Journal Pre-proof

Hybrid decision tree-based machine learning models for short-term water quality prediction

Hongfang Lu, Xin Ma

PII: S0045-6535(20)30362-3

DOI: https://doi.org/10.1016/j.chemosphere.2020.126169

Reference: CHEM 126169

To appear in: *ECSN*

Received Date: 21 October 2019

Revised Date: 4 February 2020

Accepted Date: 9 February 2020

Please cite this article as: Lu, H., Ma, X., Hybrid decision tree-based machine learning models for short-term water quality prediction, *Chemosphere* (2020), doi: https://doi.org/10.1016/j.chemosphere.2020.126169.

CRediT author statement

Hongfang Lu: Conceptualization, Methodology, Data curation, Writing- Original draft preparation

Xin Ma: Investigation, Writing- Reviewing and Editing

Hybrid decision tree-based machine learning models for short-term water quality prediction

Hongfang Lu[a,b,*], Xin Ma[c]

[a] *State Key Laboratory of Oil and Gas Reservoir Geology and Exploitation, Southwest Petroleum University, Chengdu 610500, China*

[b] *School of Science, Southwest University of Science and Technology, Mianyang 621010, China*

[c] *Trenchless Technology Center, Louisiana Tech University, Ruston LA71270, United States*

*Corresponding author: hlu006@latech.edu (H. Lu)

Address: 599 Dan Reneau Drive, Ruston, LA 71270, USA

Abstract: Water resources are the foundation of people's life and economic development, and are closely related to health and the environment. Accurate prediction of water quality is the key to improving water management and pollution control. In this paper, two novel hybrid decision tree-based machine learning models are proposed to obtain more accurate short-term water quality prediction results. The basic models of the two hybrid models are extreme gradient boosting (XGBoost) and random forest (RF), which respectively introduce an advanced data denoising technique - complete ensemble empirical mode decomposition with adaptive noise (CEEMDAN). Taking the water resources of Gales Creek site in Tualatin River (one of the most polluted rivers in the world) Basin as an example, a total of 1875 data (hourly data) from May 1, 2019 to July 20, 2019 are collected. Two hybrid models are used to predict six water quality indicators, including water temperature, dissolved oxygen, pH value, specific conductance, turbidity, and fluorescent dissolved organic matter. Six error metrics are introduced as the basis of performance evaluation, and the results of the two models are compared with the other four conventional models. The results reveal that: (1) CEEMDAN-RF performs best in the prediction of temperature, dissolved oxygen and specific conductance, the mean absolute percentage errors (MAPEs) are 0.69%, 1.05%, and 0.90%, respectively. CEEMDAN-XGBoost performs best in the prediction of pH value, turbidity, and fluorescent dissolved organic matter, the MAPEs are 0.27%, 14.94%, and 1.59%, respectively. (2) The average MAPEs of CEEMDAN-RF and CEEMMDAN-XGBoost models are the smallest, which are 3.90% and 3.71% respectively, indicating that their overall prediction performance is the best. In addition, the stability of the prediction model is also discussed in this paper. The analysis shows that the prediction stability of CEEMDAN-RF and CEEMDAN-XGBoost is higher than other benchmark models.

Keywords: decision tree-based model; short-term; water quality prediction; extreme gradient boosting; random forest; data denoising

Nomenclature

$\overline{IMF_k}$	k-th modal component
$E_k(\cdot)$	k-th modal component obtained by EMD decomposition
O_t	observation value at time t
P_t	prediction value at time t
$d^i(t)$	i-th signal sequence
f_i	final prediction result
f_n	normalized prediction result
g_i and h_i	the first and second derivatives of the loss function in the gradient direction
$wn^i(t)$	white noise
x_i	i-th eigenvector
$\hat{y}$	predicted value
y_i	the true value of the sample
z_i	raw data
z_{min} and z_{max}	minimum and maximum of the raw data, respectively
z_n	normalized data
Ω	penalty term for model complexity
F	regression tree set
I	number of tests
$IMFs$	intrinsic mode functions
K	number of trees
R	final residue
T	number of leaves in the tree
f	a regression tree in tree space F
l	loss function
r	residue
t	current iteration
w	internal split tree weight
γ and λ	configurable parameters
ε	noise standard deviation

Abbreviations

ANN	artificial neural network
BPNN	back propagation neural network
CEEMDAN	complete ensemble empirical mode decomposition with adaptive noise
CPU	central processing unit
EEMD	ensemble empirical mode decomposition
EMD	empirical mode decomposition
FDOM	fluorescent dissolved organic matter
FNU	Formazin Nephelometric Units
GMNN	Gamma memory neural network
JENN	Jordan-Elman neural network
LSSVM	least squares support vector machine

LSTM	long short-term memory
MAE	mean absolute error
MAPE	mean absolute percentage error
Max.	maximum
Min.	minimum
ppb	parts per billion
PSO	particle swarm optimization
QSE	quinine sulfate equivalents
RBFNN	radial basis function neural network
RF	random forest
RMSE	root mean square error
RMSPE	root mean squared percentage error
SD	standard deviation
SDE	standard deviation of errors
SVD	singular value decomposition
SVM	support vector machine
U1	Theil U statistic 1
U2	Theil U statistic 2
USGS	U.S. Geological Survey
XGboost	extreme gradient boosting

1. Introduction

Water is the resource that human beings depend on for survival. The prevention and control of water pollution have been a hot issue of concern in recent decades. Accidents of water pollution have been reported uninterruptedly all over the world in recent years (Hounslow, 2018). It can be seen that many water pollution accidents occur in the absence of information management. Many water pollution accidents can only be remedied afterward because of the lack of early prediction. Therefore, water quality prediction has become a hot topic in water environment science.

Water quality prediction is an essential work in water environment management. Accurate forecasting value will undoubtedly improve the management level of water resources. At present, many water resource management departments have set monitoring points to observe water quality changes, but they cannot play the role of water quality prediction. However, these monitoring data can provide a predictive basis for some data-driven models. Through real-time control of future water quality changes, water pollution degree can be accurately judged. Besides, accurate water

quality predictions can also provide a basis for policymakers and provide data to the environmental management department to act as an “early warning”.

We reviewed the literature on water quality predictions in recent years. Zhao et al. (2007) used BPNN to predict the water quality of Yuqiao Reservoir. Singh et al. (2009) used the ANN model to predict the water quality of the river and analyzed it with the case of the Gomti river in India. Liu et al. (2016) adopted a multi-task multi-view learning method to predict urban water quality. The effectiveness of the method was verified by experiments. Palani et al. (2008) applied the ANN model to the prediction of coastal water quality in Singapore. Water quality parameters considered include salinity, temperature, dissolved oxygen, and chlorophyll-a. West and Dellana (2011) used JENN and GMNN to predict the basin water quality. Chan et al. (2013) established a 3D hydrodynamic model to predict the water quality of Hong Kong beach in real time. The experiment proved that the accuracy of prediction is higher than 80%. Meyers et al. (2017) used several machine learning-based models (ANN, RF, and SVM) to predict water turbidity. Peng et al. (2019) proposed a framework for real-time prediction of daily water quality and applied it to Lake Chaohu in China, which can better predict dissolved oxygen, total phosphorus, and other parameters. Huang et al. (2019) established a prediction system for urban estuary water quality and used the gradient boosting machine model to fill and predict the flow. Alba et al. (2019) used an ANN-based model to predict the bathing water quality of the estuary, which combines laboratory analysis, machine learning, and numerical simulation to achieve real-time water quality management. According to the literature review, some researchers have established water quality prediction systems and used experimental methods to predict water quality. However, physical methods are time-consuming and labor-intensive. With the prevalence of machine learning and deep learning (Lu et al., 2020a,b; Kong and Ma, 2018; Ma et al., 2020), more and more scholars use intelligent models to predict water quality (Liang et al., 2020; Gao et al., 2019; Dabrowski et al., 2020; Zhang et al., 2017; Hussein et al., 2019; Panidhapu et al., 2020). In addition, intelligent model has been applied not only in water quality prediction, but also in the field of environment and energy engineering (Xie et al., 2020; Wu et al., 2019; Zhang et al., 2019; Yang et al., 2020; Zhao et al., 2019). Although the accuracy of water quality prediction is improving, because water quality is unstable and non-linear in time series, more accurate prediction methods are worthy of further study. Therefore, this paper proposes two hybrid models, which combine CEEMDAN method with the original models, thus enhancing their prediction accuracy.

The rest of the paper is organized as follows: Section 2 introduces six water quality indicators, Section 3 describes the collected data, data cleaning technology--CEEMDAN and two decision tree-based machine learning models--XGBoost and RF. Section 4 describes the process of

prediction and error metrics. Section 5 reveals the prediction results and discussions. Section 6 summarizes the main conclusions and future works.

2. Water quality indicators

There are many water quality indicators, which can be divided into chemical indicators, physical indicators, biological indicators, radioactive indicators, and so on (Valdivia-Garcia et al., 2019). Water for different purposes usually has different evaluation indicators. For example, the critical indicators of domestic water use include temperature, pH value, biochemical oxygen demand, while the indicators of food industry water use include suspended solid, pH value and number of Escherichia coli. This paper introduces several common water quality indicators (see Table 1), including temperature, dissolved oxygen, pH value, specific conductance, turbidity, and FDOM.

157 Table 1. Common water quality indicators and explanations.

Indicator	Explanation	Reference(s)
Water temperature	Temperature is a critical physical indicator of water. The sudden rise in water temperature during daily monitoring indicates that the water body may be contaminated by new sources of pollution. Thermal pollution may also cause biological growth to increase and cause biological pollution in the water.	Tao et al., 2020; Graf et al. 2019
Dissolved oxygen	Free oxygen dissolved in water is called dissolved oxygen, it is a parameter to measure the quality of water. In general, the concentration of dissolved oxygen in water is called equilibrium concentration when it approaches saturation. The saturation value of dissolved oxygen is 9.17 mg/L at 20℃. When water is polluted by oxygen-consuming pollutants, dissolved oxygen decreases. Oxygen-consuming pollutants include carbohydrates, proteins, oils, amino acids, fatty acids, esters and other organic compounds. These pollutants mainly come from domestic sewage and some industrial wastewater.	Larsen et al., 2019
pH value	The pH of natural water is generally between 6.5 and 8.5. A suitable pH range for drinking water is from 7 to 8.5, with a limit range of 6.5 to 9.2. Generally, fish live normally in water with a pH of 6.5 to 8.5. The crop is suitable for growth in water with a pH of 6 to 7.5. Long-term irrigation of water with a pH lower than 5.5 will cause the nitrifying bacteria in the soil to be inhibited, the nitrification will be weakened, and the nitrogen fertilizer will not be fully released.	Mosley et al., 2010
Specific conductance	Specific conductance represents the ability to conduct current in aqueous solution, and it is also an index for routine monitoring of water quality with multi-parameters. The specific conductance is proportional to the ion content in the solution, so the total soluble matter content can be estimated indirectly.	Makarewicz et al., 2012
Turbidity	Turbidity indicates the obstruction extent of light sources by suspended matter in water. The reason for the increase in turbidity in rivers and lakes is because the river water contains many suspended substances. When the turbidity is large, it will affect the photosynthesis of aquatic organisms and reduce the self-purification ability of water.	Kerr et al., 2018
FDOM	FDOM is a fast and simple method for tracking dissolved organic matter in natural water, it is often used as a biological indicator of lake water.	Liu et al., 2019

3. Material and methods

In this paper, two new hybrid models, CEEMDAN-XGBoost and CEEMMDAN-RF, are used to predict water quality indicators. This section describes the collected data and introduces the relevant theories of CEEMDAN, XGBoost, and RF.

3.1. Collected data description

The water quality data for this paper is from the Tualatin River in Oregon, USA, it is called one of the ten most dangerous rivers in the world. The Tualatin River drains 712 square miles in the northwest corner of Oregon, it is a sub-basin of the Willamette River Basin. It is about 134 kilometers long, and the slope of most sections is very flat. The main tributaries of the Tualatin River include Scoggins, Gales, Dairy, Rock and Fanno Creek. The summer water flow is discharged by water from the Scoggins Reservoir and the Barney Reservoir, which diverts water into the upper Tualatin River. Wastewater from wastewater treatment plants accounts for a large portion of summer river flows.

Before the 1970s, wastewater treatment plants discharged high concentrations of ammonia, nitrogen, and phosphorus into the mainstream of the Tualatin River. High ammonia concentration usually causes obvious nitrification in rivers, resulting in low dissolved oxygen concentration. In addition, in summer, the abundance of phytoplankton in the Tualatin River results in the river's water quality violating the requirements of minimum dissolved oxygen and maximum pH value. Later, in 1970, the Unified Sewerage Agency of Washington County was established. They used various water treatment methods to control the pollution and health problems of the Tualatin River. By 2002, many water bodies in the Tualatin River Basin had been confirmed to be damaged.

This paper chooses the time series data of the Gales Creek site of Tualatin River as the research object. The raw data of the monitoring point is from USGS (https://www.usgs.gov). The collected water quality data are temperature, dissolved oxygen, pH value, specific conductance, turbidity, and FDOM from 0:00 on May 1, 2009 to 23:00 on July 20, 2019 (the data interval is one hour). Each water quality indicator has 1875 data, and their statistical descriptions are shown in Table 2. It can be seen that the six datasets have multiple probability density function types such as Triangular, Johnson SB and Lognormal (3P), indicating that the data used in this paper is diverse and extensive, and lays a foundation for more convincing conclusions.

187 **Table 2.** Statistical descriptions of the data in this paper.

Water quality indicator	Unit	Data period (month/day/year)	Amount of data	Test set data amount	Statistical distribution		Statistical characteristics			
					Distribution	Parameters	Max.	Min.	Mean	SD
Temperature	℃	05/01/2019 00:00-07/20/2019 23:00	1875	187	Triangular	m=17.73, a=9.853, b=23.722	23.55	9.98	17.08	2.79
Dissolved oxygen	mg/L				Johnson SB	γ=0.24388, δ=1.121, λ=4.2411, ξ=7.2869	11.11	7.12	9.21	0.81
pH	Dimensionless				Error function	k=3.0988, σ=0.08784, μ=7.3089	7.52	7.08	7.31	0.09
Specific conductance	uS/cm				Johnson SB	γ=-0.25905, δ=0.51603, λ=35.138, ξ=94.969	133.9	93.8	115.24	10.70
Turbidity	FNU				Lognormal (3P)	σ=0.96807, μ=0.22892, γ=0.45261	16.0	0.5	2.25	1.72
FDOM	ppb QSE				Dagum	k=0.82881, α=15.915, β=12.418, γ=0	32.62	8.46	12.28	1.75

188

3.2. Decision tree-based machine learning models

The basic models of the two hybrid models used in this paper are XGBoost and RF. They are all belong to decision tree-based machine learning models. The decision tree-based model has many advantages:

a) Ability to handle both data and regular attributes; b) Insensitive to missing values; c) High efficiency, the decision tree only needs to be built once. In fact, there are other models in the field of machine learning, such as ANN and SVM. Compared to them, decision tree-based models may have faster calculation speed and are more conducive to short-term prediction. Moreover, water quality monitoring data sometimes have missing values due to equipment failure, the decision tree-based model has an advantage in forecasting.

3.2.1. eXtreme Gradient Boosting (XGBoost)

XGBoost was proposed by Chen and Guestrin (2016) and is based on the C++ language (Nobre and Neves, 2019). The model has achieved great success since its appearance, and it is always seen in the top models in various data mining competitions. XGBoost is able to integrate multiple weak learning machines into one strong learning machine by iterating and generating multiple trees, and it has the following features: a) It can automatically utilize the multithreading of the CPU for parallelism, while improving the algorithm to improve accuracy, and this is the most prominent feature of XGBoost; b) It is a lifting learning algorithm based on the decision tree model and can process sparse data automatically; c) large amounts of data can be processed at high speed according to block technology.

In the XGBoost model, tree model adopts additive model

$$\hat{y} = \sum_{k=1}^{K} f_k(x_i)\,, f_k \in F \tag{1}$$

The objective function is

$$L(\phi) = \sum_i l(\hat{y}, y_i) + \sum_k \Omega(f_k) \tag{2}$$

where $\Omega(f) = \gamma T + 0.5\lambda\|w\|^2$, $w = (w_1, w_2, \cdots, w_k)$.

Because learning all tree parameters at once is challenging, XGBoost uses an additive strategy that learns the parameters of one tree at a time:

$$\hat{y}_i^{(0)} = 0$$
$$\hat{y}_i^{(1)} = \hat{y}_i^{(0)} + f_1(x_i)$$
$$\hat{y}_i^{(2)} = \hat{y}_i^{(1)} + f_2(x_i)$$
$$\vdots$$

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_t(x_i) \quad (3)$$

The XGBoost algorithm uses the stepwise forward additive model as the gradient boosting algorithm. The difference is that the gradient boosting algorithm is a negative gradient that learns a weak learner to approximate the loss function. The XGBoost algorithm first finds the second-order Taylor approximation of the loss function at that point, and then minimizes the approximation loss function to train the weak learner. Therefore, the objective function can be expressed as

$$L^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) \quad (4)$$

Using the second-order Taylor expansion, the following function can be obtained

$$L^{(t)} = \sum_{i=1}^{n} [g_i f_t(x_i) + 0.5 h_i f_t^2(x_i)] + \Omega(f_t) \quad (5)$$

3.2.2. Random forest

RF is an integrated learning method for classification and regression (Liaw and Wiener, 2002). It is one of the representative of ensemble learning, and it is an additive model based on bagging algorithm. Different from bagging, when constructing each tree, RF uses a random sample predictor before each node segmentation, which can reduce bias. It has the following characteristics: a) The introduction of two randomness makes RF not easy to fall into overfitting, and has excellent noise immunity; b) It can process data of high dimension (many features) without feature selection; c) It has fast training speed and is easy to be parallelized, so it is relatively simple to implement (Wu et al., 2019). More details about RF can be found in the literature (Liaw and Wiener, 2002).

3.3. Data denoising method: CEEMDAN

Short-term water quality may be affected by factors such as temperature, industrial wastewater discharge and so on, so that the data may have large fluctuations in the time series and exhibit a high degree of nonlinear characteristics, which undoubtedly increases the difficulty of prediction. Therefore, many scholars use EMD, SVD, EEMD, wavelet decomposition, and other methods to extract feature values. Although these methods can improve prediction accuracy to some extent, they all have some limitations (see Table 2). For example, EMD is prone to mode mixing, while wavelet decomposition is often not ideal in some practical problems.

CEEMDAN is based on EEMD by adding a limited number of adaptive white noise (Torres et al., 2011; Zhou et al., 2019). Its implementation process is as follows:

(1) Add a white noise sequence to the raw signal to generate a noisy signal set

$$d^i(t) = d(t) + \varepsilon_0 wn^i(t), i = 1,2,\cdots I \quad (6)$$

(2) EMD decomposition operation is performed on the signal set to obtain

$$\overline{IMF_1}(t) = I^{-1}\sum_{i=1}^{I} IMF_1^i(t) \quad (7)$$

(3) Calculate the margin signal of the first stage ($k = 1$)

$$r_1(t) = d(t) - \overline{IMF_1}(t) \quad (8)$$

(4) Calculate the second modal component

$$\overline{IMF_2}(t) = I^{-1}\sum_{i=1}^{I} E_1\{r_1(t) + \varepsilon_1 E_1[wn^i(t)]\} \quad (9)$$

(5) For the following stages, calculate the *k*-th margin signal in the same way

$$r_k(t) = r_{k-1}(t) - \overline{IMF_k}(t) \quad (10)$$

(6) Calculate the $(k+1)$-th modal component

$$\overline{IMF_{k+1}}(t) = I^{-1}\sum_{i=1}^{I} E_1\{r_k(t) + \varepsilon_k E_k[wn^i(t)]\} \quad (11)$$

(7) Repeat Eq. (10) until the residual component no longer satisfies the decomposition condition. Finally, the original signal $d(t)$ can be expressed as

$$d(t) = \sum_{i=1}^{K} \overline{IMF_i}(t) + R(t) \quad (12)$$

4. Prediction process and error metrics

4.1. Prediction process

(1) Data decomposition

CEEMDAN is used for data decomposition and denoising, so that raw data with large fluctuations is decomposed into multiple datasets with less fluctuations. In other words, the data in the same dataset has more obvious similar features, as can be seen from Appendix 1.

(2) Data normalization

In order to eliminate the dimensional influence of the data indicators, the data after the decomposition is normalized and limited to the range of [0,1], the equation is

$$z_n = \frac{z_i - z_{min}}{z_{max} - z_{min}} \quad (13)$$

(3) Divide data into the training set and test set

In this paper, all decomposed datasets are divided into the training set and test set, and their ratios are 9:1. The sliding window length is 7, that is, the data of the first 6 hours are used to predict the next one.

(4) Prediction

XGBoost and RF are used to do the prediction, the prediction results are summarized, then denormalize the summarized data using Eq. (14) to get the ultimate result, as shown in Fig.1.

$$f_i = f_n(z_{max} - z_{min}) + z_{min} \quad (14)$$

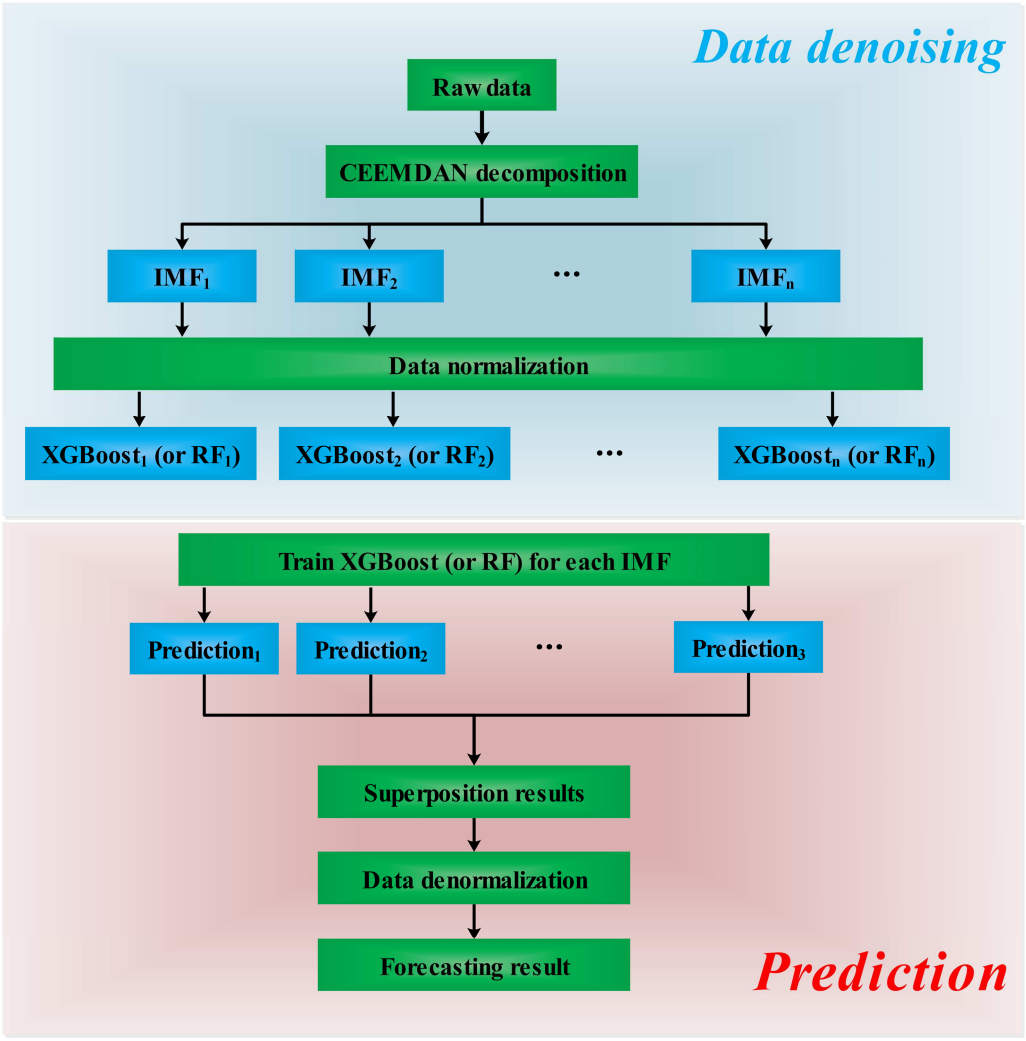


Fig.1. Data denoising process and prediction.

4.2. Error metrics

In this paper, six error metrics are used to evaluate the prediction performance, and their expressions are shown in Eqs. (15)-(20). Among them, MAE, RMSE, MAPE, RMSPE are four common error evaluation indicators (Ma et al., 2019), and the smaller their values, the smaller the error. U1 and U2 respectively represent prediction accuracy and prediction quality. The smaller the value, the higher the prediction accuracy and the better the prediction quality. In 1982, Lewis rated the prediction performance based on MAPE. The MAPE less than 10% can be considered as "excellent", the MAPE between 10% and 20% can be evaluated as "good", and the prediction performance is "reasonable" when the MAPE is in the range of 20%-50%. If the MAPE is greater than 50%, the prediction result is "inaccurate".

$$\mathrm{MAE}=\frac{1}{n}\sum_{t=1}^{n}\left|O_t-P_t\right| \tag{15}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n}\left(O_t - P_t\right)^2} \tag{16}$$

$$\mathrm{MAPE} = \frac{100\%}{n}\sum_{t=1}^{n}\left|\frac{O_t - P_t}{O_t}\right| \tag{17}$$

$$\mathrm{RMSPE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n}\left(\frac{O_t - P_t}{O_t}\right)^2} \tag{18}$$

$$\mathrm{U1} = \frac{\sqrt{\frac{1}{n}\sum_{t=1}^{n}\left(O_t - P_t\right)^2}}{\sqrt{\frac{1}{n}\sum_{t=1}^{n}O_t^2} + \sqrt{\frac{1}{n}\sum_{t=1}^{n}P_t^2}} \tag{19}$$

$$\mathrm{U2} = \sqrt{\frac{\sum_{t=1}^{n}\left(O_t - P_t\right)^2}{\sum_{t=1}^{n}O_t^2}} \tag{20}$$

5. Results and discussions

This section presents the prediction results and error analysis results of six water quality indicators, and discusses the stability of the prediction model.

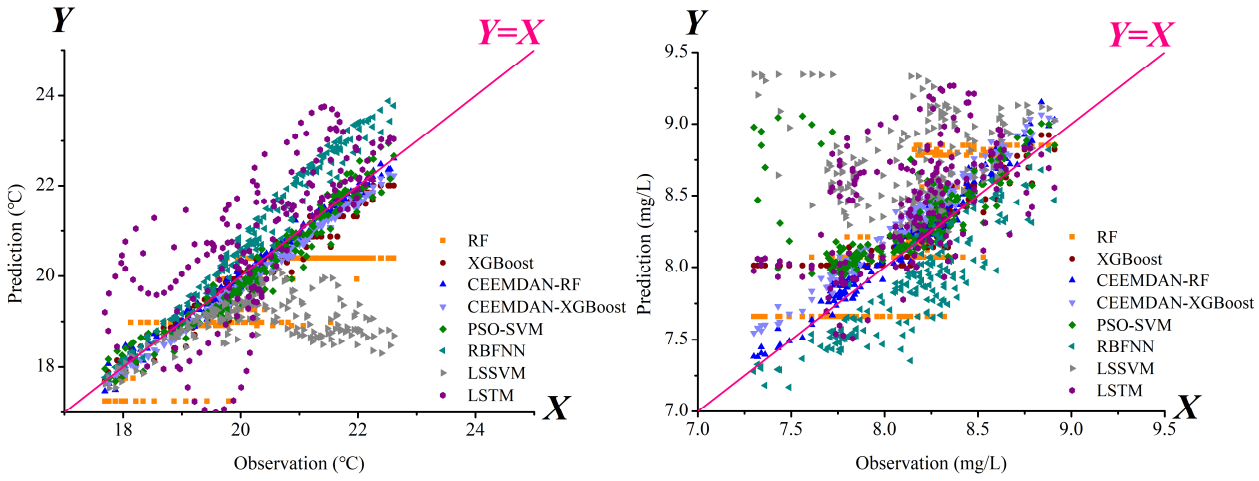
5.1. Results

Fig.2 shows the prediction results of six water quality indicators using various models, namely RF, XGboost, CEEMDAN-RF, CEEMDAN-XGBoost, PSO-SVM, RBFNN, LSSVM, and LSTM. The X-axis represents the observation value, and the Y-axis represents the prediction value. In general, the ideal prediction results will be distributed over Y=X or evenly distributed on both sides of the line, thus indicating that the error is basically obeying the Gaussian distribution. Therefore, the closer the point to the Y=X line, the smaller the error. In the prediction of the six indicators, it can be clearly seen that the points of RF, LSTM, and RBFNN are far away from Y=X. It can be seen from (d) and (f) of Fig.2 that the prediction results of some models are distributed on the same side of Y=X, which indicates that the deviation is large, and there may be problems of over-fitting or under-fitting. Table 3 lists the prediction errors of each model for each indicator. It can be concluded that CEEMDAN-RF or CEEMDAN-XGboost models have the best performance. CEEMDAN-RF performs best in the prediction of temperature, dissolved oxygen, and specific

conductance, the MAPEs are 0.69%, 1.05% and 0.90%, respectively. CEEMDAN-XGBoost performs best in the prediction of pH value, turbidity and FDOM, the MAPEs are 0.27%, 14.94% and 1.59%, respectively.

In the prediction of turbidity, the MAPEs of the eight models are generally larger. The MAPEs of RF, XGboost, CEEMDAN-RF, CEEMDAN-XGBoost, PSO-SVM, RBFNN, LSSVM and LSTM models are 18.91%, 19.13%, 18.34%, 14.94%, 44.08%, 24.55%, 21.29% and 24.22%, respectively. Moreover, we compared the average values of MAPE and RMSPE of the six water quality indicators (without considering other error metrics, because the magnitude of different water quality indicators is different, the average values for MAE, RMSE, U1, U2 are not available for reference). The results show that the average MAPEs of RF, XGboost, CEEMDAN-RF, CEEMDAN-XGBoost, PSO-SVM, RBFNN, LSSVM and LSTM models are 6.41%, 4.60%, 3.90%, 3.71%, 8.96%, 6.93%, 6.96% and 7.35%, respectively. The RMSPEs of RF, XGboost, CEDAN-XGBoost, PSO-SVM, RBFNN, SVM and SVM models are 6.41%, 4.60%, 3.90%, 3.71%, 8.96%, 6.93%, 6.96% and 7.35%, respectively. Therefore, in general, CEEMDAN-XGBoost has the best prediction performance, followed by CEEMDAN-RF.

For the prediction results of RF, CEEMDAN-RF, XGBoost, and CEEMDAN-XGBoost, the prediction results of CEEMDAN-RF are better than RF, and the improvement effect of prediction performance is noticeable. The average MAPE of CEEMDAN-RF is 38.10% lower than RF. However, some error indicators of CEEMDAN-XGBoost are not as good as XGBoost. For example, in the prediction of temperature and dissolved oxygen, the MAPEs of XGBoost are lower than that of CEEMDAN-XGBoost. However, on the whole (comprehensive consideration of the results of the six water quality indicators), the prediction performance of CEEMDAN-XGBoost is better than that of XGBoost because its average MAPE is 19.35% lower than XGBoost.



(a) (b)

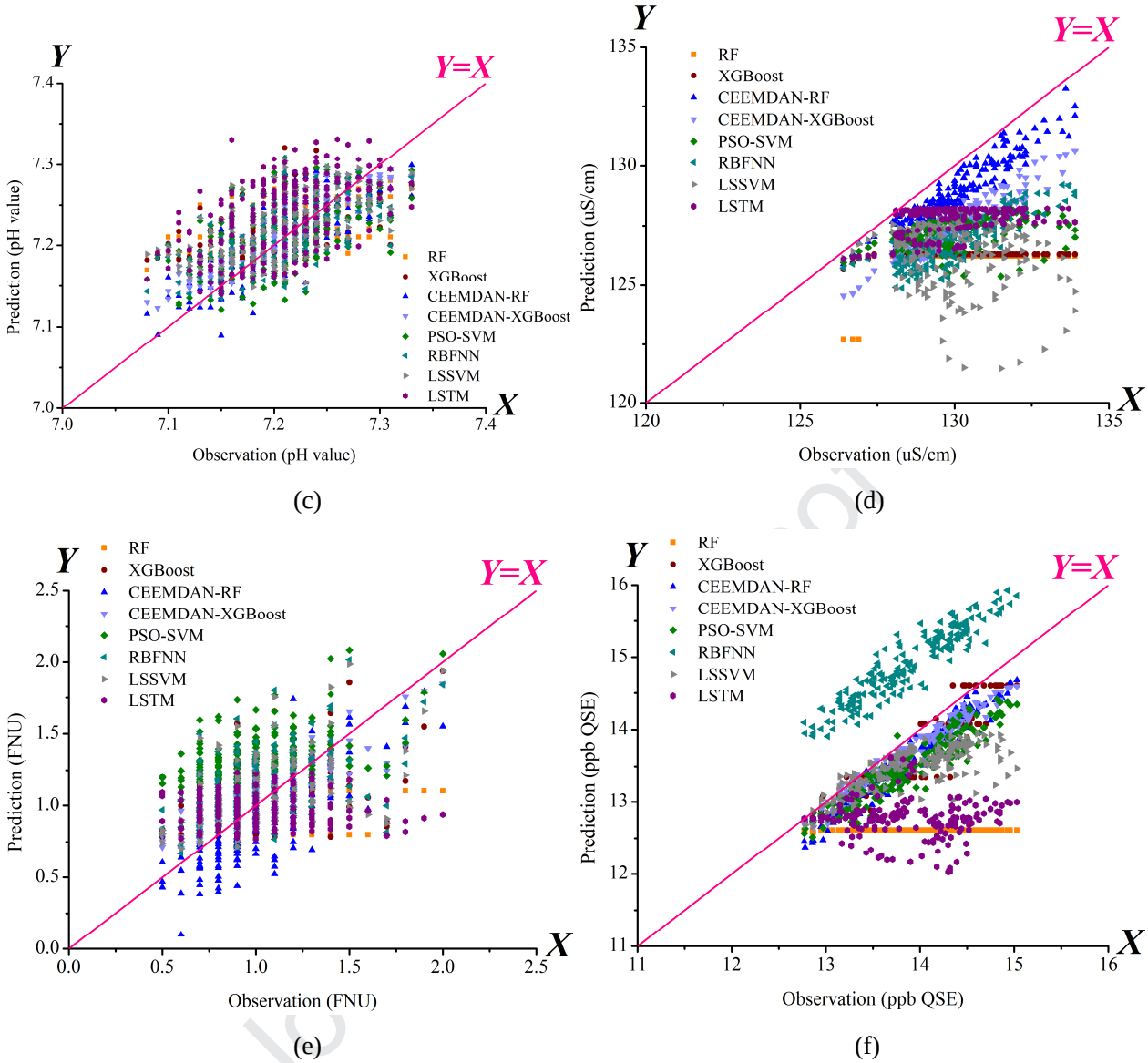


Fig.2. Water quality prediction results (test set). (a) Temperature; (b) Dissolved oxygen; (c) pH value; (d) Specific conductance; (e) Turbidity; (f) FDOM.

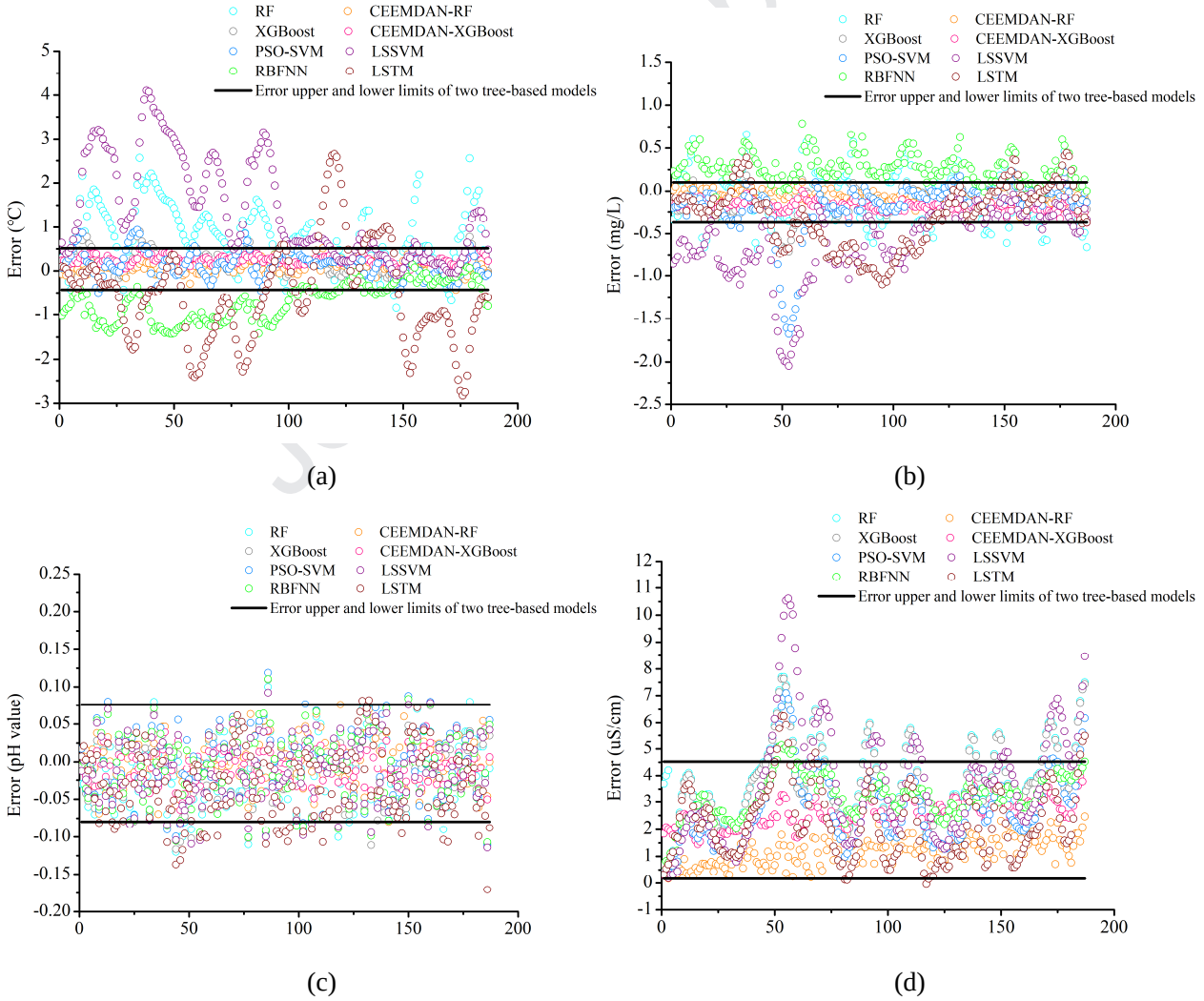
Table 3. Water quality prediction errors.

Indicator	Models	Error metrics						Indicator	Models	Error metrics					
		MAE	RMSE	MAPE (%)	RMSPE (%)	U1	U2			MAE	RMSE	MAPE (%)	RMSPE (%)	U1	U2
Temperature (Unit: □)	RF	0.85	1.04	4.08	4.97	0.026	0.051	Specific conductance (uS/cm)	RF	3.83	4.08	2.94	3.12	0.016	0.031
	XGBoost	0.22	0.29	1.07	1.38	0.007	0.014		XGBoost	3.72	3.99	2.85	3.04	0.016	0.031
	CEEMDAN-RF	**0.14**	**0.17**	**0.69**	**0.86**	**0.004**	**0.008**		CEEMDAN-RF	**1.17**	**1.27**	**0.90**	**0.98**	**0.005**	**0.010**
	CEEMDAN-XGBoost	0.24	0.26	1.18	1.26	0.006	0.013		CEEMDAN-XGBoost	2.59	2.65	1.99	2.03	0.010	0.020
	PSO-SVM	0.27	0.34	1.31	1.66	0.008	0.017		PSO-SVM	2.74	3.06	2.10	2.33	0.012	0.024
	RBFNN	0.69	0.81	3.30	3.82	0.020	0.040		RBFNN	3.22	3.32	2.47	2.55	0.013	0.026
	LSSVM	1.36	1.77	6.43	8.21	0.045	0.087		LSSVM	3.72	4.32	2.85	3.29	0.017	0.033
	LSTM	0.98	1.23	4.93	6.23	0.030	0.060		LSTM	2.21	2.60	1.69	1.97	0.010	0.020
Dissolved oxygen(Unit: mg/L)	RF	0.25	0.30	3.03	3.72	0.019	0.037	Turbidity (FNU)	RF	0.20	0.28	18.91	24.05	0.143	0.2669
	XGBoost	0.15	0.21	1.96	2.72	0.013	0.026		XGBoost	0.18	0.24	19.13	25.10	0.112	0.2270
	CEEMDAN-RF	**0.09**	**0.10**	**1.05**	**1.26**	**0.006**	**0.013**		CEEMDAN-RF	0.18	0.23	18.34	22.76	0.114	0.2186
	CEEMDAN-XGBoost	0.19	0.20	2.30	2.42	0.012	0.024		CEEMDAN-XGBoost	**0.13**	**0.16**	**14.94**	**19.18**	**0.075**	**0.1534**
	PSO-SVM	0.22	0.38	2.76	5.08	0.023	0.047		PSO-SVM	0.38	0.42	44.08	54.09	0.176	0.4063
	RBFNN	0.27	0.31	3.34	3.83	0.020	0.039		RBFNN	0.22	0.28	24.55	32.54	0.127	0.2683
	LSSVM	0.61	0.74	7.62	9.55	0.044	0.091		LSSVM	0.20	0.26	21.29	28.04	0.122	0.2478
	LSTM	0.38	0.48	4.67	5.96	0.029	0.058		LSTM	0.24	0.31	24.22	31.01	0.152	0.2936
pH value	RF	0.04	0.05	0.54	0.65	0.003	0.006	FDOM (ppb QSE)	RF	1.26	1.37	8.97	9.65	0.052	0.099
	XGBoost	0.04	0.05	0.52	0.65	0.003	0.006		XGBoost	0.29	0.33	2.07	2.38	0.012	0.024
	CEEMDAN-RF	**0.02**	0.03	0.33	0.41	0.002	0.004		CEEMDAN-RF	0.29	0.32	2.11	2.27	0.011	0.023
	CEEMDAN-XGBoost	**0.02**	**0.02**	**0.27**	**0.33**	**0.001**	**0.003**		CEEMDAN-XGBoost	**0.22**	**0.25**	**1.59**	**1.77**	**0.009**	**0.018**
	PSO-SVM	0.04	0.04	0.51	0.62	0.003	0.006		PSO-SVM	0.42	0.47	3.02	3.30	0.017	0.034
	RBFNN	0.04	0.04	0.50	0.61	0.003	0.006		RBFNN	1.02	1.04	7.39	7.54	0.036	0.075
	LSSVM	0.04	0.04	0.51	0.62	0.003	0.006		LSSVM	0.44	0.56	3.08	3.90	0.021	0.041
	LSTM	0.05	0.06	0.65	0.80	0.004	0.008		LSTM	1.12	1.28	7.95	8.98	0.048	0.092

Note: Bold represents the data with best performance in the current dataset.

5.2. Discussions

Error analysis results can only obtain the overall performance of the prediction. However, short-term water quality prediction sometimes requires accurate results at various time points. Some models have different characteristics in different datasets. For example, some models have difficulty ensuring the accuracy of time series with mutation points. Therefore, the stability of prediction is particularly critical. In this paper, the SDE is used as the criterion to evaluate the prediction stability of each model. Fig.3 shows the prediction error of each water quality indicator. The black lines represent the error upper and lower limits of CEEMDAN-XGBoost model and CEMDAN-RF model. It can be seen that many errors of the other six models lie outside the two error ranges. Moreover, Fig.4 shows the SDEs of the prediction results of various water quality indicators. It can be seen that the SDEs of CEEMDAN-RF and CEEMDAN-XGBoost are smaller than the other models, indicating that the stability of the two novel models is better.



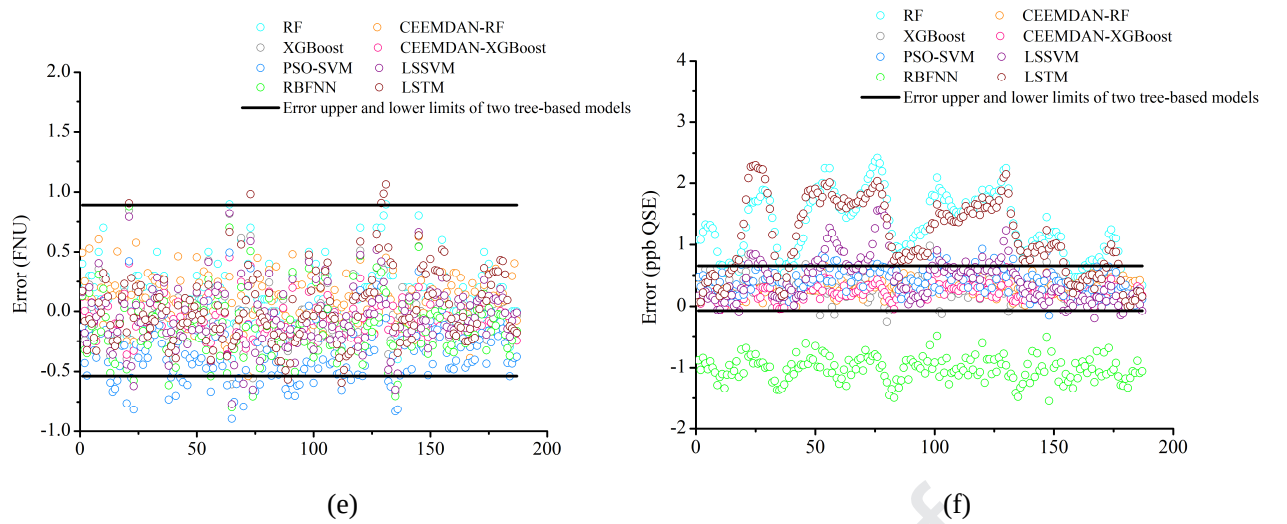


Fig.3. Water quality prediction error (test set). (a) Temperature; (b) Dissolved oxygen; (c) pH value; (d) Specific conductance; (e) Turbidity; (f) FDOM.

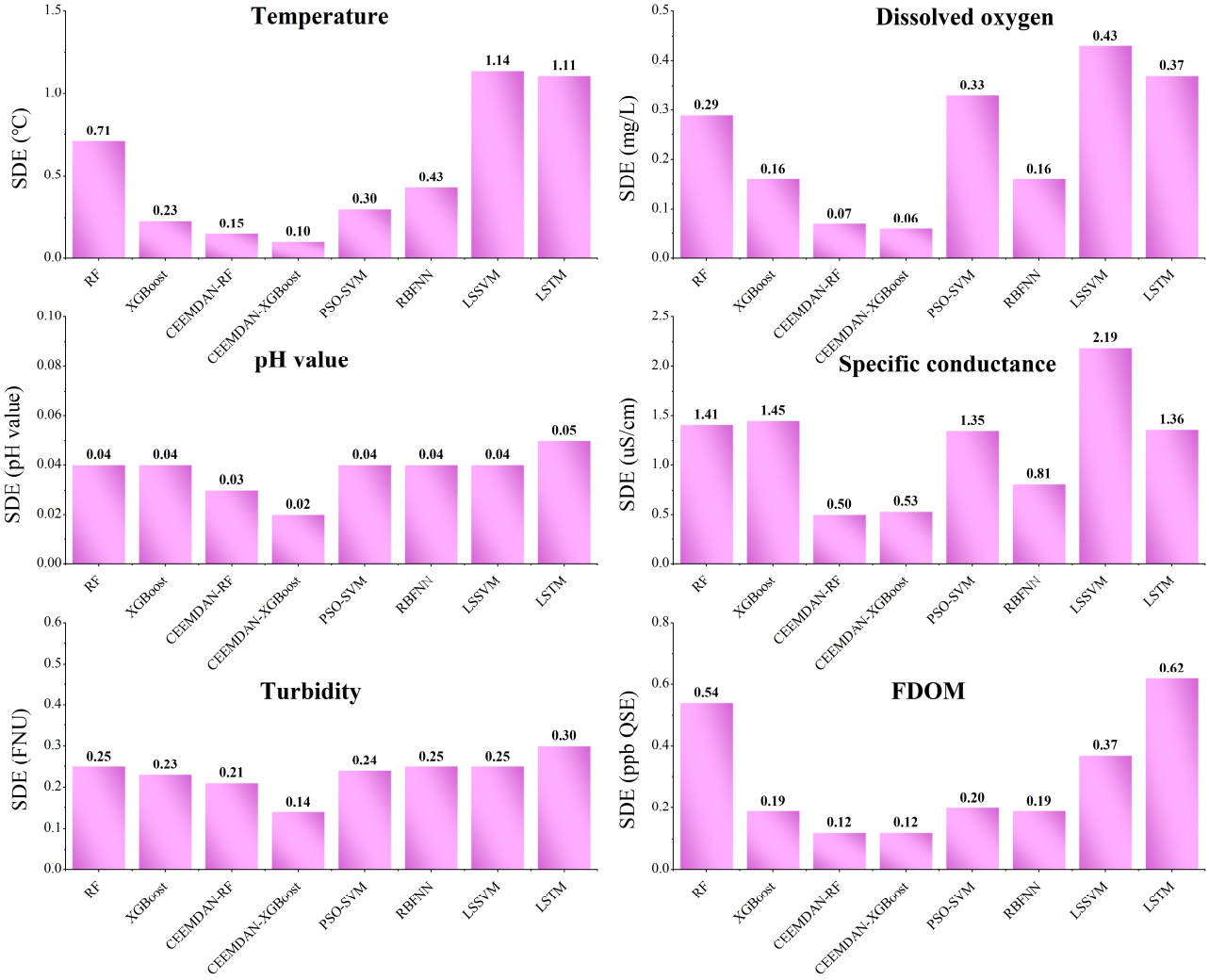


Fig.4. SDEs for water quality indicators.

6. Conclusions and future works

This paper proposes two hybrid decision tree-based models (CEEMDAN-XGBoost and CEEMDAN-RF) for the water quality prediction. The CEEMDAN in the two models is used to decompose the raw data with large fluctuations, so that the prediction performance of XGBoost and RF can be better. This paper takes the water quality of the Gales Creek site of Tualatin River in Oregon, USA as the research object, collects the data from May 1st to July 20th, 2019, and divides the raw data into training sets and test sets according to the ratio of 9:1. Two models are used to predict water temperature, dissolved oxygen, pH value, specific conductance, turbidity, and FDOM, and the prediction results are compared with other benchmark models. Besides, this paper takes SDE as an evaluation index to analyze the stability of the model. The results indicate:

a) CEEMDAN-RF performs best in the prediction of temperature, dissolved oxygen and specific conductance, the MAPEs are 0.69%, 1.05% and 0.90%, respectively. CEEMDAN-XGBoost performs best in the prediction of pH value, turbidity and FDOM, the MAPEs are 0.27%, 14.94% and 1.59%, respectively.

b) The average MAPEs of RF, XGboost, CEEMDAN-RF, CEEMDAN-XGBoost, PSO-SVM, RBFNN, LSSVM and LSTM models are 6.41%, 4.60%, 3.90%, 3.71%, 8.96%, 6.93%, 6.96% and 7.35%, respectively. Therefore, in general, CEEMDAN-XGBoost has the best predictive performance, followed by CEEMDAN-RF.

c) The SDEs of CEEMDAN-RF and CEMDAN-XGBoost is smaller than other benchmark models in predicting each water quality indicator, which implies that the stability of the two new models is better.

Although the models proposed in this paper can already achieve high prediction accuracy, the following aspects can be considered in future research: (1) Consider other factors affecting water quality in the model, rather than purely time series issues; (2) Due to the high demand of short-term prediction on computing time, parallel computing (Li et al., 2019) can be considered in the future.

Declaration of interest

None

Acknowledgments

This article is funded by Open Fund of State Key Laboratory of Oil and Gas Reservoir Geology and Exploitation (Southwest Petroleum University) (PLN201710), National Natural Science Foundation

406 of China (71901184), Humanities and Social Science Fund of Ministry of Education of China
407 (19YJCZH119), and China Scholarship Council (201708030006).

408

409 **References**

410 Chan, S. N., Thoe, W., & Lee, J. H. W. (2013). Real-time forecasting of Hong Kong beach water quality by 3D
411 deterministic model. Water research, 47(4), 1631-1647.
412 Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd
413 acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794). ACM.
414 Dabrowski, J. J., Rahman, A., Pagendam, D. E., & George, A. (2020). Enforcing mean reversion in state space
415 models for prawn pond water quality forecasting. Computers and Electronics in Agriculture, 168, 105120.
416 Gao, G., Xiao, K., & Chen, M. (2019). An intelligent IoT-based control and traceability system to forecast and
417 maintain water quality in freshwater fish farms. Computers and Electronics in Agriculture, 166, 105013.
418 García-Alba, J., Bárcena, J. F., Ugarteburu, C., & García, A. (2019). Artificial neural networks as emulators of
419 process-based models to analyse bathing water quality in estuaries. Water research, 150, 283-295.
420 Graf, R., Zhu, S., & Sivakumar, B. (2019). Forecasting river water temperature time series using a wavelet–neural
421 network hybrid modelling approach. Journal of Hydrology, 578, 124115.
422 Hounslow, A. (2018). Water quality data: analysis and interpretation. CRC press.
423 Huang, P., Trayler, K., Wang, B., Saeed, A., Oldham, C. E., Busch, B., & Hipsey, M. R. (2019). An integrated
424 modelling system for water quality forecasting in an urban eutrophic estuary: The swan-canning estuary
425 virtual observatory. Journal of Marine Systems, 103218.
426 Hussein, A. M., Elaziz, M. A., Wahed, M. S. A., & Sillanpää, M. (2019). A new approach to predict the missing
427 values of algae during water quality monitoring programs based on a hybrid moth search algorithm and the
428 random vector functional link network. Journal of Hydrology, 575, 852-863.
429 Kerr, J. G., Zettel, J. P., McClain, C. N., & Kruk, M. K. (2018). Monitoring heavy metal concentrations in turbid
430 rivers: Can fixed frequency sampling regimes accurately determine criteria exceedance frequencies,
431 distribution statistics and temporal trends?. Ecological Indicators, 93, 447-457.
432 Kong, L., & Ma, X. (2018). Comparison study on the nonlinear parameter optimization of nonlinear grey
433 Bernoulli model (NGBM (1, 1)) between intelligent optimizers. Grey Systems: Theory and Application, 8(2),
434 210-226.
435 Larsen, S. J., Kilminster, K. L., Mantovanelli, A., Goss, Z. J., Evans, G. C., Bryant, L. D., & McGinnis, D. F.
436 (2019). Artificially oxygenating the Swan River estuary increases dissolved oxygen concentrations in the
437 water and at the sediment interface. Ecological Engineering, 128, 112-121.
438 Li, F., Chen, J., & Wang, Z. (2019). Wireless MapReduce distributed computing. IEEE Transactions on
439 Information Theory, 65(10), 6101-6114.
440 Liang, Z., Zou, R., Chen, X., Ren, T., Su, H., & Liu, Y. (2020). Simulate the forecast capacity of a complicated
441 water quality model using the long short-term memory approach. Journal of Hydrology, 581, 124432.
442 Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. R news, 2(3), 18-22.

Liu, W. X., He, W., Wu, J. Y., Wu, W. J., & Xu, F. L. (2019). Effects of fluorescent dissolved organic matters (FDOMs) on perfluoroalkyl acids (PFAAs) in lake and river water. Science of the Total Environment, 666, 598-607.

Liu, Y., Zheng, Y., Liang, Y., Liu, S., & Rosenblum, D. S. (2016). Urban water quality prediction based on multi-task multi-view learning.

Lu, H., Ma, X., & Azimi, M. (2020). US natural gas consumption prediction using an improved kernel-based nonlinear extension of the Arps decline model. Energy, 116905.

Lu, H., Ma, X., Huang, K., & Azimi, M. (2020a). Carbon trading volume and price forecasting in China using multiple machine learning models. Journal of Cleaner Production, 249, 119386.

Ma, X., Mei, X., Wu, W., Wu, X., & Zeng, B. (2019). A novel fractional time delayed grey model with Grey Wolf Optimizer and its applications in forecasting the natural gas and coal consumption in Chongqing China. Energy, 178, 487-507.

Ma, X., Wu, W., Zeng, B., Wang, Y., & Wu, X. (2020). The conformable fractional grey system model. ISA transactions. DOI: 10.1016/j.isatra.2019.07.009.

Makarewicz, J. C., Lewis, T. W., Boyer, G. L., & Edwards, W. J. (2012). The influence of streams on nearshore water chemistry, Lake Ontario. Journal of Great Lakes Research, 38, 62-71.

Meyers, G., Kapelan, Z., & Keedwell, E. (2017). Short-term forecasting of turbidity in trunk main networks. Water research, 124, 67-76.

Mosley, L. M., Peake, B. M., & Hunter, K. A. (2010). Modelling of pH and inorganic carbon speciation in estuaries using the composition of the river and seawater end members. Environmental Modelling & Software, 25(12), 1658-1663.

Nobre, J., & Neves, R. F. (2019). Combining principal component analysis, discrete wavelet transform and XGBoost to trade in the financial markets. Expert Systems with Applications, 125, 181-194.

Palani, S., Liong, S. Y., & Tkalich, P. (2008). An ANN application for water quality forecasting. Marine Pollution Bulletin, 56(9), 1586-1597.

Panidhapu, A., Li, Z., Aliashrafi, A., & Peleato, N. M. (2020). Integration of weather conditions for predicting microbial water quality using Bayesian Belief Networks. Water Research, 170, 115349.

Peng, Z., Hu, W., Liu, G., Zhang, H., Gao, R., & Wei, W. (2019). Development and evaluation of a real-time forecasting framework for daily water quality forecasts for Lake Chaohu to Lead time of six days. Science of The Total Environment, 687, 218-231.

Singh, K. P., Basant, A., Malik, A., & Jain, G. (2009). Artificial neural network modeling of the river water quality—a case study. Ecological Modelling, 220(6), 888-895.

Tao, Y., Wang, Y., Rhoads, B., Wang, D., Ni, L., & Wu, J. (2020). Quantifying the impacts of the Three Gorges Reservoir on water temperature in the middle reach of the Yangtze River. Journal of Hydrology, 582, 124476.

Torres, M. E., Colominas, M. A., Schlotthauer, G., & Flandrin, P. (2011, May). A complete ensemble empirical mode decomposition with adaptive noise. In 2011 IEEE international conference on acoustics, speech and signal processing (ICASSP) (pp. 4144-4147). IEEE.

Valdivia-Garcia, M., Weir, P., Graham, D. W., & Werner, D. (2019). predicted Impact of Climate Change on trihalomethanes Formation in Drinking Water treatment. Scientific reports, 9(1), 9967.

West, D., & Dellana, S. (2011). An empirical analysis of neural network memory structures for basin water quality forecasting. International Journal of Forecasting, 27(3), 777-803.

Wu, L. F., Li, N., & Zhao, T. (2019). Using the seasonal FGM (1, 1) model to predict the air quality indicators in Xingtai and Handan. Environmental Science and Pollution Research, 26(14), 14683-14688.

Wu, L., Huang, G., Fan, J., Zhang, F., Wang, X., & Zeng, W. (2019). Potential of kernel-based nonlinear extension of Arps decline model and gradient boosting with categorical features support for predicting daily global solar radiation in humid regions. Energy conversion and management, 183, 280-295.

Xie, M., Wu, L., Li, B., & Li, Z. (2020). A novel hybrid multivariate nonlinear grey model for forecasting the traffic-related emissions. Applied Mathematical Modelling, 77, 1242-1254.

Yang, W., Wang, J., Niu, T., & Du, P. (2020). A novel system for multi-step electricity price forecasting for electricity market management. Applied Soft Computing, 88, 106029.

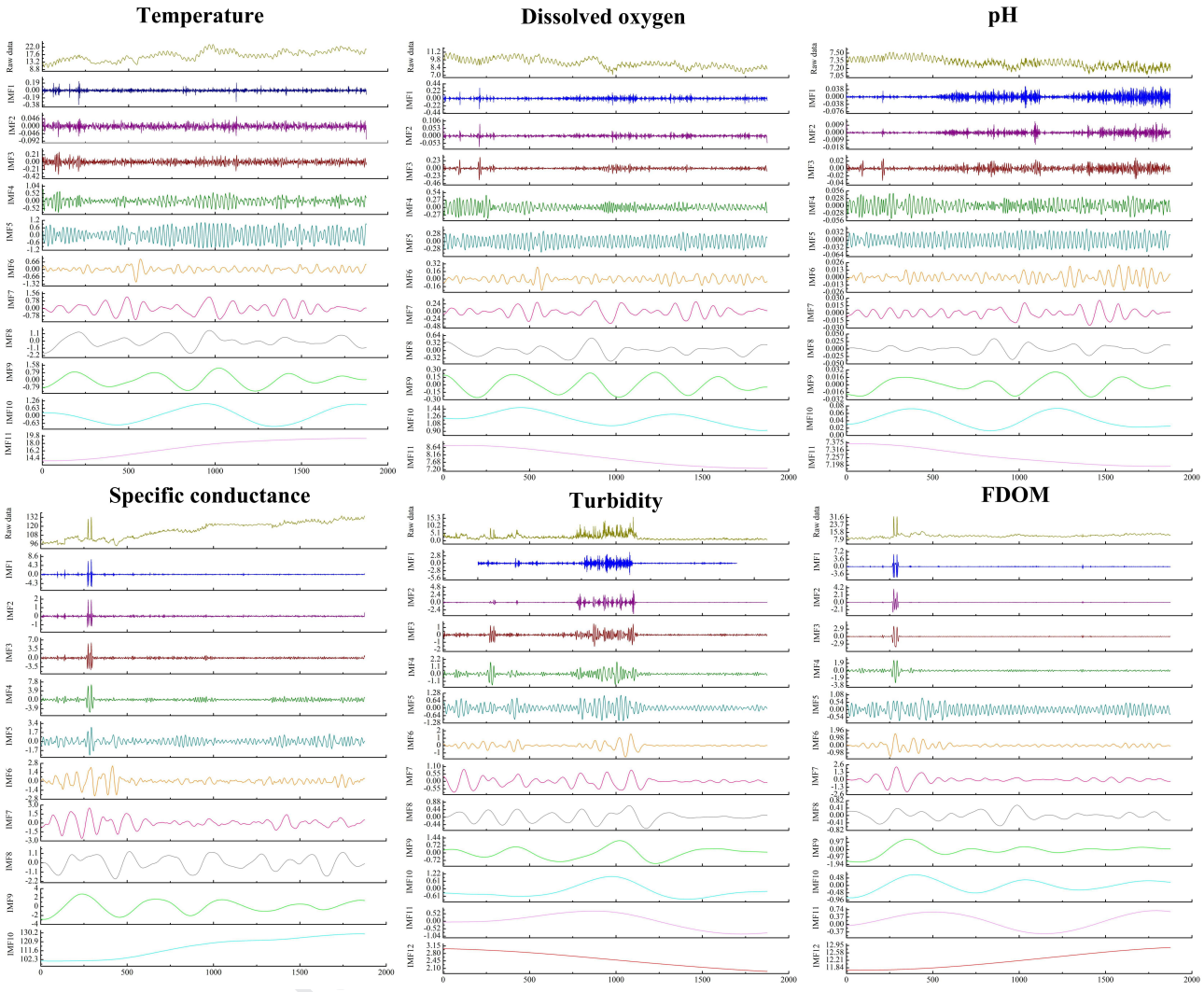
Zhang, P., Ma, X., & She, K. (2019). A Novel Power-Driven Grey Model with Whale Optimization Algorithm and Its Application in Forecasting the Residential Energy Consumption in China. Complexity, 2019.

Zhao, J., Wang, J., Guo, Z., Guo, Y., Lin, W., & Lin, Y. (2019). Multi-step wind speed forecasting based on numerical simulations and an optimized stochastic ensemble method. Applied Energy, 255, 113833.

Zhao, Y., Nan, J., Cui, F. Y., & Guo, L. (2007). Water quality forecast through application of BP neural network at Yuqiao reservoir. Journal of Zhejiang University-Science A, 8(9), 1482-1487.

Zhou, L., Xu, C., Yuan, Z., Lu, T. (2019). Dam Deformation Prediction Based on CEEMDAN-PSR-KELM Model. Yellow River, 41(6), 138-142.

Appendix



Appendix 1. Decomposition of raw data by CEEMDAN.

Highlights

1. Two hybrid decision tree-based models are proposed to predict the water quality.

2. An advanced denoising method is used to preprocess raw data.

3. The case study was conducted on the most polluted river Tualatin River in Oregon, USA.

4. The prediction stability of the model is analyzed.

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: