Chinese Journal of Chemical Engineering 87 (2025) 405–417

Contents lists available at ScienceDirect

Chinese Journal of Chemical Engineering

journal homepage: www.elsevier.com/locate/CJChE

Full Length Article

Prediction of wastewater treatment plant influent quality based on discrete wavelet transform and convolutional enhanced transformer

Lili Ma [1], Danxia Li [1,*], Jinrong He [1], Zhirui Niu [2], Zhihua Feng [3]

[1] *School of Mathematics and Computer Science, Yan'an University, Yan'an 716000, China*
[2] *School of Petroleum and Environment Engineering, Yan'an University, Yan'an 716000, China*
[3] *Yan'an CECEP Sewage Treatment Co., Ltd., Yan'an 716000, China*

A R T I C L E I N F O

Article history:
Received 20 February 2025
Received in revised form
28 April 2025
Accepted 6 June 2025
Available online 19 August 2025

Keywords:
Wastewater treatment plant
Influent quality prediction
Discrete wavelet transform
Transformer
Local feature
Long-term dependencies

A B S T R A C T

Accurate prediction of wastewater treatment plants (WWTPs) influent quality can provide valuable decision-making support to facilitate operations and management. However, since existing methods overlook the data noise generated from harsh operations and instruments, while the local feature pattern and long-term dependency in the wastewater quality time series, the prediction performance can be degraded. In this paper, a discrete wavelet transform and convolutional enhanced Transformer (DWT-CeTransformer) method is developed to predict the influent quality in WWTPs. Specifically, we perform multi-scale analysis on time series of wastewater quality using discrete wavelet transform, effectively removing noise while preserving key data characteristics. Further, a tightly coupled convolutional-enhanced Transformer model is devised where convolutional neural network is used to extract local features, and then these local features are combined with Transformer's self-attention mechanism, so that the model can not only capture long-term dependencies, but also retain the sensitivity to local context. In this study, we conduct comprehensive experiments based on the actual data from a WWTP in Shaanxi Province and the simulated data generated by BSM2. The experimental results show that, compared to baseline models, DWT-CeTransformer can significantly improve the prediction performance of influent COD and NH_3-N. Specifically, MSE, MAE, and RMSE improve by 78.7%, 79.5%, and 53.8% for COD, and 79.4%, 70.2%, and 54.5% for NH_3-N. On simulated data, our method shows strong improvements under various weather conditions, especially in dry weather, with MSE, MAE, and RMSE for COD improving by 68.9%, 48.0%, and 44.3%, and for NH_3-N by 78.4%, 54.8%, and 53.2%.

1. Introduction

With the acceleration of urbanization and industrialization, the volume and pollution level of wastewater discharge have been increasing, which results in higher demands on the operational efficiency of wastewater treatment plants (WWTPs) [1]. Accurate prediction of influent quality in WWTPs can deeply optimize their daily operational processes [2,3]. By conducting a detailed analysis of influent characteristics, treatment strategies can be flexibly adjusted to enhance treatment efficiency and effectively reduce costs, such as precisely controlling the aeration intensity, optimizing the sludge return ratio, and reasonably arranging the dosage of chemicals [4]. Furthermore, accurate predictions can guarantee the stable compliance of effluent quality, provide early warnings, and prevent potential equipment failures [5,6]. Currently, the various influent quality prediction methods proposed by researchers can be broadly classified into two categories: mathematical methods and data-driven methods [7].

Mathematical methods simulate the complex interactions among surface runoff, precipitation, sewer transport, and pollutant migration by constructing mechanistic hydrological–hydraulic coupled models, aiming to enable dynamic prediction of influent quality [8]. These methods employ continuity equations, momentum equations, and advection–diffusion–reaction equations for pollutants to characterize the spatiotemporal coupling and evolution of fluids and

* Corresponding author.
E-mail address: ldx@yau.edu.cn (D. Li).

https://doi.org/10.1016/j.cjche.2025.06.028
1004-9541/

contaminants [9]. Although such models provide clear physical interpretability and a solid theoretical basis, they still face considerable challenges in practical engineering applications. On one hand, model development depends on domain expertise and often involves idealized assumptions about system conditions [10]. On the other hand, these models require high-quality data and many parameters that are difficult to obtain under complex and dynamic conditions. Moreover, parameter identification and calibration are time-consuming, which hinders the ability to support real-time monitoring and control. This limitation is particularly evident when dealing with strong temporal variability, nonlinearity, and multivariable coupling.

Data-driven methods, with their powerful data processing capabilities, build models using numerous sensor monitoring parameters without the need for in-depth analysis of complex biological, chemical, and physical mechanisms. They are increasingly becoming the mainstream approach for predicting influent quality [11,12]. Influenced by living standards, climatic environment, and other factors, the influent quality often presents high fluctuations in inflow rate, pollutant concentration, chemical environment, and hydraulic conditions [13–15]. Nowadays, the remarkable adaptive learning ability of machine learning and artificial neural networks (ANNs) makes them the most widely applied water quality indicator monitoring models [16]. Ref. [17] effectively predicted chemical oxygen demand (COD) and biochemical oxygen demand (BOD) in wastewater by tuning the parameters of the random forest algorithm. Ref. [18] proposed a hybrid model combining extreme learning machine (ELM) with the bat algorithm to develop a predictive framework for BOD concentrations. Ref. [19] designed a fully connected network with only one hidden layer to predict the concentration of *E. coli* in WWTPs, which can highlight more importance of nonlinearities and interactions in WWTPs compared with other linear regression models. Ref. [20] derived a novel L1 norm-based ELM to overcome the problem of over-fitting in the prediction of key quality variables in the process industry. However, when dealing with complex problems with strong coupling and high nonlinearity, machine learning and shallow ANNs show limited feature representation ability and insufficient generalization ability, which may lead to the insufficient prediction accuracy of these models in practical application.

In recent years, numerous studies have demonstrated that deep neural networks can capture more complex feature patterns through multi-level abstraction and representation learning compared with shallow neural networks [21,22]. Ref. [23] proposed a hybrid model based on multimodal data and ensemble deep learning for predicting influent wastewater quality, outperforming five recurrent neural network (RNN) based reference models. Ref. [24] developed a long short-term memory (LSTM) based neural network model for detecting time-delayed water quality indicators in influent, achieving 46.20%% to 61.22% higher R^2 than models based on gated recursive unit (GRU), RNN, extreme gradient boosting (XGBoost), and random sample consistency (RANSAC). Ref. [25] proposed a residual LSTM-based model for predicting influent wastewater quality, showing superior level and directional accuracy compared to baseline models. Ref. [26] constructed a hybrid deep learning framework integrating graph convolution network (GCN), convolutional neural network (CNN), GRU and attention mechanisms for predicting industrial wastewater influent quality, successfully capturing high-dimensional features and nonlinear patterns in the data. However, the deep neural network based on recurrent structures faces two major challenges: one is the problem of gradient disappearance, which hinders the network from effectively transmitting error signals in the deep layer; second, due to the limitation in capturing the long-distance time correlation in time series data, it is difficult for the model to capture the global dependency [27].

Recently, Transformer models have shown great potential in time series prediction tasks, such as water quality forecasting in WWTPs, due to their ability to extract long-term temporal correlations and support parallel computing [28]. Ref. [29] introduced Transformer network to construct a soft sensor model of the key effluent index in WWTPs, and the effectiveness and feasibility of the method were verified on the benchmark simulation model. Ref. [30] constructed Transformer-based model to predict the influent conductivity in WWTPs, which obtained a RMSE of 155.2 $\mu S\cdot cm^{-1}$. The aforementioned work effectively characterizes the data's nonlinearity and long-term temporal correlations. However, as most of other time series, the influent wastewater quality has the multiscale property, which means the subsequences exhibit distinct patterns in different time scales. In addition, under the joint effects of environmental vibrations, temperature changes, humidity impacts, and electromagnetic interference, data generated by sensors operating in harsh detection environments typically contain noise. This noise can not only increase the complexity of data processing and modeling but also reduce the accuracy and stability of model predictions.

Therefore, in this work, in order to effectively remove noise from wastewater quality time series data and accurately capture its inherent local and global dependencies, we design an innovative method for predicting the influent quality of WWTPs, named DWT-CeTransformer (discrete wavelet transform and convolutional enhanced Transformer). The key contributions of this study are as follows.

(1) We utilize discrete wavelet transform (DWT) to isolate and diminish noise in the time series through the steps of decomposition, thresholding, and reconstruction, thereby enhancing the quality of the data and laying the foundation for subsequent predictive modeling.
(2) We construct a tightly coupled convolutional-enhanced Transformer (CeTransformer) architecture to capture both long-term and short-term temporal dependencies in the time series data after denoising with DWT. In the CeTransformer, the enhanced multi-head self-attention (EMSA) mechanism we designed extracts local features via depth-wise convolution and captures global dependencies through multi-head attention, thereby enhancing the precision of feature extraction. Meanwhile, the depth-wise separable convolution (DWSConv) layer we designed further strengthens the model's ability to capture complex features.
(3) We perform experimental validation on both actual data and the BMS2 [31] based simulated data. The results demonstrate that the proposed model significantly outperforms the baseline models in terms of prediction performance, substantially improving the accuracy of COD and ammonia nitrogen (NH_3-N) predictions.

2. Materials and Methods

2.1. Data analysis

The study is verified in a WWTP, which is located in Yan'an City, Shaanxi Province, China. This WWTP has the capacity to treat 25000 m^3 of municipal wastewater daily, and the treatment process it adopted is A2O + magnetic coagulation + aerobic filter + ultrafiltration membrane.

Table 1
Statistics of WWTP influent data for the study site.

Parameter	Unit	Min	Max	Median	Mean	Std	CV
COD	$mg\cdot L^{-1}$	45.2	555.6	206.7	214.1	52.8	0.25
NH_3-N	$mg\cdot L^{-1}$	14.3	60.1	38.1	37.9	7.0	0.18
TN	$mg\cdot L^{-1}$	10.9	78.5	45.0	45.2	10.5	0.23
TP	$mg\cdot L^{-1}$	0.29	11.6	4.63	4.88	1.74	0.35
pH	$mg\cdot L^{-1}$	6.8	8.5	7.7	7.7	0.27	0.03
SS	$mg\cdot L^{-1}$	52.9	5855.5	304.2	394.0	377.4	0.96
Influent flow	$m^3\cdot h^{-1}$	1304.2	3858.1	2678.2	2631.1	464.7	0.17

In this study, we collect three types of online quality monitoring data including the influent data, treatment process data, and effluent data to predict the quality of influent. For influent, there are 7 on-line parameters including COD^I, NH_3-N^I, total nitrogen (TN^I), total phosphorus (TP^I), pH^I, suspended solids (SS^I), and influent flow. For effluent, there are 6 on-line parameters including COD^E, TN^E, TP^E, pH^E, SS^E, and effluent flow. There are also five types of on-line monitoring devices for treatment process, including 4 dissolved oxygen (DO) meters, 8 oxidation–reduction potential (ORP) meters, 4 mixed liquor suspended solids meters (MLSS), 4 NH_3-N meters, and 4 nitrate nitrogen (NO_3-N) meters in the biological reaction tanks. All these on-line monitoring data are collected at half-hourly interval from April 1, 2022, to June 30, 2023. The statistical characteristics of the collected influent data are presented in Table 1. It can be seen from the table that there are considerable variations in influent concentrations of the WWTP in terms of both water quantity and water quality.

2.2. Data preprocessing

Data preprocessing is a crucial step in enhancing the quality of data prior to the construction of a prediction model [32]. Firstly, we remove consecutive missing values resulting from equipment maintenance or accidental events. Secondly, the 3σ principle is used to remove abnormal values. Thirdly, the individual missing values that account for 2.43% of the whole data are supplemented using the k-nearest neighbors (KNN) method:

$$x_e^* = \frac{\sum_{n=1}^{K} x_{e-n} + \sum_{n=1}^{K} x_{e+n}}{2K} \tag{1}$$

where x_e^* represents the estimated value for the missing data point, x_{e-n} and x_{e+n} represent the n-th data points before and after the missing data point, respectively. Finally, to accelerate model convergence and enhance performance, the min-max normalization method is applied to scale data into a consistent range, as defined in Eq. (2).

$$x_t = \frac{x_t - \min(\boldsymbol{X})}{\max(\boldsymbol{X}) - \min(\boldsymbol{X})} \tag{2}$$

where t is the time step, x_t is the indicator value at t, and $\max(\boldsymbol{X})$ and $\min(\boldsymbol{X})$ are its maximum and minimum values, respectively.

2.3. Correlation analysis

We employ mutual information (MI) and random forest (RF) to assess the correlation between feature variables, and integrate the results of these two methods to determine the final set of key features. The number of key features is adjusted according to the specific prediction objectives and the practical requirements of the model, ensuring that the feature set can be flexibly adapted to various prediction tasks, thereby ensuring the accuracy and robustness of the model's predictions. The calculation formula for quantifying the correlation between each feature variable and the target variable using MI is defined as:

$$\mathrm{MI}_{x,y} = \sum_{(x,y)\in X\times Y} p(x,y)\lg\left(\frac{p(x,y)}{p(x)p(y)}\right) \tag{3}$$

where $p(x,y)$ represents the joint probability distribution of the two variables, while $p(x)$ and $p(y)$ represent their respective marginal probability distributions. The calculation formula for quantifying the importance of a feature variable using RF is defined as:

$$I(x_i) = \frac{1}{n_{\text{trees}}} \sum_{t=1}^{n_{\text{trees}}} \Delta\mathrm{MSE}(t, x_i) \tag{4}$$

where n_{trees} represents the number of trees in RF, and $\Delta\mathrm{MSE}(t, x_i)$ denotes the reduction in MSE when the feature variable is used for splitting in the t-th tree. In this study, the influent COD and NH_3-N serve as predictive indicators in the developed model. To identify their key influencing factors, the correlations between COD^I and NH_3-N^I with other feature variables are separately analyzed using MI and RF, as illustrated in Fig. 1. After the above correlation analysis, the key feature sets ultimately used for predicting influent COD and NH_3-N are {PH^I, COD^E, TN^E, COD^I, influent flow, TP^E } and {TN^I, TP^I, COD^I, NH_3-N^I, NO_3-N, SS^I }, respectively. The input data of the prediction method is described as a matrix $\boldsymbol{X}' = (x'_1, x'_2, \cdots, x'_t, \cdots)$, $x'_t \in \boldsymbol{R}^D$ and $t\in[1, N]$. Here D denotes the

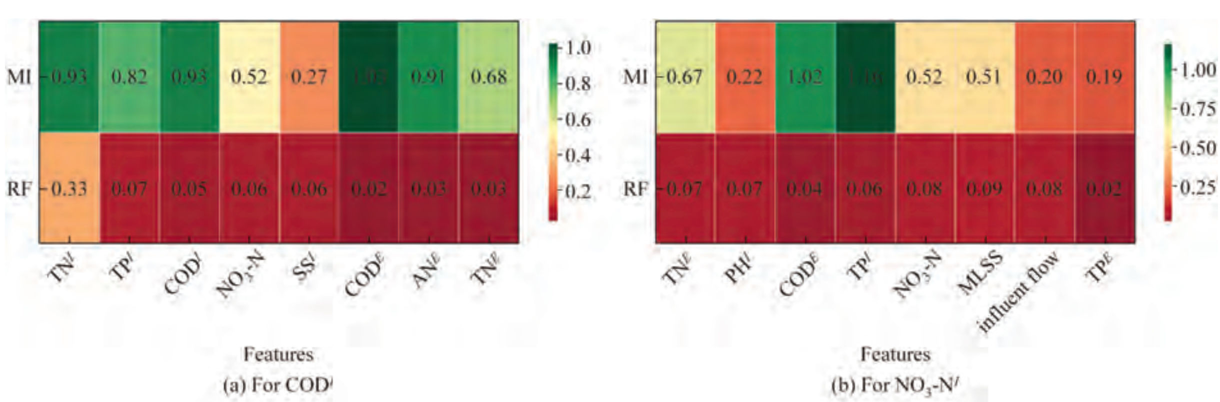

Fig. 1. The correlation of COD^I and NH_3-N^I with other feature variables is separately calculated using MI and RF.

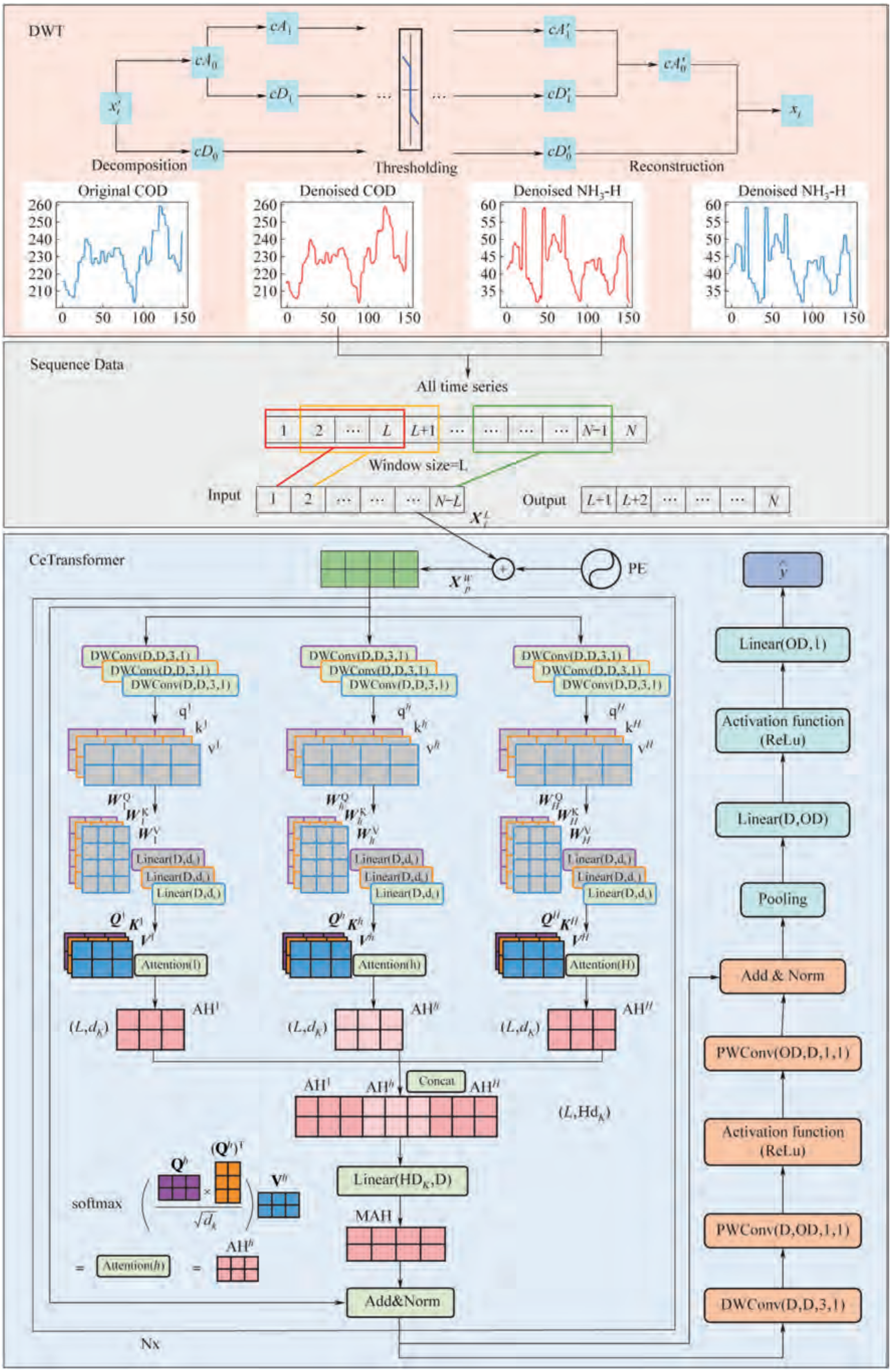

Fig. 2. Overall architecture of the proposed DWT-CeTransformer method.

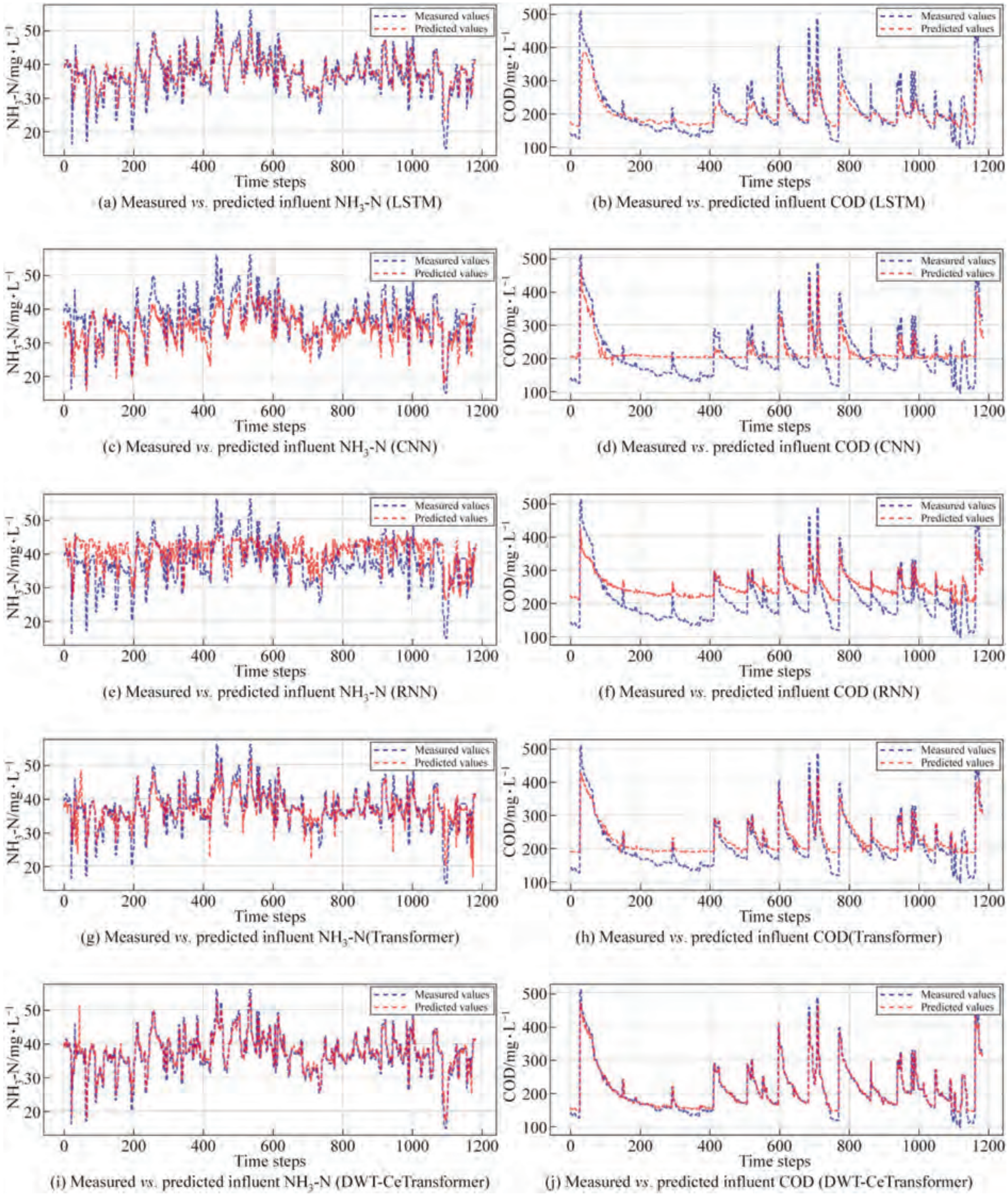

Fig. 3. Comparison of predicted vs measured influent COD and NH_3-N based on actual data.

dimensionality of the features and N denotes the number of the time steps.

2.4. Overall architecture of DWT-CeTransformer

Fig. 2 illustrates the overall architecture of the proposed DWT-CeTransformer approach for predicting influent wastewater quality. The DWT-CeTransformer is comprised of two essential phases: the data denoising phase with DWT and the predictive modeling phase with CeTransformer.

In the denosing phase, the denoising process for wastewater quality time series can be refined into three key steps: decomposition, thresholding, and reconstruction. Firstly, in the decomposition step, the original data x'_t is decomposed into two main components: a approximation coefficient cA_0 and a detail coefficient cD_0. Then, cA_0 serves as the input for the next iteration, continuing the decomposition process, which generates new approximation coefficients cA_i and detail coefficients cD_i at each iteration, until the preset decomposition level is reached. Secondly, in the thresholding step, the detail coefficients cD_i are processed with a threshold to identify and suppress noise components while maximizing the retention of useful information in the wastewater quality time series. Lastly, in the reconstruction step, the thresholded approximation coefficients cA_i' and detail coefficients cD_i' are recombined to restore the denoised data sequence.

As shown in Fig. 2, the input of the prediction model CeTransformer is a matrix X_i^L obtained from the denoised historical time series X' using a sliding window. $X_i^L = (x_{t-L+1}, x_{t-L+2}, \ldots, x_t)$, where $x_t \in \boldsymbol{R}^D$, $t \in [1, N]$, $i \in [1, N - L]$, and L denotes the size of sliding window.

In the modeling phase, CeTransformer comprises three core components: positional encoding (PE), enhanced multi-head self-attention (EMSA), and depth-wise separable convolution (DWSConv). The role of PE is to enable the model to capture and understand the order and relative positional relationships of various time points in the wastewater quality sequence, which is crucial for predicting future influent quality. By encoding positional information into vectors and adding them to the input time series X_i^L to obtain X_P^W, which can be served as the input to the model. Our proposed EMSA mechanism can employ depth-wise convolution to extract local features from the input data, and then use multiple attention heads to focus on different parts of the input sequence, thereby enhancing both the precision of feature extraction and the ability to capture global dependencies. In the DWSConv layer, we replace the traditional feed-forward network (FFN) in the Transformer with DSConv to receive the output from the EMSA layer and perform nonlinear transformations, thereby further enhancing the model's ability to capture more complex and rich features.

2.5. Data denoising with DWT

DWT is a mathematical transform for discrete time domains. Traditional Fourier transform, using global sinusoidal basis functions, cannot capture local variations in non-stationary signals [33]. Kalman filtering, limited by linear Gaussian assumptions, cannot track nonlinear fluctuations [34]. In contrast, DWT's adaptive time-frequency decomposition retains long-term trends while suppressing transient noise, addressing both non-stationarity and nonlinearity [35]. DWT produces a variety of types based on differing basis functions, with daubechies wavelet, symlets wavelet, and coiflets wavelet being some of the more common ones [36]. After numerous trials, we adopt the coiflets wavelet to perform denoising on the time series of wastewater quality. DWT relies on two fundamental functions: the scaling function and the wavelet function, and it can improve computational efficiency by discretizing the scaling and shifting operations. Both the scaling function and the wavelet function utilize a scaling factor a that is typically a power of 2, $a = 2^i$, which is employed to analyze data at various frequency levels. The translation factor b is commonly an integer multiple of a, $b = k \cdot a$, and is used to analyze data at different temporal positions. The scaling function, realized by a low-pass filter, and the wavelet function, realized by a high-pass filter, are mathematically expressed as:

$$\phi_{i,k}(t) = 2^{i/2}\phi\left(2^i t - k\right) \tag{5}$$

$$\psi_{i,k}(t) = 2^{i/2}\psi\left(2^i t - k\right) \tag{6}$$

where i represents the scale parameter, which controls the reduction of frequency; k denotes the position parameter, which controls the temporal shift.

(1) **Decomposition:** In DWT, the time series of wastewater quality is analyzed through a layer-by-layer decomposition approach. This decomposition process is carried out iteratively. DWT consists of a maximum of $\log_2 N$ levels, thus the scale parameter $i \in [0, \lfloor \log_2 N \rfloor - 1]$, and the position parameter $k \in [0, 2^i - 1]$. The sequence data is decomposed into two parts at each level: approximation coefficients and detail coefficients. Approximation coefficients are obtained by convolving the sequence data with the scaling function, representing the low-frequency part, that is, the main trend and smooth variations of the data. Detail coefficients are obtained by convolving the sequence data with the wavelet function, capturing the high-frequency part of the data, which is the details or rapid changes. Specifically, for the j-th level, the approximation coefficients $cA_{i,k}$ are obtained by convolving the original data x_t' with the scaling function $\phi_{i,k}(t)$, as shown in Eq. (7). Correspondingly, the detail coefficients $cD_{i,k}$ for the j-th layer are obtained by convolving the original data x_t' with the wavelet function $\psi_{i,k}(t)$, as depicted in Eq. (8).

$$cA_{i,k} = \sum_{t=0}^{N_j - 1} x_t' \cdot \phi_{j,k}(t) \tag{7}$$

$$cD_{i,k} = \sum_{t=0}^{N_j - 1} x_t' \cdot \psi_{j,k}(t) \tag{8}$$

where $N_j = N/2^i$ denotes the length of the original time series after j downsamplings.

(2) **Thresholding:** The time series of wastewater quality typically exhibit sparsity in the wavelet domain, which means that most detail coefficients are close to zero, while those caused by noise are more evenly distributed. By applying thresholding to the detail coefficients, noise can be effectively suppressed [37]. In this study, we adopt the soft thresholding method, with the specific steps as follows: first, calculate the median absolute deviation (MAD) value of the detail coefficients as an estimate of the noise level. The MAD is defined as:

$$\mathrm{MAD} = \mathrm{median}\left(\left|cD_{i,k} - \mathrm{median}\left(cD_{i,k}\right)\right|\right) \tag{9}$$

Next, the threshold is determined as $\lambda = 1.4826 \times \mathrm{MAD}$, where the constant factor 1.4826 corresponds to the 75th percentile of the standard normal distribution. Finally, soft thresholding is applied to each detail coefficient $cD_{i,k}$ as shown in Eq. (10). This process effectively suppresses noise while preserving the essential characteristics of the time series.

$$\mathrm{cD}_{i,k}' = \begin{cases} \mathrm{sgn}\left(\mathrm{cD}_{i,k}\right)\left(|\mathrm{cD}_{i,k}| - \lambda\right), & \text{if } |\mathrm{cD}_{i,k}| \geq \lambda \\ 0, & \text{if } |\mathrm{cD}_{i,k}| < \lambda \end{cases} \tag{10}$$

(3) Reconstruction:

Reconstruction is the process of transforming the time series of wastewater quality from wavelet domain to time domain. After I-level wavelet decomposition and thresholding, a coefficient list $\boldsymbol{H} = [cA_{I-1}, cD'_{I-1}, cD'_{I-2}, \ldots, cD'_1, cD'_0]$ is generated. And the reconstruction of the time series data at level I is calculated as:

$$x_t = \sum_k \mathrm{cA}_{I-1,k}\phi_{I-1,k}(t) + \sum_{i=0}^{I-1}\sum_k \mathrm{cD}'_{i,k}\psi_{i,k}(t) \tag{11}$$

Here, the coefficients in $\boldsymbol{H}$ are upsampled (zero-padding) and subsequently processed through convolution with the corresponding wavelet functions, ultimately resulting in the denoised and reconstructed original data x'_t as x_t.

2.6. Predictive modeling with CeTransformer

In this section, the three core components of the CeTransformer model will be introduced in sequence.

(1) **Positional Encoding (PE):** PE embeds positional information into the input time series data, enabling the model to recognize the temporal order of the data and capture temporal correlations at different positions within the sequence, thereby more effectively processing time-dependent wastewater quality sequences [38].

This paper employs sinusoidal positional encoding, which calculates the positional encoding by applying sine and cosine functions of different frequencies to the positions in the sequence. Specifically, the positional encoding consists of sine and cosine functions with varying frequencies, with the formulas as follows:

$$\mathrm{PE}_{\mathrm{pos},2j} = \sin\left(\frac{\mathrm{pos}}{10000^{2j/D}}\right) \tag{12}$$

$$\mathrm{PE}_{\mathrm{pos},2j+1} = \cos\left(\frac{\mathrm{pos}}{10000^{\frac{2j+1}{D}}}\right) \tag{13}$$

where pos denotes the position index in the sequence matrix $\boldsymbol{X}_i^L$ and j denotes the dimension index in the feature dimension D.

(2) **Enhanced multi-head self-attention (EMSA):** As shown in Fig. 2, the computation of each attention head in the EMSA mechanism is independent and parallel, and the process is as follows. First, by applying depth-wise convolution (DWConv) operations to the input sequence X_P^W, intermediate feature vector $\boldsymbol{q}^h$, $\boldsymbol{k}^h$ and $\boldsymbol{v}^h$ are obtained, and their calculation formulas are given by

$$\boldsymbol{q}^h = \mathrm{DWConv}_{1\times3}\left(X_P^W\right) \tag{14}$$

$$\boldsymbol{k}^h = \mathrm{DWConv}_{1\times3}\left(X_P^W\right) \tag{15}$$

$$\boldsymbol{v}^h = \mathrm{DWConv}_{1\times3}\left(X_P^W\right) \tag{16}$$

DWConv independently applies a convolutional kernel to each feature dimension, which can effectively extract local features while simultaneously reduce the number of model parameters and computational load.

Second, by performing linear transformations on the feature vectors $\boldsymbol{q}^h$, $\boldsymbol{k}^h$ and $\boldsymbol{v}^h$, the query matrix $\boldsymbol{Q}^h$, key matrix $\boldsymbol{K}^h$, and value matrix $\boldsymbol{V}^h$ are derived, with the corresponding calculation formulas provided below:

$$\boldsymbol{Q}^h = \boldsymbol{q}^h\boldsymbol{W}_h^Q, \quad \boldsymbol{K}^h = \boldsymbol{k}^h\boldsymbol{W}_h^K, \quad \boldsymbol{V}^i = \boldsymbol{v}^h\boldsymbol{W}_h^V \tag{17}$$

where $\boldsymbol{W}_h^Q, \boldsymbol{W}_h^K, \boldsymbol{W}_h^V \in \mathbb{R}^{D\times d_k}$ are the weight matrices. Linear transformations enhance the model's expressive power and computational efficiency by providing differentiated feature representations for $\boldsymbol{Q}^h$, $\boldsymbol{K}^h$, and $\boldsymbol{V}^h$.

Third, the similarity matrix $\boldsymbol{S}^h$ between $\boldsymbol{Q}^h$ and $\boldsymbol{K}^h$ is calculated through the dot product:

$$\boldsymbol{S}^h = \begin{pmatrix} \boldsymbol{Q}_1^h \\ \boldsymbol{Q}_2^h \\ \vdots \\ \boldsymbol{Q}_L^h \end{pmatrix} \left(\left(\boldsymbol{K}_1^h\right)^{\mathrm{T}}, \left(\boldsymbol{K}_2^h\right)^{\mathrm{T}}, \cdots, \left(\boldsymbol{K}_L^h\right)^{\mathrm{T}}\right) = \left(\boldsymbol{Q}_i^h\left(\boldsymbol{K}_j^h\right)^{\mathrm{T}}\right)_{ij} \tag{18}$$

where the attention score $\boldsymbol{S}^h$ is divided by the scaling factor $\sqrt{d_k}$ to prevent numerical instability, and then the softmax function is applied to generate the attention weights matrix:

$$\boldsymbol{A}^h = \mathrm{Softmax}\left(\frac{1}{\sqrt{d_k}}\boldsymbol{S}^h\right) \tag{19}$$

As shown in Eq. (20), $\boldsymbol{A}^h$ is used to perform a weighted sum on the value matrix $\boldsymbol{V}^h$, yielding the output of the h-th attention head:

$$\boldsymbol{AH}^h = \begin{pmatrix} \boldsymbol{A}_{11}^h & \boldsymbol{A}_{12}^h & \ldots & \boldsymbol{A}_{1L}^h \\ \boldsymbol{A}_{21}^h & \boldsymbol{A}_{22}^h & \ldots & \boldsymbol{A}_{2L}^h \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{A}_{L1}^h & \boldsymbol{A}_{L2}^h & \ldots & \boldsymbol{A}_{LL}^h \end{pmatrix} \begin{pmatrix} \boldsymbol{V}_1^h \\ \boldsymbol{V}_2^h \\ \vdots \\ \boldsymbol{V}_L^h \end{pmatrix} = \begin{pmatrix} \sum_{j=1}^{L} \boldsymbol{A}_{1j}^h\boldsymbol{V}_j^h \\ \sum_{j=1}^{L} \boldsymbol{A}_{2j}^h\boldsymbol{V}_j^h \\ \vdots \\ \sum_{j=1}^{L} \boldsymbol{A}_{Lj}^h\boldsymbol{V}_j^h \end{pmatrix} \tag{20}$$

Fourth, as shown in Eq. (21), the outputs from each attention head $\mathbf{AH}^h$ are concatenated, then processed through a linear layer to yield the final multi-head attention matrix **MAH**. This process enhances the model's ability to capture a variety of dependency relationships, while preserving the original dimensionality of the input, thus offering a rich feature representation for subsequent model layers.

$$\mathbf{MAH}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \mathrm{Concat}\left(\mathbf{AH}^1, \mathbf{AH}^2, \cdots, \mathbf{AH}^H\right)\boldsymbol{W}^O \tag{21}$$

where $\boldsymbol{W}^O \in \boldsymbol{R}^{Hd_k\times D}$ is the weight matrice. The multi-head attention score **MAH** is added to the original input vector $\boldsymbol{X}_P^W$ to form a residual connection, which alleviates the vanishing gradient problem in deep networks, allowing the model to learn features more deeply. Subsequently, layer normalization (LN) is applied to the result of the residual connection to stabilize the training process and enhance the model's generalization capability. The mathematical formula for this process is expressed as follows:

Table 2
Performance comparison of baseline models and the proposed DWT-CeTransformer based on actual data.

	Model	MSE	MAE	RMSE	Training time/ms	Inference time/ms
Influent COD	RNN	0.0202	0.1261	0.1421	509	0.207
	CNN	0.0141	0.0932	0.1186	528	0.490
	LSTM	0.0125	0.0652	0.1120	3296	0.226
	Transformer	0.0089	0.0722	0.0945	34233	23.2
	DWT-CeTransformer	0.0043	0.0258	0.0657	753534	757.0
Influent NH_3-N	RNN	0.0228	0.1251	0.1509	1873	0.231
	CNN	0.0136	0.0892	0.1164	522	0.837
	LSTM	0.0092	0.0651	0.0958	3303	0.229
	Transformer	0.0087	0.0595	0.0931	45494	21.2
	DWT-CeTransformer	0.0047	0.0373	0.0686	757761	757.7

$$\boldsymbol{X}^{RL} = \mathrm{LN}\left(\boldsymbol{X}_P^W + \mathbf{MAH}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V})\right) \tag{22}$$

(3) **Depth-wise separable convolution (DWSConv):** DWSConv is a combination of DWConv and point-wise convolution (PWConv), which efficiently extracts features while reducing the number of parameters and computational load. DWConv applies a convolutional operation independently to each feature to capture its local temporal dependencies. PWConv, on the other hand, fuses and combines the sequences of features obtained from the independent convolutions to generate a richer feature representation [39]. As shown in Eq. (23), vector $\boldsymbol{X}^{RL}$ undergoes DWSConv, followed by a nonlinear transformation and PWConv, ultimately resulting in feature vector:.

$$\boldsymbol{X}^D = \mathrm{PWConv}_{1\times1}\left(\phi\left(\mathrm{PWConv}_{1\times1}\left(\mathrm{DWConv}_{1\times3}\left(X^{RL}\right)\right)\right)\right) \tag{23}$$

where ϕ deonotes the activation function (ReLu).

As shown in Eq. (24), the feature vector $\boldsymbol{X}^D$ undergoes residual connection, LN, and adaptive average pooling (AAP), and then passes through a feed-forward network (FFN) to ultimately produce the prediction output $\widehat{y}$.

$$\widehat{y} = \mathrm{FFN}\left(\left(\mathrm{AAP}\left(\mathrm{LN}\left(\boldsymbol{X}^{RL} + \boldsymbol{X}^D\right)\right)\right)\right) \tag{24}$$

where FFN consists of two linear transformation layers, with the ReLU activation function in between.

2.7. Model training, test and evaluation

To validate the performance of our proposed DWT-CeTransformer method, RNN, CNN, LSTM, and Transformer are selected as comparison models. The dataset is partitioned into training and testing subsets with a 9:1 ratio to ensure robust model evaluation. The specific configurations of each model are as follows: the RNN model employs a hidden layer size of 32; the CNN model is designed with convolutional layers that comprise 64 output channels; the LSTM model utilizes a hidden layer size of 16. Both the Transformer and DWT-CeTransformer models are configured with 4 attention heads and 2 stacked encoder layers, where the feed-forward network employs a hidden dimension of 128 with ReLU activation. All models are trained using the Adam optimizer to enhance convergence and performance.

To comprehensively evaluate the performance of each model, MSE, MAE, and RMSE are regarded as evaluation metrics in all experiments. The formulas are as follows:

$$\mathrm{MSE} = \frac{1}{S}\sum_{k=1}^{S}\left(y_k - \widehat{y}_k\right)^2 \tag{25}$$

$$\mathrm{MAE} = \frac{1}{S}\sum_{k=1}^{S}\left|y_k - \widehat{y}_k\right| \tag{26}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{S}\sum_{k=1}^{S}\left(\widehat{y}_k - y_k\right)^2} \tag{27}$$

where S represents the total number of samples, y_k is the measured value of the k-th sample, and $\widehat{y}_k$ is the predicted value of the k-th sample.

3. Results and Discussion

The inflow volume, composition, and pollutant concentration of wastewater are highly uncertain due to unpredictable factors such as weather and temperature. To verify the accuracy, stability, and robustness of the proposed DWT-CeTransformer method, we conducted experiments using actual WWTP data and simulated data from BSM2 [31]. In BSM2, influent data are categorized into dry, rainy, and stormy conditions, each with 14 days of records. Data are collected every 15 min, yielding 96 points per day and a total of 1344 sample sets over 14 days.

3.1. Based on actual data

3.1.1. Performance comparison of methods

Table 2 shows the performance comparison between DWT-CeTransformer and baseline models (RNN, CNN, LSTM, and Transformer) for influent COD and NH_3-N prediction tasks. The results indicate that DWT-CeTransformer demonstrates a significant advantage in predicting both water quality indicators, with

Table 3
Ablation results for DWT-CeTransformer based on actual data.

	Model	MSE	MAE	RMSE
Influent COD	Transformer	0.0089	0.0722	0.0945
	CeTransformer	0.0077	0.0539	0.0876
	DWT-CeTransformer	0.0043	0.0258	0.0657
Influent NH_3-N	Transformer	0.0087	0.0595	0.0931
	CeTransformer	0.0055	0.0462	0.0739
	DWT-CeTransformer	0.0047	0.0373	0.0686

Table 4
Paired t-test results (MSE and MAE) based on actual data.

	Model comparison	MSE			MAE		
		t-value	p-value	Sign	t-value	p-value	Sign
Influent COD	RNN *vs*. DWT-CeTransformer	26.5674	2.29×10^{-122}	Yes	57.2418	1.00×10^{-300}	Yes
	LSTM *vs*. DWT-CeTransformer	7.0129	3.91×10^{-12}	Yes	18.6676	2.27×10^{-68}	Yes
	CNN *vs*. DWT-CeTransformer	24.8414	5.55×10^{-110}	Yes	60.3431	1.00×10^{-300}	Yes
	Transformer *vs*. DWT-CeTransformer	12.9786	4.14×10^{-36}	Yes	38.4668	1.07×10^{-210}	Yes
Influent NH_3-N	RNN *vs*. DWT-CeTransformer	14.7964	1.40×10^{-45}	Yes	25.5562	4.41×10^{-115}	Yes
	LSTM *vs*. DWT-CeTransformer	10.1982	1.82×10^{-23}	Yes	17.5816	1.17×10^{-61}	Yes
	CNN *vs*. DWT-CeTransformer	14.1596	3.65×10^{-42}	Yes	23.5710	4.92×10^{-101}	Yes
	Transformer *vs*. DWT-CeTransformer	7.0868	2.35×10^{-12}	Yes	14.0319	1.72×10^{-41}	Yes

better performance in NH_3-N prediction. Compared to RNN, DWT-CeTransformer reduces the RMSE by 53.8% (for COD) and 54.5% (for NH_3-N), and decreases the MSE by 78.7% and 79.4%, respectively. This difference may result from the more stable characteristics of NH_3-N. As shown in Table 1, its coefficient of variation ($CV = 0.18$) is lower than that of COD ($CV = 0.25$), indicating a more concentrated and less volatile distribution, which helps the model capture trends more reliably. In contrast, COD shows greater fluctuation and noise, increasing prediction uncertainty. Moreover, the DWT-CeTransformer leverages DWT and enhanced multi-head attention to capture both global patterns and local features, maintaining strong performance under noisy conditions.

Table 2 shows that DWT-CeTransformer requires much longer training and inference times than baseline models. Training exceeds 750000 ms and inference approaches 757 ms, while most baselines train under 50000 ms and infer in less than 1 ms. For instance, in COD prediction, training and inference take 753534 ms and 757.0 ms, with CNN (528 ms) and LSTM (0.23 ms) being the fastest. The computational performance available for modeling in this study reaches 100975 GFLOPS. The increase in processing time is mainly attributed to the intensive frequency-domain multi-scale transformations involved in the DWT denoising process [40], as well as the high complexity and large number of parameters introduced by the self-attention mechanism in the CeTransformer phase. Although EMSA and DWSConv further increase the model complexity, the additional overhead compared to the base Transformer is hardly noticeable. Notably, this increase in inference time has almost no impact on the operational execution of WWTPs, as in most WWTPs using the activated sludge process, the hydraulic retention time from the inlet to the aeration tank typically exceeds 1 h [6].

From Fig. 3, it is evident that different models perform quite differently on COD and NH3-N prediction tasks. The RNN model responds slowly to steep changes and fails to capture rapid fluctuations, resulting in overly smooth predictions and poor local feature representation. The LSTM model, aided by its memory mechanism, improves global trend modeling and captures some long-term dependencies, but still lags in responding to sharp fluctuations, limiting its accuracy on local dynamics. The CNN model performs better on local features and short-term variations but lacks the ability to capture long-term dependencies and global trends. The Transformer, with multi-head attention, effectively models global trends and closely follows the overall COD and NH3-N patterns, but still struggles with fine details and rapid changes. However, the Transformer still has errors and limitations in handling details and capturing rapid changes. In contrast, the DWT-CeTransformer performs best among all models, with predictions closely matching measured values, especially in regions with complex steep changes and local features. This is due to several factors: First, DWT removes high-frequency noise while preserving key dynamics, improving stability and accuracy. For COD data with large fluctuations ($CV = 0.25$), denoising helps the model focus on essential features. Second, convolution-enhanced attention captures global trends and local dynamics. For less volatile NH_3-N data ($CV = 0.18$), the model achieves high local feature accuracy. Finally, multi-scale denoising and DWSConv ensure robust performance under rapid fluctuations, enhancing adaptability to complex time series.

3.1.2. Ablation experiment

Table 3 shows the ablation study results, comparing the performance of three models—Transformer, CeTransformer, and

Table 5
Performance of proposed DWT-CeTransformer approach compared to baseline methods based on simulated data (dry, rainy, and stormy weather).

Weather	Model	Influent COD			Influent NH_3-N			Training time/ms	Inference time/ms
		MSE	MAE	RMSE	MSE	MAE	RMSE		
Dry	RNN	0.0212	0.1061	0.1455	0.0199	0.1094	0.1410	66.7	0.245
	CNN	0.0195	0.1119	0.1397	0.0182	0.1052	0.1350	42.2	14.3
	LSTM	0.0135	0.0980	0.1160	0.0122	0.0830	0.1104	239	0.230
	Transformer	0.0093	0.0667	0.0962	0.0110	0.0826	0.1050	3615	21.9
	DWT-CeTransformer	0.0066	0.0552	0.0810	0.0043	0.0494	0.0659	54293	132.7
Rainy	RNN	0.0243	0.1222	0.1558	0.0225	0.1062	0.1499	38.4	0.288
	CNN	0.0188	0.1086	0.1372	0.0221	0.1280	0.1486	54.2	21.7
	LSTM	0.0148	0.0935	0.1216	0.0146	0.0893	0.1208	211	0.315
	Transformer	0.0123	0.0867	0.1111	0.0131	0.0848	0.1144	2930	0.541
	DWT-CeTransformer	0.0083	0.0695	0.0913	0.0075	0.0582	0.0864	74362	118.0
Stormy	RNN	0.0267	0.1230	0.1635	0.0250	0.1224	0.1581	66.2	0.223
	CNN	0.0226	0.1083	0.1502	0.0226	0.1174	0.1504	44.7	8.34
	LSTM	0.0170	0.0827	0.1303	0.0194	0.1067	0.1392	261	0.241
	Transformer	0.0136	0.0959	0.1165	0.0161	0.0913	0.1267	3877	0.601
	DWT-CeTransformer	0.0114	0.0800	0.1067	0.0097	0.0760	0.0985	73755	123.1

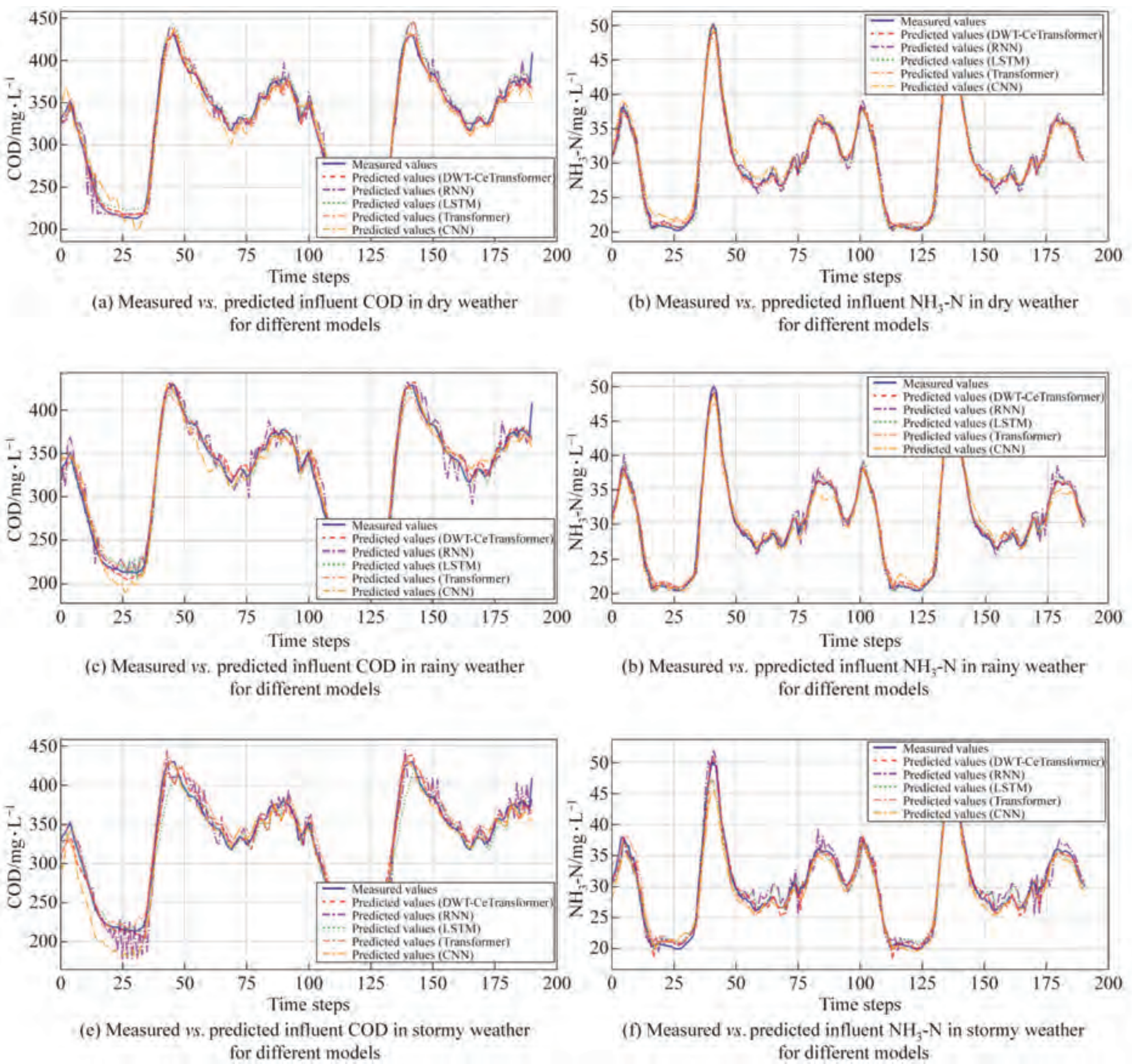

Fig. 4. Comparison of predicted vs measured influent COD and NH_3-N under different weather conditions (simulated data).

DWT-CeTransformer—on the influent COD and NH_3-N prediction tasks. After introducing DWConv and DWSConv operations, CeTransformer reduces RMSE by 7.3% and 20.6% for the COD and NH_3-N tasks, respectively, compared to Transformer. This result shows that combining DWConv and DWSConv improves the model's ability to capture local dynamic variations in time series, especially in regions with rapid fluctuations, enhancing dynamic feature extraction. The convolution mechanism models local dynamics well, enabling finer feature extraction and improving robustness against complex changes, thus boosting short-term dynamic prediction accuracy.

Table 3 shows that the RMSE of DWT-CeTransformer drops by 25.1% and 7.2% for the COD and NH_3-N prediction tasks, respectively, compared to CeTransformer. This result shows that DWT effectively filters out high-frequency noise by decomposing the time series into multiple frequency components while preserving key dynamic features. This enables the model to better capture important patterns in the data. Moreover, the decomposition not only reduces noise interference but also enhances the model's sensitivity to changes at different time scales, improving both stability and prediction accuracy.

3.1.3. *Significance analysis*

Table 4 shows that DWT-CeTransformer consistently outperforms baseline models in COD and NH_3-N prediction, with p-values for MSE and MAE mostly below 0.001. This robustness mainly comes from DWT-based frequency-domain denoising, which suppresses high-frequency noise, and the combination of DWSConv and EMSA modules, which enhances local feature capture and global temporal modeling. Specifically, in the COD prediction task, DWT-CeTransformer achieves the greatest improvement over CNN (with t-values of 60.3431 and 24.8414 for MAE and MSE, respectively). Even when compared with Transformer, although the improvement is relatively smaller, the t-values still reach 38.4668 and 12.9786, highlighting its superiority in capturing high-frequency variations and modeling complex dynamic sequences. In the NH_3-N prediction task, the model shows the most substantial enhancement over RNN (with t-values of 25.5562 and 14.7964), and maintains notable advantages compared to Transformer, with t-values of 14.0319 and 7.0868, further validating its robustness and generalization ability in modeling relatively stable sequences.

Table 6
Ablation results for DWT-CeTransformer based on simulated data (dry, rainy, and stormy weather).

Weather	Model	Influent COD			Influent NH_3-N		
		MSE	MAE	RMSE	MSE	MAE	RMSE
Dry	Transformer	0.0093	0.0667	0.0962	0.0110	0.0826	0.1050
	CeTransformer	0.0080	0.0621	0.0895	0.0082	0.0691	0.0906
	DWT-CeTransformer	0.0066	0.0552	0.0810	0.0043	0.0494	0.0659
Rainy	Transformer	0.0123	0.0867	0.1111	0.0131	0.0848	0.1144
	CeTransformer	0.0101	0.0809	0.1005	0.0115	0.0824	0.1073
	DWT-CeTransformer	0.0083	0.0695	0.0913	0.0075	0.0582	0.0864
Stormy	Transformer	0.0136	0.0959	0.1165	0.0161	0.0913	0.1267
	CeTransformer	0.0112	0.0856	0.1060	0.0152	0.0934	0.1234
	DWT-CeTransformer	0.0114	0.0800	0.1067	0.0097	0.0760	0.0985

Table 7
Paired *t*-test results for (MSE and MAE) based on simulated data (dry, rainy, and stormy weather).

Weather	Model comparison	Influent COD						Influent NH_3-N					
		MSE			MAE			MSE			MAE		
		t-value	*p*-value	sign	*t*-value	*p*-value	sign	*t*-value	*p*-value	sign	*t*-value	*p*-value	sign
Dry	RNN vs DWT-CeTransformer	6.6285	3.41×10^{-10}	Yes	9.2838	3.78×10^{-17}	Yes	6.3178	1.84×10^{-9}	Yes	9.0757	1.44×10^{-16}	Yes
	LSTM vs DWT-CeTransformer	5.0839	8.82×10^{-7}	Yes	8.3324	1.56×10^{-14}	Yes	4.6498	6.21×10^{-6}	Yes	5.8741	1.87×10^{-8}	Yes
	CNN vs DWT-CeTransformer	6.5413	5.50×10^{-10}	Yes	8.6331	2.39×10^{-15}	Yes	7.9481	1.64×10^{-13}	Yes	9.7839	1.45×10^{-18}	Yes
	Transformer vs DWT-CeTransformer	2.5883	1.04×10^{-2}	Yes	2.4636	1.46×10^{-2}	Yes	4.3428	2.29×10^{-5}	Yes	6.1969	3.50×10^{-9}	Yes
Rainy	RNN vs DWT-CeTransformer	6.3058	$1.96e \times 10^{-9}$	Yes	7.1926	1.42×10^{-11}	Yes	5.2744	3.61×10^{-7}	Yes	6.2256	3.01×10^{-9}	Yes
	LSTM vs DWT-CeTransformer	4.4875	1.25×10^{-5}	Yes	4.0627	7.09×10^{-5}	Yes	4.6951	5.09×10^{-6}	Yes	5.5464	9.67×10^{-8}	Yes
	CNN vs DWT-CeTransformer	5.2453	4.14×10^{-7}	Yes	5.6915	4.71×10^{-8}	Yes	6.6835	2.52×10^{-10}	Yes	9.5163	8.36×10^{-18}	Yes
	Transformer vs DWT-CeTransformer	3.7036	2.79×10^{-4}	Yes	3.4456	7.01×10^{-4}	Yes	3.4474	6.97×10^{-4}	Yes	5.2205	4.65×10^{-7}	Yes
Stormy	RNN vs DWT-CeTransformer	4.7269	4.43×10^{-6}	Yes	5.1612	6.15×10^{-7}	Yes	5.4188	1.80×10^{-7}	Yes	6.0226	8.71×10^{-9}	Yes
	LSTM vs DWT-CeTransformer	2.9920	3.14×10^{-3}	Yes	2.8554	4.78×10^{-3}	Yes	3.5134	5.53×10^{-4}	Yes	4.1374	5.27×10^{-5}	Yes
	CNN vs DWT-CeTransformer	3.1122	2.14×10^{-3}	Yes	2.4877	1.37×10^{-2}	Yes	4.9451	1.67×10^{-6}	Yes	5.1926	5.31×10^{-7}	Yes
	Transformer vs DWT-CeTransformer	1.2465	2.14×10^{-1}	No	2.5324	1.21×10^{-2}	Yes	2.1920	2.96×10^{-2}	Yes	2.0037	4.65×10^{-2}	Yes

3.2. *Based on simulated data*

3.2.1. *Performance comparison of methods*

Table 5 compares the prediction performance of the DWT-CeTransformer and baseline models (RNN, CNN, LSTM, and Transformer) for influent COD and NH_3-N concentrations under dry, rainy, and stormy conditions. Results show that DWT-CeTransformer consistently outperforms the baselines, particularly in NH_3-N prediction, where its errors are significantly lower. Specifically, under dry conditions, it reduces the RMSE of COD and NH_3-N by 44.3% and 53.2% compared to RNN. Under rainy and stormy conditions, COD RMSE is reduced by 41.5% and 34.8%, and NH_3-N RMSE by 42.4% and 37.7%, respectively. DWT-CeTransformer also achieves major improvements in MSE, especially under dry conditions, reducing COD MSE by 68.9% and NH_3-N MSE by 78.4%. These gains are partly due to the more stable variation of NH_3-N across weather conditions, making its trend easier to capture. In contrast, influent COD concentration fluctuates more, especially during rainy and stormy conditions, increasing prediction uncertainty. Second, DWT-CeTransformer combines DWT with convolution-enhanced attention mechanisms, enabling the model to accurately extract key features from the time series amid complex weather changes and noise interference, thereby maintaining high prediction accuracy.

Table 5 shows that the training and inference times for simulated data are both shorter than those for real-world data, mainly due to the higher quality and smaller volume of the simulated dataset. The training time of DWT-CeTransformer under various weather conditions is approximately 50000−70000 ms, and the inference time ranges from 118 to 133 ms. In contrast, most baseline models complete training in less than 100 ms and inference in less than 1 ms. For example, under highly volatile stormy conditions, the training and inference times of DWT-CeTransformer are 73755 ms and 123.1 ms, respectively, while those of RNN are 66.2 ms and 0.223 ms, respectively. Based on the previous analysis, we found that the DWT-CeTransformer has an inference time of less than 1 min on both simulated and real-world data. With such a level of time overhead, wastewater treatment plants have sufficient time to respond and adjust promptly according to the predicted influent water quality.

Fig. 4 shows clear differences in model performance for influent COD and NH_3-N predictions under varying weather conditions. Under dry conditions, with small fluctuations and low noise, the DWT-CeTransformer accurately captures both global trends and local details, especially in regions with rapid changes. The Transformer models global trends well but struggles with local details and rapid changes. LSTM fits global trends but is affected by noise, resulting in delayed responses to COD fluctuations. RNN performs poorly due to noise, leading to low local accuracy. CNN performs the worst, with large errors in COD fluctuation regions. Under rainy conditions, with increased fluctuations and noise, the DWT-CeTransformer still outperforms others, maintaining accuracy in trends and local details with minimal noise impact. The Transformer captures global trends but shows deviations in peaks and valleys due to noise. LSTM outperforms RNN and CNN but still lags in response to COD fluctuations. RNN and CNN perform poorly, especially in COD's rapidly changing regions. Under stormy conditions, with the highest fluctuations and noise, the DWT-CeTransformer remains robust, accurately capturing COD and NH_3-N dynamics. The Transformer fits trends but responds slowly to severe fluctuations. LSTM is stable for NH_3-N but lacks adaptability for COD. RNN and CNN perform worst, with large prediction errors in complex COD fluctuation regions and poor dynamic feature capture. These performance differences arise from several factors. First, DWT-CeTransformer enhances its adaptability to dynamic data by combining effective noise handling mechanisms with the integration of DWConv and DWSConv. Second, under rainy and stormy weather conditions, where data exhibits significant fluctuations and high noise levels, the model's multi-scale feature extraction mechanism enables it to capture both global trends and local details simultaneously.

3.2.2. Ablation experiment

Table 6 summarizes the ablation results, comparing the Transformer, CeTransformer, and DWT-CeTransformer models on COD and NH_3-N prediction tasks. After introducing DWConv and DWSConv, the RMSE of CeTransformer decreases by 6.9% and 25.5% (dry weather), 9.5% and 6.2% (rainy weather), and 9.0% and 2.6% (stormy weather) for COD and NH_3-N, respectively, compared to the Transformer. This improvement shows that the deep convolution mechanism strengthens the model's ability to capture local features and short-term dynamics, achieving a better balance between global trend modeling and local detail extraction. In addition, deep convolution more effectively extracts fine-grained time series features, significantly boosting short-term prediction accuracy and detail fitting.

Table 6, the RMSE of DWT-CeTransformer compared to CeTransformer decreases by 9.5% and 27.5% for COD and NH_3-N prediction tasks (under dry weather), 9.2% and 19.5% (under rainy weather), and 0.9% and 19.9% (under stormy weather). These results show that DWT effectively reduces noise by decomposing the time series into low-frequency approximations and high-frequency details while preserving key dynamic features. The multi-scale decomposition successfully filters out high-frequency noise, enhancing the model's ability to capture variations at different time scales. Especially under rainy and stormy conditions, wavelet denoising greatly improves model robustness in complex, high-noise environments.

3.2.3. Significance analysis

Table 7 shows that under Dry, Rainy, and Stormy conditions, DWT-CeTransformer consistently outperforms baseline models in predicting COD and NH_3-N, with most paired t-tests yielding p-values below 0.001, confirming its robustness to weather disturbances. This performance mainly benefits from DWT denoising, which suppresses high-frequency interference, and EMSA's deep convolution operations, which enhance local feature capture. Under Dry conditions, DWT-CeTransformer achieves the largest improvements over CNN, with t-values of 8.6331/6.5413 for COD and 9.7839/7.9481 for NH_3-N, demonstrating advantages in stable environments. Even under Rainy conditions with intensified fluctuations, it remains stable, outperforming CNN on NH_3-N prediction ($t = 9.5163$ for MAE, $t = 6.6835$ for MSE), showing adaptability to moderate disturbances. The only non-significant result appears in the comparison with the Transformer on the MSE metric for COD, where the t-value is 1.2465 with $p = 0.214$, marked as "No", suggesting that performance gains may be limited under extreme weather scenarios.

4. Conclusions and Future Work

This study proposes a novel prediction method, DWT-CeTransformer, to address the limitations of traditional models in analyzing wastewater treatment time series data. Traditional models often struggle with noise interference and fail to capture both global trends and local details. DWT enhances robustness by removing high-frequency noise through multi-scale decomposition while preserving key dynamic features. CeTransformer architecture innovatively designs the EMSA mechanism and DWSConv layers. EMSA mechanism employs depthwise convolution to extract local features while utilizing multi-head attention to model global dependencies. DWSConv layer adopts a channel separation and spatial aggregation strategy to enhance the model's ability to capture complex features, collectively optimizing network performance. Actual data from WWTPs in Yan'an, Shaanxi Province, and simulated BSM2 data under dry, rainy, and stormy conditions are used to evaluate the performance of DWT-CeTransformer. The experimental results show that DWT-CeTransformer reduces MSE by 16.2% to 79.4% compared to baseline models. Under storm conditions, its MSE values still remain steadily between 0.0097 and 0.0114. Additionally, the R^2 values for DWT-CeTransformer are 0.78 to 0.86 on actual data and 0.97 to 0.99 on simulated data.

Building on these results, future work will optimize the model for cross-regional and multi-operational scenarios by integrating long-term monitoring data and multi-source water quality features. We will explore adaptive mechanisms for dynamic wavelet adjustment and enhance generalization to influent fluctuations. Additionally, fine-tuning or even retraining with larger, more diverse datasets and developing more robust architectures will be pursued to improve resilience to data quality variations.

CRediT Authorship Contribution Statement

Lili Ma: Writing – original draft. Danxia Li: Writing – review & editing. Jinrong He: Conceptualization. Zhirui Niu: Investigation. Zhihua Feng: Data curation.

Declaration of Competing Interest

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.

Acknowledgements

This research is funded by the Natural Science Basic Research Program of Shaanxi (2024JCYBMS576), and the National Natural Science Foundation of China (62366053).

References

[1] H. Han, Y. Liu, J. Qiao, Mechanism-data-driven multiobjective optimization for wastewater treatment process, *IEEE Trans. Ind. Inf.* 20 (5) (2024) 7810–7819.
[2] S. Hamidoud, M. Bendjaballah, I. Kouadri, M.R. Makhlouf, Multi-objective optimization of wastewater treatment using electrocoagulation, *Chin. J. Chem. Eng.* 75 (2024) 152–160.
[3] M.B. Aregu, Industrial wastewater treatment efficiency of mixed substrate (pumice and scoria) in horizontal subsurface flow constructed wetland: comparative experimental study design, *Air Soil. Water Res.* 15 (2022) 11786221211063888.
[4] H. Han, Y. Zhao, X. Wu, H. Yang, Multi-timescale feature extraction method of wastewater treatment process based on adaptive entropy, *Chin. J. Chem. Eng.* 76 (2024) 264–271.
[5] W. Wei, N. Chen, Z. Zhang, Z. Liu, M. Zuo, K. Liu, Y. Xia, A scalable-bandwidth extended state observer-based adaptive sliding-mode control for the dissolved oxygen in a wastewater treatment process, *IEEE Trans. Cybern.* 52 (12) (2021) 13448–13457.
[6] Y. Zhang, J. Wang, C. Li, H. Duan, W. Wang, Attention-based deep learning models for predicting anomalous shock of wastewater treatment plants, *Water Res.* 275 (2025) 123192.
[7] J. Lv, L. Du, H. Lin, B. Wang, W. Yin, Y. Song, J. Chen, J. Yang, A. Wang, H. Wang, Enhancing effluent quality prediction in wastewater treatment plants through the integration of factor analysis and machine learning, *Bioresour. Technol.* 393 (2024) 130008.
[8] A. Elsayed, M. Ghaith, A. Yosri, Z. Li, W. El-Dakhakhni, Genetic programming expressions for effluent quality prediction: towards AI-driven monitoring and management of wastewater treatment plants, *J. Environ. Manag.* 356 (2024) 120510.

[9] D. Wang, S. Thunéll, U. Lindberg, L. Jiang, J. Trygg, M. Tysklind, N. Souihi, A machine learning framework to improve effluent quality control in wastewater treatment plants, *Sci. Total Environ.* 784 (2021) 147138.
[10] H. Zhang, C. Yang, X. Shi, H. Liu, Effluent quality prediction in papermaking wastewater treatment processes using dynamic Bayesian networks, *J. Clean. Prod.* 282 (2021) 125396.
[11] H. Han, M. Sun, F. Li, Z. Liu, C. Wang, Self-supervised deep clustering method for detecting abnormal data of wastewater treatment process, *IEEE Trans. Ind. Inf.* 20 (2) (2024) 1155–1166.
[12] K. Li, L. Zhang, J. Qiao, Multi-task stochastic configuration network with autonomous linking and its application in wastewater treatment processes, *Inf. Sci.* 662 (2024) 120195.
[13] M.A. Saleem, F. Harrou, Y. Sun, Explainable machine learning methods for predicting water treatment plant features under varying weather conditions, *Results Eng.* 21 (2024) 101930.
[14] H. Han, M. Sun, H. Han, X. Wu, J. Qiao, Univariate imputation method for recovering missing data in wastewater treatment process, *Chin. J. Chem. Eng.* 53 (2023) 201–210.
[15] N. Aryanti, A. Nafiunisa, T.D. Kusworo, Recent study on hydrophilization of polyvinylidene fluoride membrane for oily–wastewater treatment, *Chin. J. Chem. Eng.* 76 (2024) 157–186.
[16] M. Bahramian, R.K. Dereli, W. Zhao, M. Giberti, E. Casey, Data to intelligence: the role of data-driven models in wastewater treatment, *Expert Syst. Appl.* 217 (2023) 119453.
[17] S. Soni, A. Khurshid, A.M. Minase, A. Bonkinpelliwar, A tinyml approach for quantification of bod and cod in water, *Proc. Int. Conf. Paradigm Shifts Commun. Embedded Syst. Mach. Learn. Signal Process. (PCEMS)* 2 (2023) 1–6.
[18] H. Mekaoussi, S. Heddam, N. Bouslimanni, S. Kim, M. Zounemat-Kermani, Predicting biochemical oxygen demand in wastewater treatment plant using advance extreme learning machine optimized by bat algorithm, *Heliyon* 9 (11) (2023) e21351.
[19] J. Foschi, A. Turolla, M. Antonelli, Soft sensor predictor of e. coli concentration based on conventional monitoring parameters for wastewater disinfection control, *Water Res.* 191 (2021) 116806.
[20] X. Shi, Q. Kang, J. An, M. Zhou, Novel l1 regularized extreme learning machine for soft-sensing of an industrial process, *IEEE Trans. Ind. Inf.* 18 (2) (2022) 1009–1017.
[21] C. Tao, M. Jia, G. Wang, Y. Zhang, Q. Zhang, X. Wang, Q. Wang, W. Wang, Time-sensitive prediction of no2 concentration in China using an ensemble machine learning model from multi-source data, *J. Environ. Sci.* 137 (2024) 30–40.
[22] T. Cheng, F. Harrou, F. Kadri, Y. Sun, T. Leiknes, Forecasting of wastewater treatment plant key features using deep learning-based models: a case study, *IEEE Access* 8 (2020) 1–11.
[23] S. Heo, K. Nam, J. Loy-Benitez, C. Yoo, Data-driven hybrid model for forecasting wastewater influent loads based on multimodal and ensemble deep learning, *IEEE Trans. Ind. Inf.* 17 (10) (2021) 6925–6934.
[24] Y. Zhang, C. Li, H. Duan, K. Yan, J. Wang, W. Wang, Deep learning based data-driven model for detecting time-delay water quality indicators of wastewater treatment plant influent, *Chem. Eng. J.* 467 (2023) 143483.
[25] W. Zhang, J. Zhao, P. Quan, J. Wang, X. Meng, Q. Li, Prediction of influent wastewater quality based on wavelet transform and residual lstm, *Appl. Soft Comput.* 148 (2023) 110858.
[26] J. Cao, A. Xue, Y. Yang, R. Lu, X. Hu, L. Zhang, W. Cao, G. Cao, X. Geng, L. Wang, A hybrid deep learning framework for predicting industrial wastewater influent quality based on graph optimisation, *J. Water Process Eng.* 65 (2024) 105831.
[27] W. Cao, W. Qi, P. Lu, Air quality prediction based on time series decomposition and convolutional sparse self-attention mechanism transformer model, *IEEE Access* 12 (2024) 155340–155350.
[28] R. Wang, Y. Qi, Q. Zhang, F. Wen, A multi-step water quality prediction model based on the savitzky-golay filter and transformer optimized network, *Environ. Sci. Pollut. Res.* 30 (50) (2023) 109299–109314.
[29] P. Chang, S. Zhang, Z. Wang, Soft sensor of the key effluent index in the municipal wastewater treatment process based on transformer, *IEEE Trans. Ind. Inf.* 20 (3) (2024) 4021–4028.
[30] J. Pereira, P. Oliveira, M.S. Duarte, G. Martins, P. Novais, Using Deep Learning Models to Predict the Electrical Conductivity of the Influent in a Wastewater Treatment Plant, 2023, pp. 130–141.
[31] F. Manenti, G.V. Reklaitis, Big data analytics for advanced fault detection in wastewater treatment plants, *Aided Chem. Eng.* 53 (2024) 1831–1836, of Comput.
[32] L. Gu, J. Wang, J. Liu, A combined system based on data preprocessing and optimization algorithm for electricity load forecasting, *Comput. Ind. Eng.* 191 (2024) 110114.
[33] M.S. Priyadarshini, M. Bajaj, L. Prokop, M. Berhanu, Perception of power quality disturbances using fourier, short-time fourier, continuous and discrete wavelet transforms, *Sci. Rep.* 14 (2024) 3443.
[34] W. Liu, Z. Lai, K. Bacsa, E. Chatzi, Neural extended kalman filters for learning and predicting dynamics of structural systems, *Struct. Health Monit.* 23 (2024) 1037–1052.
[35] P. Khetarpal, N. Nagpal, H.H. Alhelou, P. Siano, M. Al-Numay, Noisy and non-stationary power quality disturbance classification based on adaptive segmentation empirical wavelet transform and support vector machine, *Comput. Electr. Eng.* 118 (2024) 109346.
[36] C. Dominguez-Monferrer, A. Guerra-Sancho, A. Caggiano, L. Nele, M.H. Miguélez, J.L. Cantero, Multiresolution analysis for tool failure detection in cfrp/ti6al4v hybrid stacks drilling in aircraft assembly lines, *Mech. Syst. Signal Process.* 206 (2024) 110925.
[37] X. Li, A. Wang, Forest pest monitoring and early warning using UAV remote sensing and computer vision techniques, *Sci. Rep.* 15 (1) (2025) 401.
[38] D. Zhang, J. Gao, X. Li, Multivariate time series classification with crucial timestamps guidance, *Expert Syst. Appl.* 255 (2024) 124591.
[39] Q. Yu, W. Wei, D. Li, Z. Pan, C. Li, D. Hong, Hypersinet: a synergetic interaction network combined with convolution and transformer for hyperspectral image classification, *IEEE Trans. Geosci. Rem. Sens.* 62 (2024) 1–18.
[40] Z. Hajiabotorabi, A. Kazemi, F. Samavati, F. Ghaini, Improving dwt-rnn model *via* b-spline wavelet multiresolution to forecast a high-frequency time series, *Expert Syst. Appl.* 138 (2019) 112842.