoff, N.P. and Millard, R.C., 1983. Algorithms for computation of fundamental properties of seawater. Unesco Technical Papers in Marine Science, 44, 53 pp.
Fu, B., Merritt, W. S., Croke, B. F., Weber, T., & Jakeman, A. J., 2018. A review of catchment-scale water quality and erosion models and a synthesis of future prospects. Environ. Modell. Software,, 114, 75-97.
Galelli, S., Humphrey, G. B., Maier, H. R., Castelletti, A., Dandy, G. C., & Gibbs, M. S., 2014. An evaluation framework for input variable selection algorithms for environmental data-driven models. Environ. Model. Soft. 62, 33-51.
Gao, T., & Wang, H., 2017. Testing backpropagation neural network approach in interpolating missing daily precipitation. Water, Air, Soil Pollut., 228(10), 404.
García-Alba, J., Bárcena, J. F., Ugarteburu, C., & García, A., 2019. Artificial neural networks as emulators of process-based models to analyse bathing water quality in estuaries. Water Res., 150, 283-295.
Gallego, A. J., Gil, P., Pertusa, A., & Fisher, R. B., 2019. Semantic Segmentation of SLAR Imagery with Convolutional LSTM Selectional AutoEncoders. Remote Sens., 11(12), 1402.
Guo, D., Lintern, A., Webb, J. A., Ryu, D., Liu, S., Bende - Michl, U., & Western, A. W., 2019. Key factors affecting temporal variability in stream water quality. Water Resour. Res., 55(1), 112-129.
Gupta, J., Paul, S., & Ghosh, A., 2019. A Novel Transfer Learning-Based Missing Value Imputation on Discipline Diverse Real Test Datasets—A Comparative Study with Different Machine Learning Algorithms. In Emerging Technologies in Data Mining and Information Security (pp. 815-826). Springer, Singapore.
Gneiting, T., 2008. Probabilistic forecasting. Journal of the Royal Statistical Society: Series A (Statistics in Society), 171(2), 319-321.
Helbich, M., Yao, Y., Liu, Y., Zhang, J., Liu, P., & Wang, R., 2019. Using deep learning to examine street view green and blue spaces and their associations with geriatric depression in Beijing, China. Environ. Int., 126, 107–117.
Herr, H. D., & Krzysztofowicz, R., 2015. Ensemble Bayesian forecasting system Part I: Theory and algorithms. J. Hydrol., 524, 789-802.
Hochreiter, S., & Schmidhuber, J., 1997. Long short-term memory. Neural Comput., 9(8), 1735-1780.
Isiyaka, H. A., Mustapha, A., Juahir, H., & Phil-Eze, P., 2019. Water quality modelling using artificial neural network and multivariate statistical techniques. Model. Earth Syst. Environ., 5(2), 583-593.
Jardim W.F., 2014. Medicao e interpretacao de valores do potecial redox (E_H) em matrizes ambientais, Quim. Nova., 37(7), 1233-1235.
Jiang, G., Keller, J., Bond, P. L., & Yuan, Z., 2016. Predicting concrete corrosion of sewers using artificial neural network. Water Res., 92, 52-60.
Kao, I.-F., Zhou, Y., Chang, L.-C., & Chang, F.-J., 2020. Exploring a Long Short-Term Memory based Encoder-Decoder Framework for Multi-Step-Ahead

Flood Forecasting. J. Hydrol., 124631.
Krapu, C., & Borsuk, M., 2019. Probabilistic programming: A review for environmental modellers. Environ. Model. Softw., 114, 40-48.
Krzysztofowicz, R., 1999. Bayesian theory of probabilistic forecasting via deterministic hydrologic model. Water Resour. Res., 35(9), 2739-2750.
Krzysztofowicz, R., 2002. Bayesian system for probabilistic river stage forecasting. J. Hydrol., 268(1-4), 16-40.
Krzysztofowicz, R., & Maranzano, C. J., 2004. Hydrologic uncertainty processor for probabilistic stage transition forecasting. J. Hydrol., 293(1-4), 57-73.
Kim, J. W., & Pachepsky, Y. A., 2010. Reconstructing missing daily precipitation data using regression trees and artificial neural networks for SWAT streamflow simulation. J. Hydrol., 394(3-4), 305-314.
Lepot, M., Aubin, J. B., & Clemens, F., 2017. Interpolation in time series: An introductive overview of existing methods, their performance criteria and uncertainty assessment. Water, 9(10), 796.
Liang, S., Jia, H., Xu, C., Xu, T., & Melching, C., 2016. A Bayesian approach for evaluation of the effect of water quality model parameter uncertainty on TMDLs: a case study of Miyun Reservoir. Sci. Total Environ., 560, 44-54.
Liang, Z., Zou, R., Chen, X., Ren, T., Su, H., & Liu, Y., 2019. Simulate the forecast capacity of a complicated water quality model using the long short-term memory approach. J. Hydrol., 124432.
Libera, D. A., & Sankarasubramanian, A., 2018. Multivariate bias corrections of mechanistic water quality model predictions. J. Hydrol., 564, 529-541.
Liu, C., Wang, Q., Zou, C., Hayashi, Y., & Yasunari, T., 2015. Recent trends in nitrogen flows with urbanization in the Shanghai megacity and the effects on the water environment. Environ. Sci. Pollut. Res., 22(5), 3431-3440.
Maidment, D., Stedinger, J., Vogel, R., Foufoulageogious, E., Pilgrim, D., & Cordery, I., et al., 1993. Handbook of hydrology, 24, 227-229.
Mok, K. M., Yuen, K. V., Hoi, K. I., Chao, K. M., & Lopes, D., 2018. Predicting ground-level ozone concentrations by adaptive Bayesian model averaging of statistical seasonal models. Stoc. Environ. Res. Risk Assess., 32(5), 1283-1297.
Moreno-Rodenas, A. M., Tscheikner-Gratl, F., Langeveld, J. G., & Clemens, F. H., 2019. Uncertainty analysis in a large-scale water quality integrated catchment modelling study. Water Res., 158, 46-60.
Mian, H. R., Hu, G., Hewage, K., Rodriguez, M. J., & Sadiq, R., 2018. Prioritization of unregulated disinfection by-products in drinking water distribution systems for human health risk mitigation: A critical review. Water Res.. 147, 112-131.
Newhart, K. B., Holloway, R. W., Hering, A. S., & Cath, T. Y., 2019. Data-driven performance analyses of wastewater treatment plants: A review. Water Res., 157, 498-513.
Olsen, R. L., Chappell, R. W., & Loftis, J. C., 2012. Water quality sample collection, data treatment and results presentation for principal components analysis–literature review and Illinois River watershed case study. Water Res., 46(9), 3110-3122.

Pan, S. J., & Yang, Q., 2009. A survey on transfer learning. IEEE Transactions on Knowledge and Data Engineering, 22(10), 1345-1359.

Pearce, A. R., Rizzo, D. M., Watzin, M. C., & Druschel, G. K., 2013. Unraveling Associations between Cyanobacteria Blooms and In-Lake Environmental Conditions in Missisquoi Bay, Lake Champlain, USA, Using a Modified Self-Organizing Map. Environ. Sci. Technol., 47(24), 14267–14274.

Perelman, L., Arad, J., Housh, M., & Ostfeld, A., 2012. Event Detection in Water Distribution Systems from Multivariate Water Quality Time Series. Environ. Sci. Technol., 46(15), 8212–8219.

Peleato, N. M., Legge, R. L., & Andrews, R. C., 2018. Neural networks for dimensionality reduction of fluorescence spectra and prediction of drinking water disinfection by-products. Water Res., 136, 84-94.

Rajakumar, A. G., Mohan Kumar, M. S., Amrutur, B., & Kapelan, Z., 2019. Real-time water quality modeling with ensemble Kalman filter for state and parameter estimation in water distribution networks. J. Water Resour. Plann. Manage., 145(11), 04019049.

Regina P. Stefan P., 2019. Using artificial intelligence to forecast water oxidation catalysts. Environ. Sci. Technol., 9, 8383-8387.

Schmidhuber J., 2015. Deep learning in neural networks: an overview. Neural Networks, 61, 85-117.

Sharma, A., 2000. Seasonal to interannual rainfall probabilistic forecasts for improved water supply management: Part 1-A strategy for system predictor identification. . J. Hydrol. 239(1-4), 232-239.

Shrestha, S., & Kazama, F., 2007. Assessment of surface water quality using multivariate statistical techniques: A case study of the Fuji river basin, Japan. Environ. Model. Softw., 22(4), 464-475.

Tencaliec, P., Favre, A. C., Prieur, C., & Mathevet, T., 2015. Reconstruction of missing daily streamflow data using dynamic regression models. Water Resour. Res., 51(12), 9447-94.

Tian, Y., Zhang, K., Li, J., Lin, X., & Yang, B., 2018. LSTM-based traffic flow prediction with missing data. Neurocomputing, 318, 297-305.

Tiyasha, Minh Tung, T., Mundher Yaseen, Z., 2020. A survey on river water quality modelling using artificial intelligence models: 2000-2020. J. Hydrol., 585, 124670.

Wallace, J., Champagne, P., & Hall, G., 2016. Multivariate statistical analysis of water chemistry conditions in three wastewater stabilization ponds with algae blooms and pH fluctuations. Water Res., 96, 155-165.

Xiong, L., & O'Connor, K. M., 2008. An empirical method to improve the prediction limits of the GLUE methodology in rainfall-runoff modeling. J. Hydrol., 349(1-2), 115-124.

Yann L.C., Yoshua B., Geoffrey H., 2015. Deep Learning. Nature, 521, 436-444.

Yang, J. H., Cheng, C. H., & Chan, C. P., 2017. A time-series water level forecasting model based on imputation and variable selection method. Comput. Intelli. Neurosci., 9, 8734214.

Yang, L., Zhao, X., Peng, S., & Li, X., 2016. Water quality assessment analysis by using combination of Bayesian and genetic algorithm approach in an urban lake, China. Ecol. Modell., 339, 77-88.

Yaseen, Z. M., Sulaiman, S. O., Deo, R. C., & Chau, K.-W., 2019. An enhanced extreme learning machine model for river flow forecasting: State-of-the-art, practical applications in water resource engineering area and future research direction. J. Hydrol., 569, 387-408.

Yi, J., Wen, Z., Tao, J., Ni, H., & Liu, B., 2018. CTC Regularized Model Adaptation for Improving LSTM RNN Based Multi-Accent Mandarin Speech Recognition. J. Signal Process. Syst., 90(7), 985-997.

Zhao, J., Lin, L., Yang, K., Liu, Q., & Qian, G., 2015. Influences of land use on water quality in a reticular river network area: A case study in Shanghai, China. Landscape Urban Plan., 137, 20-29.

Zhao, J., Qu, H., Zhao, J., & Jiang, D., 2018. Towards traffic matrix prediction with LSTM recurrent neural networks. Electron. Lett., 54(9), 566-568.

Zhang, W., Li, T., & Dai, M., 2015. Uncertainty assessment of water quality modeling for a small-scale urban catchment using the GLUE methodology: a case study in Shanghai, China. Environ. Sci. Pollut. Res. 22(12), 9241-9249.

Zhang, J., Qiu, H., Li, X., Niu, J., Nevers, M. B., Hu, X., & Phanikumar, M. S., 2018. Real-Time Nowcasting of Microbiological Water Quality at Recreational Beaches: A Wavelet and Artificial Neural Network-Based Hybrid Modeling Approach. Environ. Sci. Technol., 52(15), 8446–8455.

Zhou, Y., Chang, F. J., Chang, L. C., Kao, I. F., & Wang, Y. S., 2019a. Explore a deep learning multi-output neural network for regional multi-step-ahead air quality forecasts. J. Clean. Prod., 209, 134-145.

Zhou, Y., Chang, F. J., Chang, L. C., Kao, I. F., Wang, Y. S., & Kang, C. C., 2019b. Multi-output support vector machine for regional multi-step-ahead $PM_{2.5}$ forecasting. Sci. Total Environ., 651, 230-240.

Zhou, Y., Guo, S., Xu, C. Y., Chang, F. J., & Yin, J., 2020. Improving the reliability of probabilistic multi-step-ahead flood forecasting by fusing unscented Kalman filter with recurrent neural network. Water, 12(2), 578.

Abstract

Quantifying the uncertainty of probabilistic water quality forecasting induced by missing input data is fundamentally challenging. This study introduced a novel methodology for probabilistic water quality forecasting conditional on point forecasts. A Multivariate Bayesian Uncertainty Processor (MBUP) was adopted to probabilistically model the relationship between the point forecasts made by a deep learning artificial neural network (ANN) and their corresponding observed water quality. The methodology was tested using hourly water quality series at an island of Shanghai City in China. The novelties relied upon: firstly, the use of a transfer learning algorithm to overcome flatten- and under-prediction bottlenecks of river water quality raised in artificial neural networks, and secondly, the use of the MBUP to capture the dependence structure between observations and forecasts. Two deep learning ANNs were used to make the point forecasts. Then the MBUP approach driven by the point forecasts demonstrated its competency in improving the accuracy of probabilistic water quality forecasts significantly, where predictive distributions encountered in multi-step-ahead water quality forecasts were effectively reduced to small ranges. The results demonstrated that the deep learning plus the post-processing approach suitably extracted the complex dependence structure between the model's output and observed water quality so that model reliability (Containing Ratio > 85% and average Relative Band-width < 0.25) as well as forecast accuracy (Nash-Sutcliffe Efficiency coefficient > 0.8 and Root-Mean-Square-Error < 0.4 mg/l) for future horizons from 1 hour up to 10 hours were significantly improved, even if the input data missing rate reaches 50%.

Highlights

- For the first time a TL-LSTM model is proposed to model water quality forecasts.
- Deep learning plus post-processing enhances probabilistic water quality forecasts.
- Deep learning improves accuracy of deterministic water quality forecasts.
- Transfer learning overcomes flatten/under-predictions induced by missing input data.

Figures

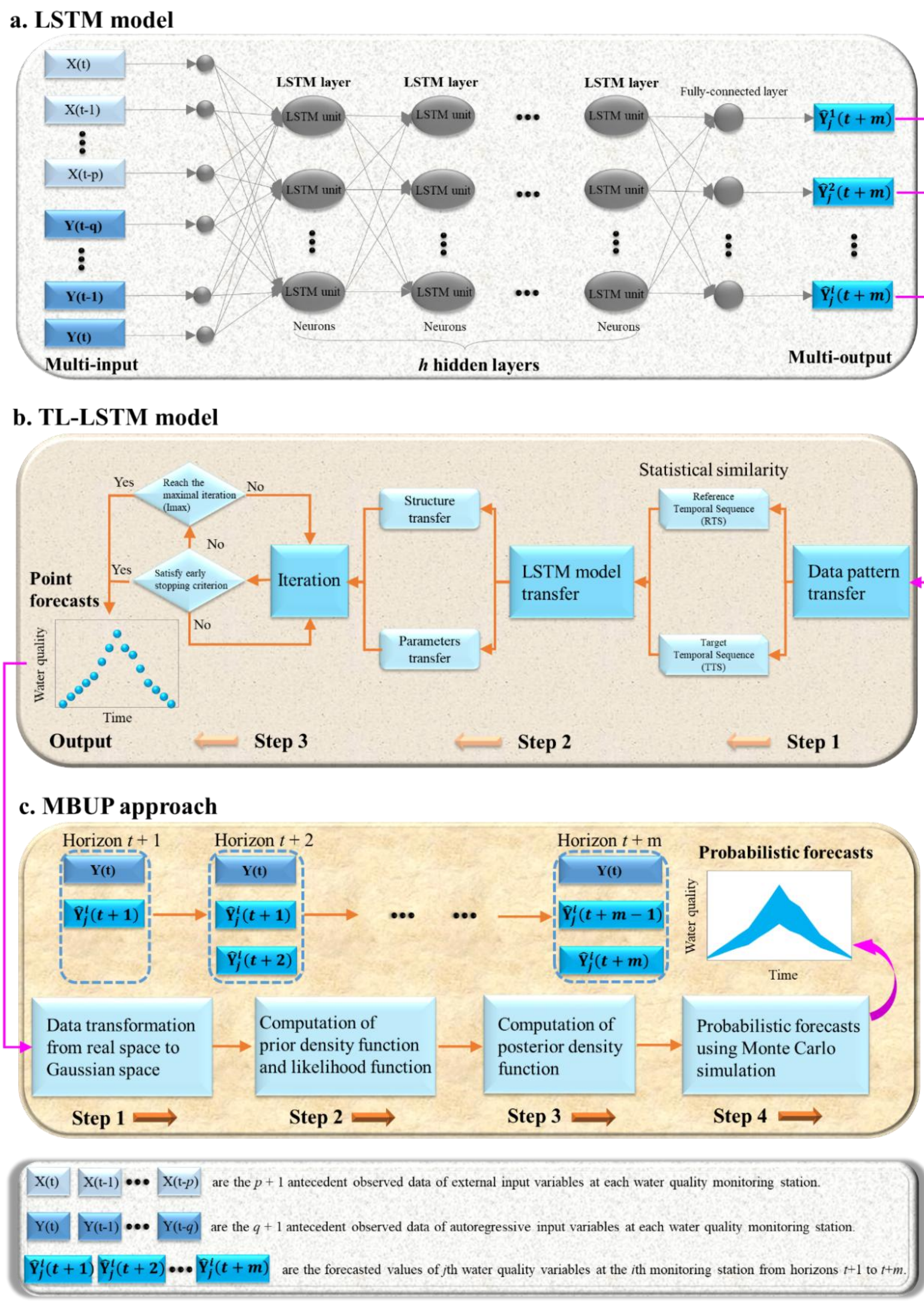

Fig. 1. Probabilistic forecast architecture. (a) LSTM neural network model. (b) Hybrid of Transfer Learning and LSTM model (TL-LSTM). (c) MBUP approach.

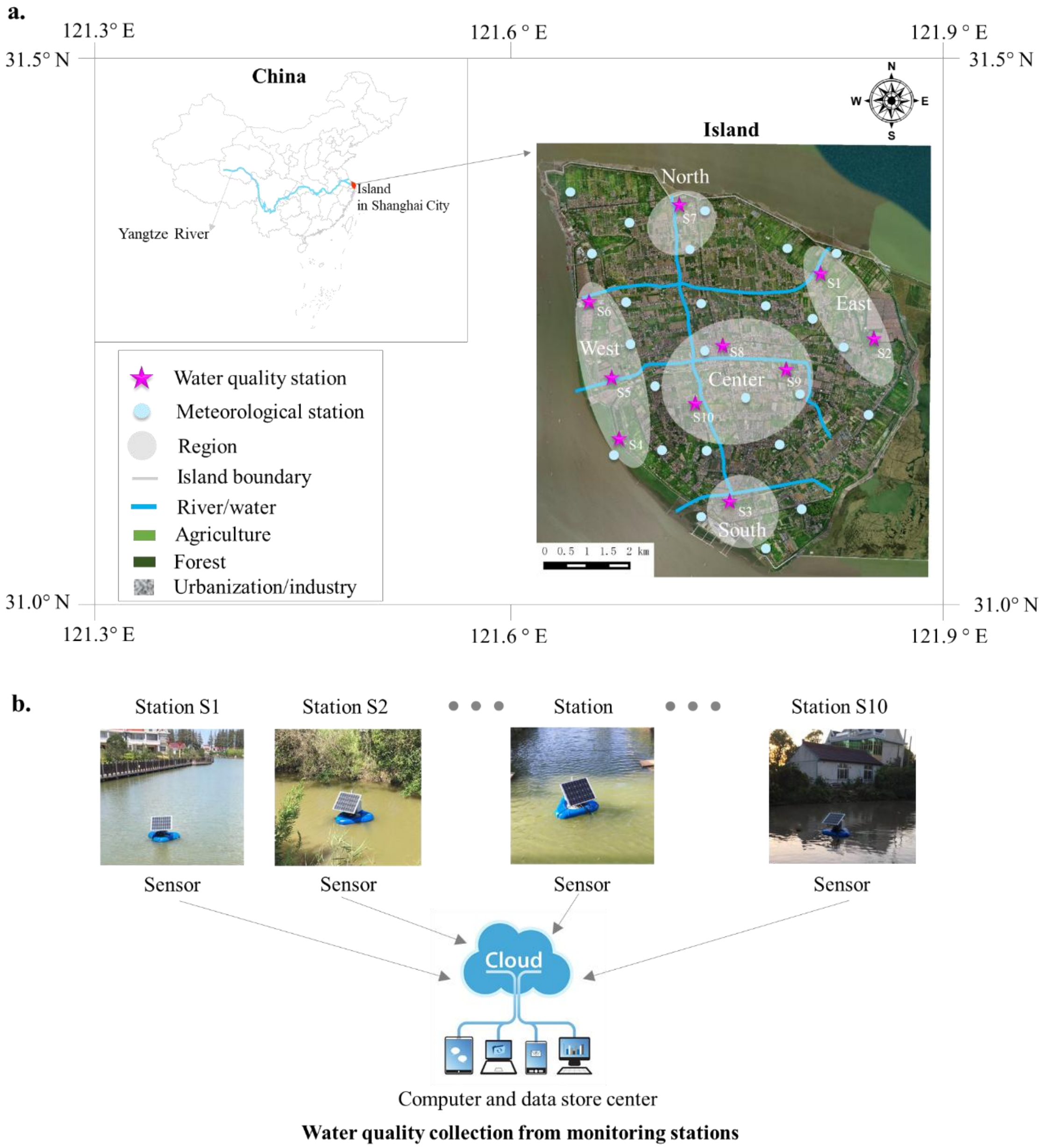

Fig. 2. Study area and water quality data collection. (a) Meteorological and river water quality monitoring stations in the island of Shanghai City. (b) Water quality data collection from monitoring stations.

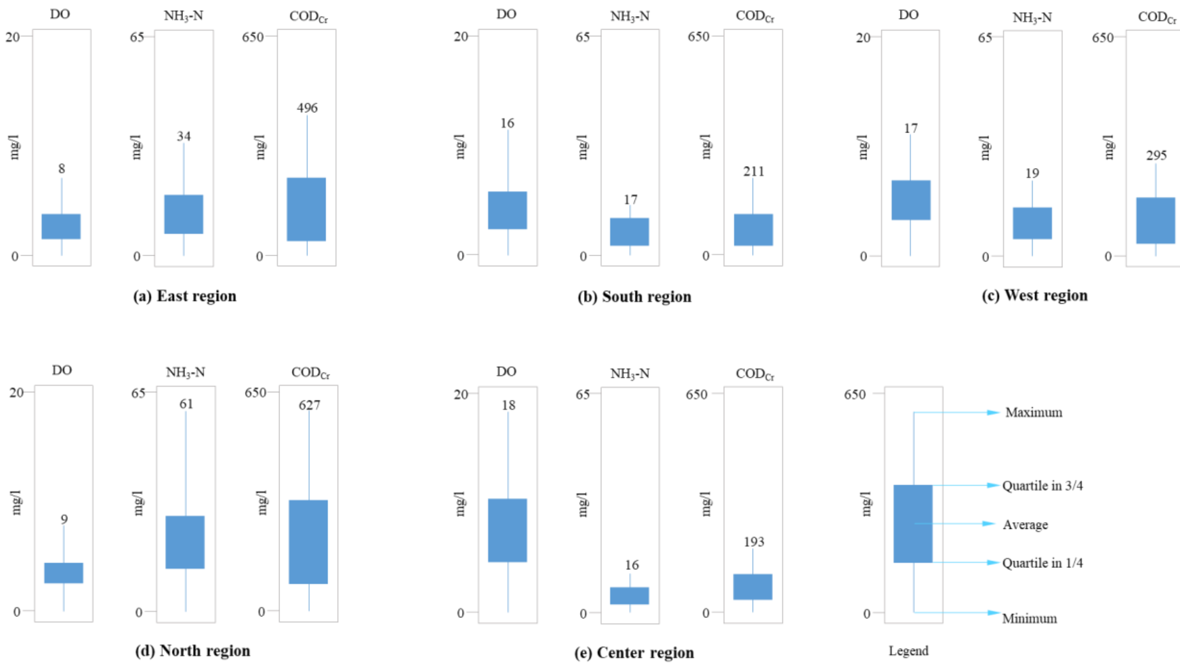

Fig. 3. Statistic indexes of DO, NH_3-N and COD_{Cr} concentrations at five regions (a–e) in the island. The abbreviations (max, ave, min, std) denote the maximum, average, minimum and standard deviation respectively. The time period of statistic covers four years (31/08/2015-31/08/2019).

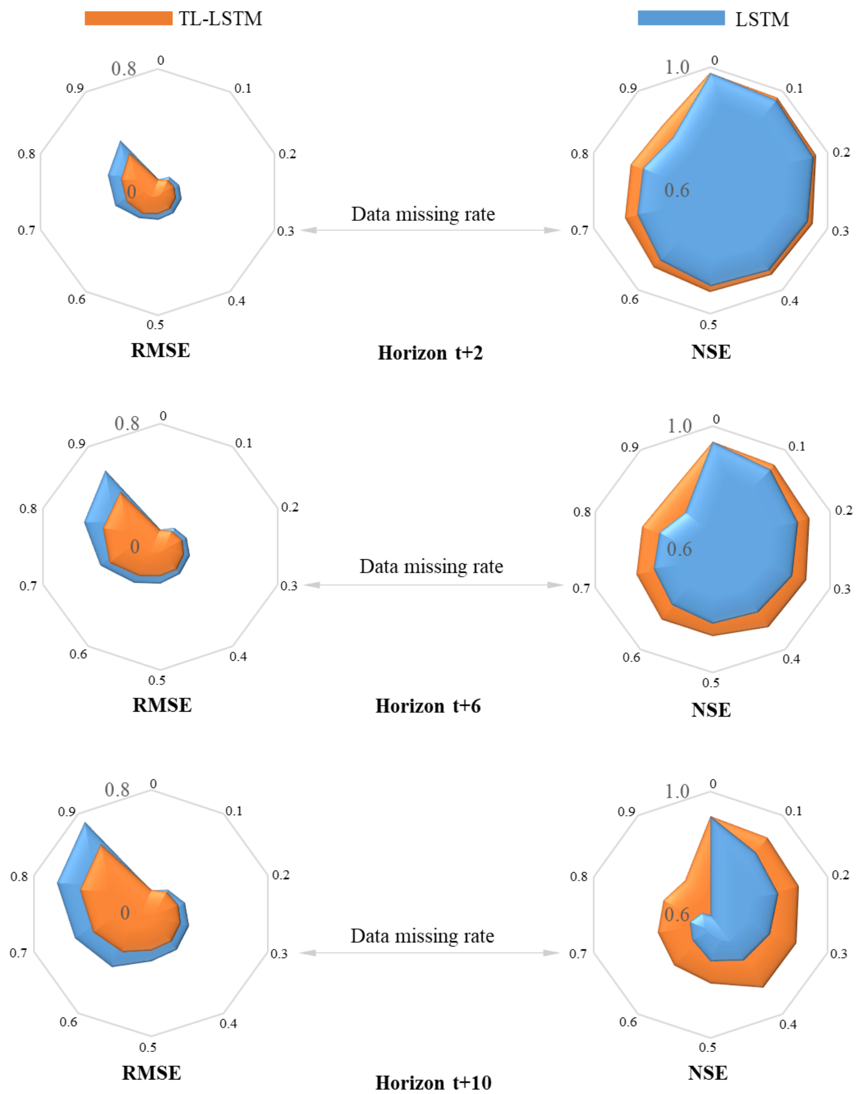

Fig. 4. Model performance of deterministic forecasts concerning water quality under different data missing rates (0 − 0.9, step = 0.1) at horizons t+2, t+6, t+10 at the Station S10 in the testing stage. In comparison analysis between TL-LSTM and LSTM models, the position of data missing in the initial data input always kept consistent in both models. That was to say, the position of data missing was randomly generated for the TL-LSTM model while the LSTM model had the same position of data missing with the TL-LSTM model. The computation result was the average result of 10 runs of each model. The value of RMSE was the average RMSE of water quality forecasts (DO, NH_3-N and COD_{Cr} values with standardization) while the value of NSE was the average NSE of water quality forecasts (DO, NH_3-N and COD_{Cr} values with standardization).

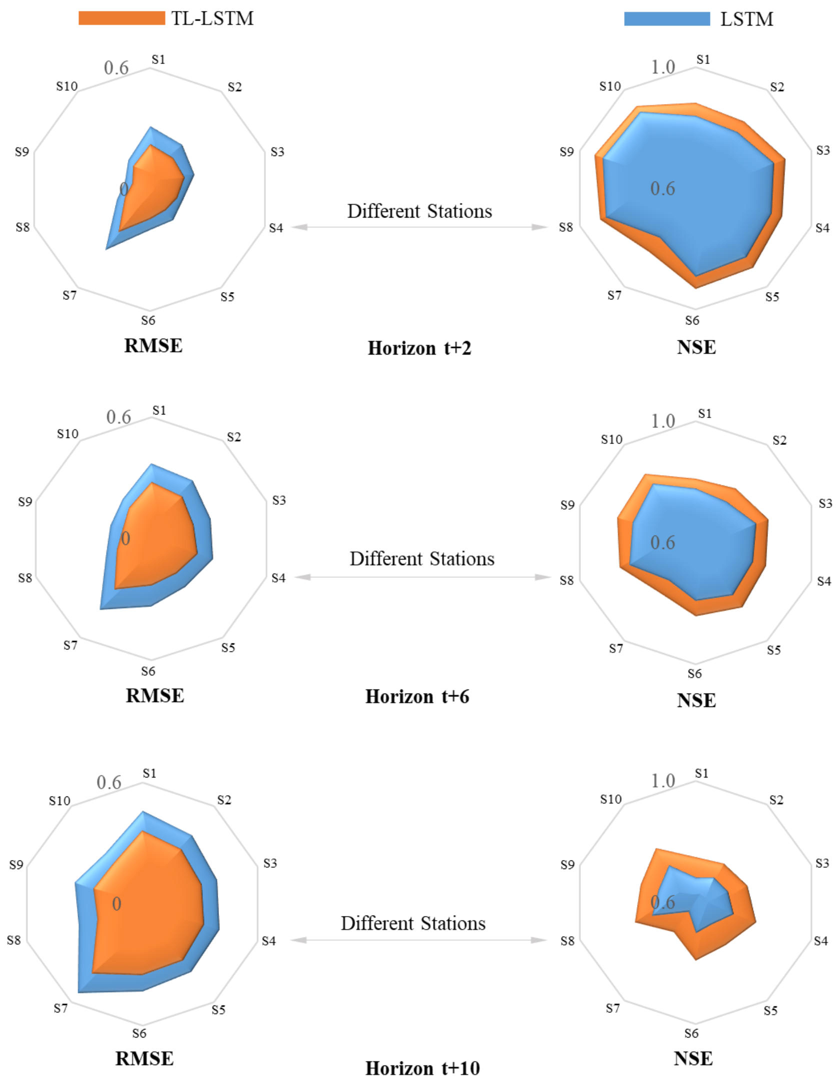

Fig. 5. Model performance of deterministic forecasts concerning water quality (DO, NH_3-N, and COD_{Cr}) under the data missing rate (= 0.5) at horizons t+2, t+6, t+10 at different stations (S1 – S10) in the testing stages. In comparison analysis between TL-LSTM and LSTM models, the position of data missing in the initial data input always kept consistent in both models. The computation result was the average result of 10 runs of each model. The value of RMSE was the average RMSE of water quality forecasts (DO, NH_3-N and COD_{Cr} values with standardization) while the value of NSE was the average NSE of water quality forecasts (DO, NH_3-N and COD_{Cr} values with standardization).

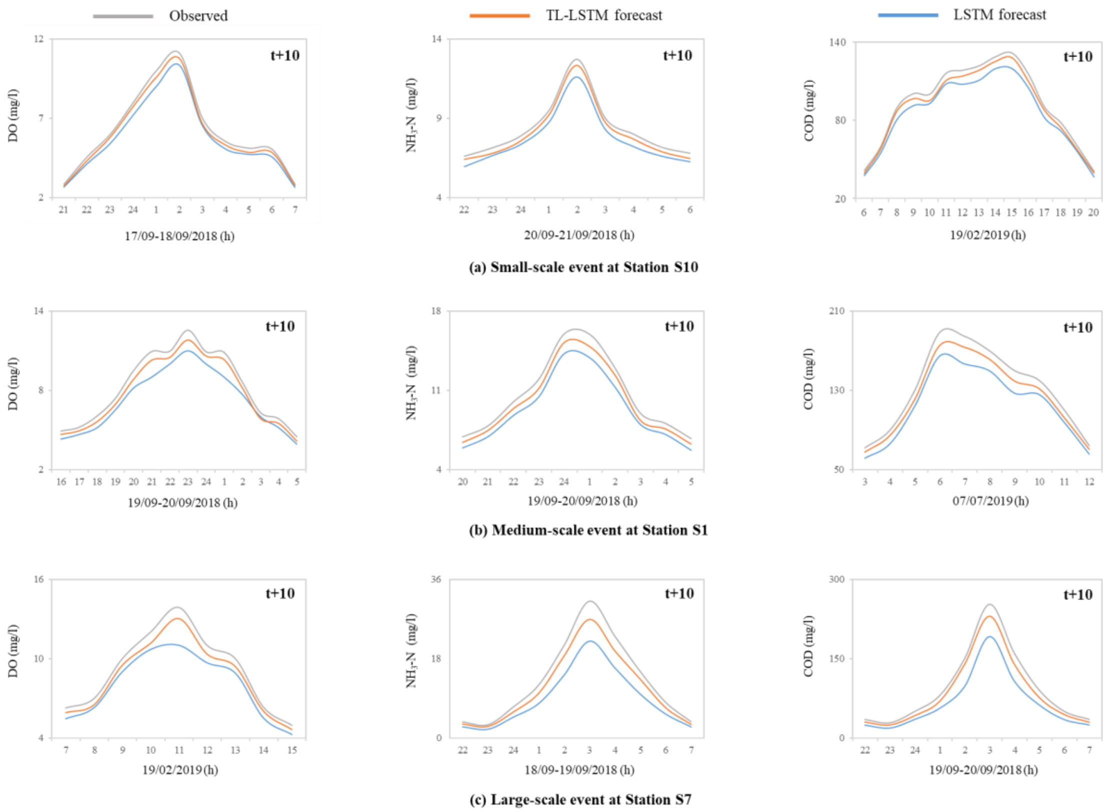

Fig. 6. Deterministic water quality forecast results (DO, NH_3-N and COD_{Cr}) of LSTM and TL-LSTM models under the data missing rate (= 0.5) at horizon t+10 in the testing stages at the Station S1 (East region), the Station S7 (North region) and the Station S10 (Center region) respectively. In comparison analysis between TL-LSTM and LSTM models, the position of data missing in the initial data input always kept consistent in both models. The computation result was the average result of 10 runs of each model. The test event with small-scale (a) occurred at the Station S10. The test event with medium-scale (b) occurred at the Station S1. The test event with high-scale (c) occurred at the Station S7.

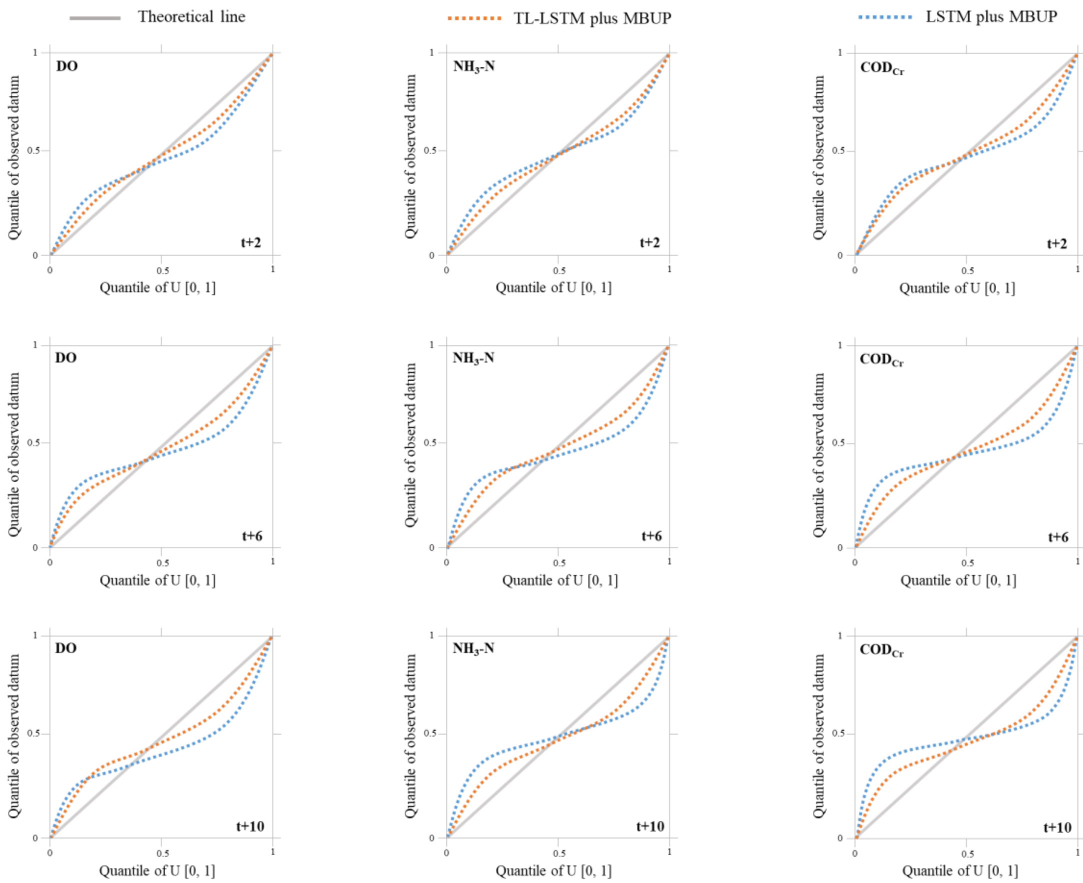

Fig. 7. Quantile-Quantile (QQ) plots of probabilistic water quality (DO, NH_3-N and COD_{Cr}) forecasts at the Station S7 under the data missing rate (= 0.5) at horizons t+2, t+6, t+10 in the testing stages.

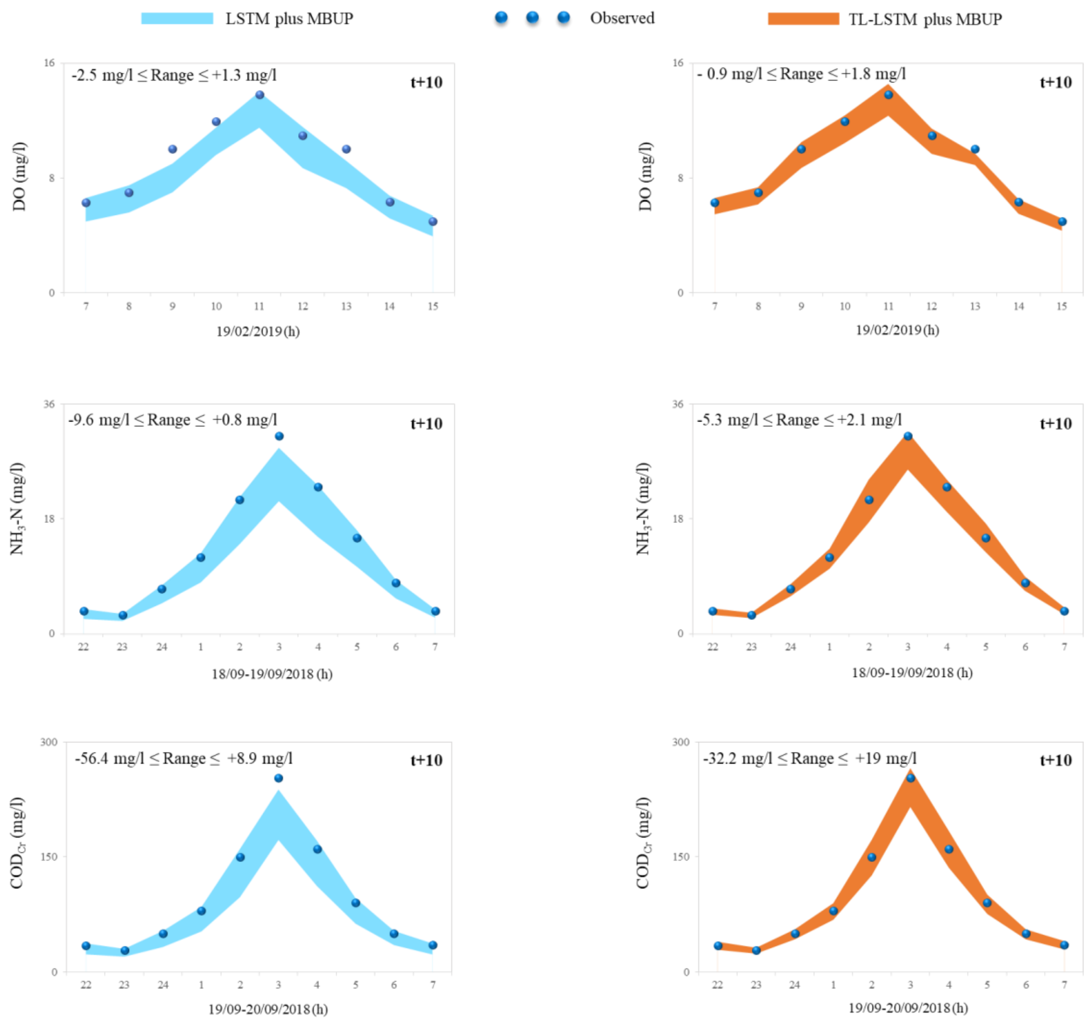

Fig. 8. Probabilistic water quality (DO, NH_3-N and COD_{Cr}) forecasts for Station S7 under the data missing rate (= 0.5) at horizon t+10 in the testing stages. The range is equal to the forecast minus the observation.

Tables

Click here to download Table: Tables.docx

Table 1. Input data of deterministic forecast models under missing data conditions

Stage	TL-LSTM model	LSTM model
Training	$\{[S^{R}_{t-q}, S^{R}_{t-q+1}, \cdots, S^{R}_{t-1}, S^{RR}_{t}] \rightarrow [S^{R}_{t+m}]\}$	
Validating	$\{[S^{T}_{t-p}, S^{T}_{t-p+1}, \cdots, S^{T}_{t-1}, S^{R}_{t}] \rightarrow [S^{T}_{t+m}]\}$	$\{[S^{T}_{t-p}, S^{T}_{t-p+1}, \cdots, S^{T}_{t-1}, S^{R}_{t}] \rightarrow [S^{T}_{t+m}]\}$
Testing	$\{[S^{T}_{t-p}, S^{T}_{t-p+1}, \cdots, S^{T}_{t-1}, S^{R}_{t}] \rightarrow [S^{T}_{t+m}]\}$	

Notes: Each stage (training, validating and testing) of the dataset was erased with one percentage (e.g. 50%) during the establishment and application of the LSTM models. S^{R} and S^{RR} were the selected RTSs. S^{T} was the incomplete TTS. Take the incomplete TTS with one missing segment $S^{T} = [S^{T}_{1}, S^{M}_{2}, S^{T}_{3}]$ for example, S^{M}_{2} was the missing segment, S^{T}_{1} and S^{T}_{3} were the complete segments. If S^{M}_{2} was at the beginning or the end of S^{T}, S^{T}_{1} or S^{T}_{3} was empty dataset. S^{R} was the highest correlation complete sequence of S^{T} while S^{RR} was the highest correlation complete sequence of S^{R}.

Table 2. Basic information on ten monitoring stations in five regions

Region	Station	Type of pollution	Source
East	S1 & S2	Nonpoint source	Aquaculture or natural area
South	S3	Point source	Industry
West	S4-S6	Point source	Industry
North	S7	Nonpoint source	Farmland and livestock
Center	S8-S10	Point source	Urban domestic sewage

Table 3. Statistic indexes of the other 9 input factors at five regions

Region	Index	Factor								
		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
East	Max.	13.6	1380.6	1524.8	1087.3	14.9	39.8	53.2	9.2	20.7
	Ave.	7.5	271.0	351.3	241.6	12.9	17.6	22.7	3.4	16.5
	Min.	6.9	200.0	5.8	7.1	4.3	0.0	0.0	0.0	0.0
South	Max.	11.5	1460.7	1337.0	948.5	14.7	35.8	47.8	4.4	24.5
	Ave.	7.5	219.7	334.4	229.4	12.0	18.1	19.7	1.8	17.8
	Min.	6.2	153.5	4.2	6.4	3.6	0.0	0.0	0.0	0.0
West	Max.	13.7	1388.6	1358.2	980.3	15.2	39.8	43.6	4.3	21.2
	Ave.	7.5	217.1	462.6	315.3	12.0	17.9	14.6	1.6	16.9
	Min.	6.3	128.1	7.6	5.8	2.1	0.0	0.0	0.0	0.0
North	Max.	19.3	1402.1	1679.2	1191.5	14.3	39.8	51.7	6.7	19.7
	Ave.	7.5	234.5	453.9	307.4	11.3	17.8	20.5	2.4	15.3
	Min.	6.9	180.1	8.4	10.3	1.8	0.0	0.0	0.0	0.0
Center	Max.	11.5	1400.9	1113.8	751.8	14.6	32.2	45.2	3.6	22.3
	Ave.	7.5	229.9	439.4	298.7	11.3	17.6	15.9	0.7	17.2
	Min.	6.8	180.3	7.9	4.7	0.9	1.3	0.0	0.0	0.0

The abbreviations of Max, Ave and Min denoted the maximum, average and minimum. The factors in columns No. (1)-(9) were pondus hydrogenii (/), oxidation-reduction potential (mV), conductivity (S/m), turbidity (mg/l), water level (m), water temperature (℃), precipitation (mm/h), wind speed (m/s) and light intensity (mega-joule/m^2) respectively.

Table 4. Parameters of the LSTM and TL-LSTM models at horizon t+10 in the training and validating stages

Model	Data missing rate	Parameters					Training		Validating	
		G_{max}	Neurons	Hidden layer	Learning rate	Dropout probability	RMSE	NSE	RMSE	NSE
LSTM [a]	0.5	1000	20	1	0.001	0.5	0.65	0.64	0.68	0.62
			30				0.43	0.72	0.42	0.73
			40				0.58	0.67	0.61	0.64
			50				0.71	0.61	0.71	0.61
LSTM	0.5	1000	30	2	0.001	0.5	0.37	0.71	0.39	0.70
				3			0.31	0.75	0.29	0.76
				4			0.49	0.68	0.51	0.67
TL-LSTM [b]	0.5	1000	30	3	0.001	0.5	0.24	0.88	0.23	0.89

A value in bold indicated the optimal parameter. The data missing rate (= 0.5) denoted that all DO, NH_3-N and COD_{Cr} time series at 10 stations missed 50% of datasets and each stage (training, validating and testing) of the dataset was erased with the same percentage (i.e. 50%) during the establishment and application of the LSTM models. The computation result was the average result of 10 runs of each model. The value of RMSE was the average RMSE of water quality forecasts (DO, NH_3-N and COD_{Cr} values with standardization) while the value of NSE was the average NSE of water quality forecasts (DO, NH_3-N and COD_{Cr} values with standardization).

[a] LSTM denoted the long-short term memory model.

[b] TL-LSTM denoted the hybrid of transfer learning and long-short term memory model.

Table 5. Impact of data missing in meteorological and water quality factors on the performance of LSTM models at the Station S7 in the testing stage.

Scenario: missing factor	Model	Indicator	Horizon		
			***t*+2**	***t*+6**	***t*+10**
No.1: Precipitation	TL-LSTM	RMSE	0.19	0.24	0.29
		NSE	0.87	0.82	0.77
	LSTM	RMSE	0.22	0.28	0.33
		NSE	0.83	0.78	0.73
No.2: Wind speed	TL-LSTM	RMSE	0.16	0.21	0.27
		NSE	0.92	0.86	0.81
	LSTM	RMSE	0.19	0.26	0.32
		NSE	0.89	0.84	0.79
No.3: NH_3-N	TL-LSTM	RMSE	0.22	0.27	0.32
		NSE	0.88	0.83	0.79
	LSTM	RMSE	0.31	0.38	0.48
		NSE	0.83	0.78	0.69
No.4: COD_{Cr}	TL-LSTM	RMSE	0.21	0.28	0.30
		NSE	0.90	0.85	0.81
	LSTM	RMSE	0.32	0.36	0.46
		NSE	0.84	0.80	0.71
No.5: All meteorological and water quality factors	TL-LSTM	RMSE	0.26	0.31	0.39
		NSE	0.86	0.81	0.76
	LSTM	RMSE	0.37	0.43	0.54
		NSE	0.80	0.75	0.64

The value of RMSE was the average RMSE of water quality forecasts (DO, NH_3-N and COD_{Cr} values with standardization) while the value of NSE was the average NSE of water quality forecasts (DO, NH_3-N and COD_{Cr} values with standardization).

Table 6. Impact of data missing positions on the performance of LSTM models at the Station S7 in the testing stage

Scenario: data missing position	Model	Indicator	Horizon		
			***t*+2**	***t*+6**	***t*+10**
No.1: Peak data possessing the missing rate (0.5)	TL-LSTM	RMSE	0.22	0.29	0.37
		NSE	0.84	0.80	0.75
	LSTM	RMSE	0.26	0.38	0.49
		NSE	0.81	0.76	0.69
No.2: Trough data possessing the missing rate (0.5)	TL-LSTM	RMSE	0.17	0.22	0.27
		NSE	0.93	0.87	0.83
	LSTM	RMSE	0.21	0.25	0.30
		NSE	0.90	0.84	0.80
No.3: Peak and trough data possessing the missing rate (0.25) respectively	TL-LSTM	RMSE	0.20	0.25	0.31
		NSE	0.89	0.84	0.78
	LSTM	RMSE	0.24	0.28	0.35
		NSE	0.85	0.80	0.73
No.4: Non-peak and non-trough data possessing the missing rate (0.5)	TL-LSTM	RMSE	0.19	0.24	0.29
		NSE	0.91	0.85	0.80
	LSTM	RMSE	0.22	0.27	0.33
		NSE	0.88	0.82	0.77

The value of RMSE was the average RMSE of water quality forecasts (DO, NH_3-N and COD_{Cr} values with standardization) while the value of NSE was the average NSE of water quality forecasts (DO, NH_3-N and COD_{Cr} values with standardization).

Table 7. Results of probabilistic water quality forecasting under the data missing rate (= 0.5) at horizons t+2, t+6, t+10 in the testing stages

Station	Model	Indicator	Horizon		
			***t*+2**	***t*+6**	***t*+10**
S1	TL-LSTM plus MBUP	CR(%)	96.17	92.39	88.62
		RB	0.09	0.18	0.25
	LSTM plus MBUP	CR(%)	95.22	90.04	83.56
		RB	0.12	0.22	0.30
S7	TL-LSTM plus MBUP	CR(%)	95.07	91.43	85.96
		RB	0.13	0.21	0.30
	LSTM plus MBUP	CR(%)	94.24	89.25	80.07
		RB	0.15	0.27	0.38
S10	TL-LSTM plus MBUP	CR(%)	98.63	93.17	89.66
		RB	0.08	0.15	0.22
	LSTM plus MBUP	CR(%)	97.48	91.24	84.39
		RB	0.10	0.21	0.26

The value of CR was the average CR of water quality forecasts (DO, NH_3-N and COD_{Cr}) while the value of RB was the average RB of water quality forecasts (DO, NH_3-N and COD_{Cr}).

Declaration of interests

☒ 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.

☐ The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: