

Article

Multi-Scale Temporal Convolutional Networks for Effluent COD Prediction in Industrial Wastewater

Yun Geng ¹ , Fengshan Zhang ^{2,3}  and Hongbin Liu ^{1,2,3,*} 

¹ Jiangsu Co-Innovation Center of Efficient Processing and Utilization of Forest Resources, Nanjing Forestry University, Nanjing 210037, China; ggloiy@njfu.edu.cn

² Shandong Huatai Paper Co., Ltd., Dongying 257335, China; htjszx@163.com

³ Shandong Yellow Triangle Biotechnology Industry Research Institute Co., Ltd., Dongying 257399, China

* Correspondence: hongbinliu@njfu.edu.cn; Tel.: +86-25-85427620

Abstract: To identify the complex time patterns in the process data and monitor the effect of wastewater treatment by predicting effluent chemical oxygen demand more accurately, a soft-sensor modeling method based on the multi-scale temporal convolutional network (MSTCN) was proposed in this paper. Data at different time scales are reconstructed according to the main frequencies determined by the Fourier transform approach, and the correlations between variables during that period are calculated and stored in the corresponding adjacency matrix. The specific temporal convolutional network (TCN) is built to learn the temporal dependencies within each sequence at the current scale, while the graph convolutional layer (GCN) captures the relationships among variables. Finally, predictions with less error can be obtained by integrating output features from GCN and TCN layers. The proposed model is validated on an annual dataset collected from a wastewater treatment plant employing biological processes for organic matter removal. The experimental results indicate that the proposed MSTCN reduces RMSE by 35.71% and 22.56% compared with the convolutional neural network and TCN, respectively. Moreover, MSTCN shortens the training period by 6.3 s and improves RMSE by 30.41% when compared to the long short-term memory network, which is outperformed in extracting temporal dynamic characteristics.

Keywords: soft sensor; wastewater treatment; modeling; indicator prediction; deep learning



Citation: Geng, Y.; Zhang, F.; Liu, H. Multi-Scale Temporal Convolutional Networks for Effluent COD Prediction in Industrial Wastewater. *Appl. Sci.* **2024**, *14*, 5824. <https://doi.org/10.3390/app14135824>

Academic Editor: Marco Carnevale Miino

Received: 28 May 2024

Revised: 24 June 2024

Accepted: 2 July 2024

Published: 3 July 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

As the scale of industrial development expands, environmental problems caused by production activities have become increasingly prominent. Industrial wastewater from sectors such as metallurgy, dyeing, and pulp and papermaking contains excessive amounts of heavy metals and refractory organics, which severely pollute soil and water resources, destroy ecosystems, and lead to negative impacts on human health [1,2]. Additionally, the harmful substances in industrial wastewater can volatilize with the rising temperature, resulting in air pollution. Thus, industrial wastewater has become one of the significant global environmental challenges. Effective treatment and enhanced regulatory mechanisms are crucial for environmental protection and the sustainable development of industries [3].

It is essential to regulate specific wastewater discharge standards based on industrial categories and pollutant types to control water pollution and maintain ecological balance. Because of the large amount of wastewater discharge, the complex composition of pollutants, and poor biodegradability [4], pre-treatment is required before carrying out the biochemical treatment again for the converted small molecular substances [5]. By combining hardware sensors with data-driven soft measurement methods, wastewater treatment plants (WWTPs) can monitor complex processes in real-time and realize the indirect estimation of key indicators that are difficult to measure in practice [6,7]. Furthermore, this enables early warning of abnormal conditions, leading to more efficient and reliable process control.

Chemical oxygen demand (COD) is the amount of consumed oxygen when the dissolved and suspended matter are exposed to a specific oxidizing agent, which is considered a critical index for evaluating the organic pollution level in wastewater [8]. However, monitoring the COD concentrations online using hardware sensors involves high costs, and some substances in the water body can react irreversibly with the sensor elements, which has negative impacts on their accuracy and lifespan [9]. Thus, data-driven soft sensor methods for modeling COD concentrations in wastewater are widely applied in practical industrial operations.

Since machine learning methods can systematically learn relationships between variables from historical data, there have been a number of studies and applications of them in recent years [10]. Alavi and co-workers [11] provided a new insight in real-time COD prediction using several novel models combining hybridizing kernel-based extreme learning machines (ELM) with intelligent optimization algorithms. To make reliable predictions with time series data, Lotfi et al. [12] utilized autoregressive integrated moving average (ARIMA) as a fundamental structure and incorporated the outlier robust ELM technique to model the nonlinear and linear variables separately, thereby improving the accuracy of predicted effluent COD. Although ARIMA can explain the relationship between current and historical values using autoregression, differencing, and moving average calculations, its reliability significantly decreases when the input is an irregularly sampled time series with multiple data patterns. The appearance of deep learning methods provides an effective direction to these problems.

Deep learning models based on the recurrent neural network (RNN), which has a chain structure, have been applied in smart WWTPs in recent years because of their ability to learn historical trends in the temporal dimension [13,14]. Examples include the long short-term memory network (LSTM), which introduces gating mechanisms into the RNN units [15–17], and gated recurrent units (GRU), which improves gating structure in LSTM to reduce the number of parameters in training [18–20]. These models generate future trends by historical dependencies in time series, and all of them achieved accurate estimations of effluent COD, heavy metal, and sludge volume index with the inputs observed in both single and entire wastewater treatment processes.

Due to the advantages of the convolution calculation, models based on convolutional neural network (CNN) have performed well in specific time series tasks involving multi-scale patterns [21]. In [22], the CNN was combined with LSTM and GRU to compensate for their ability to abstract spatial features and thereby realize more reliable water quality predictions. Xie et al. [23] considered the temporal convolutional network (TCN) an effective framework for processing sequence data since it can perform dilated causal convolution in parallel. The experimental results proved that TCN can significantly improve the simulation of total nitrogen concentrations for the next eight hours in WWTP effluent. A study on forecasting the crowd flows utilized a graph convolutional network (GCN) to capture the complicated interactive relationships among external factors and integrated it with a prediction method for exceptional accomplishment in extracting spatio-temporal dependencies [24]. Thus, it is crucial for achieving accurate predictions to select and build the most suitable model according to the various data patterns and characteristics.

In this paper, a novel deep learning framework based on time scale decomposition and spatio-temporal feature extraction, which is denoted as MSTCN, is proposed to achieve more accurate predictions of the wastewater effluent indicator COD by identifying more complex time patterns in the data and realizing hierarchical feature extractions. The Fourier transform method is applied first to determine the top three main frequencies in the original data, according to which the inputs can be reconstructed into series at multiple time scales. After that, the historical dependencies and interactions between variables at each time scale can be captured by the TCN and GCN layers, respectively. The feature fusion across different time scales allows the proposed model to forecast utilizing information from various temporal dimensions and spatial interactions. Based on the actual wastewater quality data, the effectiveness and accuracy of the MSTCN model are demonstrated compared to widely applied prediction methods for time series.

The aim of this study is to achieve stable and accurate predictions of the COD concentrations in water bodies after treatment. This research makes efforts to develop an effective soft-sensor model to analyze monitoring data and provide reliable predictions for managing wastewater treatment processes.

This paper is organized as follows: Section 2 introduces the data collection, detailed methods, and the proposed model. In Section 3, the results of data preprocessing, parameter optimization, and experiments are shown and discussed. Finally, Section 4 presents the conclusion of this paper.

2. Materials and Methods

2.1. Data Collection

The dataset used for training and validating the predictive performance of models in this work was sampled from a wastewater treatment plant in South Korea [25] between 9 March 2007 and 29 February 2008. This plant employs advanced biological treatment processes to remove suspended solids, organic matter, and nutrients from the wastewater gradually, which mainly consists of a secondary clarifier and four basins of the denitrification, anaerobic, anoxic, and oxic processes. The detailed schematic is plotted in Figure 1.

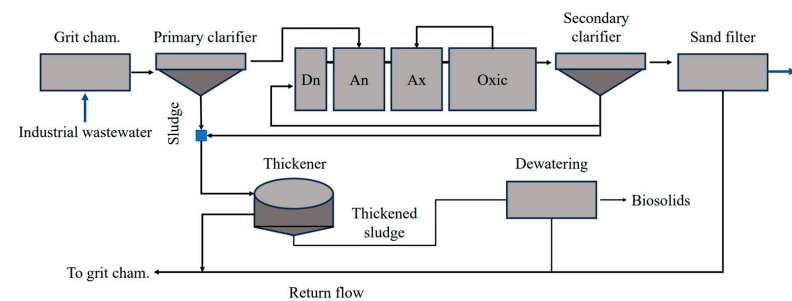


Figure 1. The schematic of the wastewater treatment plant.

Parameters monitored by hardware sensors during the whole process are influent flow, suspended solids (SS), biochemical oxygen demand (BOD), chemical oxygen demand (COD), total nitrogen (TN), total phosphorus (TP), and COD of the effluent (COD-E). Measurements in this dataset are the daily average values of seven variables sampled in a year. In the following training, the first six are chosen to be predictive variables, while COD-E is the target variable, which is to assess whether the water quality meets discharge standards. The curve of each variable is shown in Figure 2.

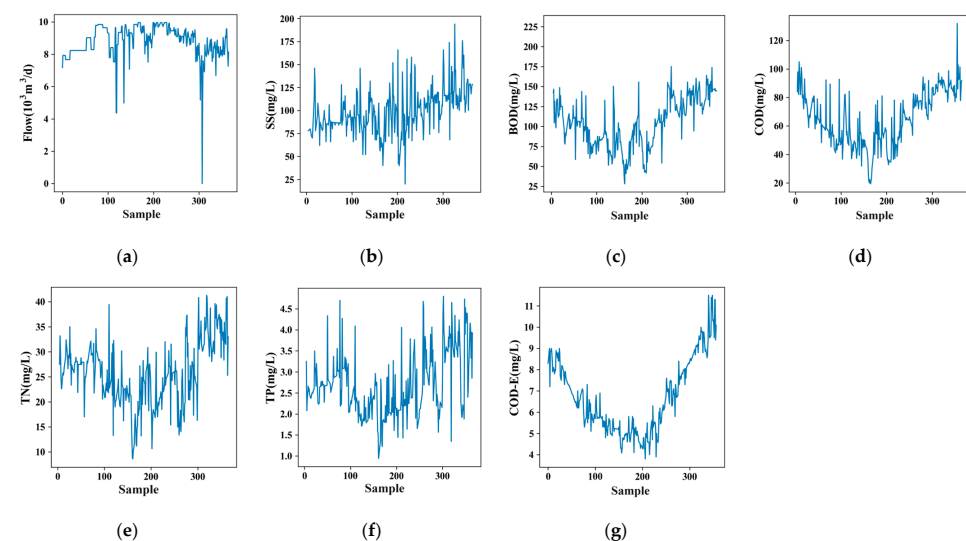


Figure 2. Curves of the wastewater treatment process data: (a) flow; (b) SS; (c) BOD; (d) COD; (e) TN; (f) TP; and (g) COD-E.

2.2. Fourier Transform

The Fourier transform is a mathematical method that can convert the time-domain signal $f(t)$ into the frequency-domain signal $F(\omega)$ and is widely applied in fields such as physics, engineering, and digital signal processing [26]. Since any periodic signal can be represented as a sum of simple oscillating functions of different frequencies, complex signals that meet the Dirichlet convergence condition can be decomposed into the superposition of infinitely many discrete sine or cosine waves using Fourier series.

The period of a continuous signal is T_1 , then its Fourier expansion is defined in Equation (1) as follows [27]:

$$f(t) = a_0 + \sum_{n=1}^{\infty} a_n \cos(n\omega_1 t) + \sum_{n=1}^{\infty} b_n \sin(n\omega_1 t) \quad (1)$$

where ω_1 is the fundamental frequency of $f(t)$, and each decomposed quantity is an integer multiple of ω_1 . While a_0 represents the direct current component, a_n and b_n are the amplitudes of the cosine and sine signals, respectively. In $[t_0, t_0 + T_1]$, they can be calculated as follows:

$$a_0 = \frac{1}{T_1} \int_{t_0}^{t_0+T_1} f(t) dt \quad (2)$$

$$a_n = \frac{2}{T_1} \int_{t_0}^{t_0+T_1} f(t) \cos(n\omega_1 t) dt \quad (3)$$

$$b_n = \frac{2}{T_1} \int_{t_0}^{t_0+T_1} f(t) \sin(n\omega_1 t) dt \quad (4)$$

However, many variables sampled from the wastewater treatment process in practice have no significant periodicity, in which case it can be equivalent to applying the Fourier transform to a function with an infinitely large period. Thus, the value of ω_1 approaches zero, and the Fourier transform can be described as follows:

$$F(\omega) = \int_{-\infty}^{\infty} f(t) e^{-i\omega t} dt \quad (5)$$

where $e^{-i\omega t}$ is a rotation vector for time-domain transformation and i in it is a unit of the complex number [28]. Each frequency component $F(\omega)$ contains the information of amplitude and phase, which determine the height and position of the wave, respectively. According to the top k $F(\omega)$ with the largest amplitudes, the most predominant k periods in the time series can be identified.

Thereby, the original data can be reconstructed across different time scales utilizing these periods obtained by the Fourier transform approach, which is essential for extracting specific features at each time scale and then enhancing the prediction accuracy.

2.3. Temporal Convolutional Network

Recurrent neural networks and their variants, such as LSTM and GRU, are structures that have demonstrated superior capabilities of capturing dynamic characteristics compared to convolutional models when dealing with time-series data. However, serial computing during forward propagation requires excessive memory to store intermediate results for each time step, which leads to more inefficient training.

TCN is a time-series modeling method based on convolutional operations that can work on entire sequences in parallel, enhancing the utilization of computing resources. Compared to LSTM, the architecture of TCN is simpler and there are fewer parameters, which further reduces the risk of overfitting [29].

TCN combines the causal convolution and dilated convolution to expand the receptive field of the filters without losing any input information. Additionally, layers are stacked

as the residual structure to eliminate the instability that arises from the increase in model depth. Figure 3 displays the TCN architecture.

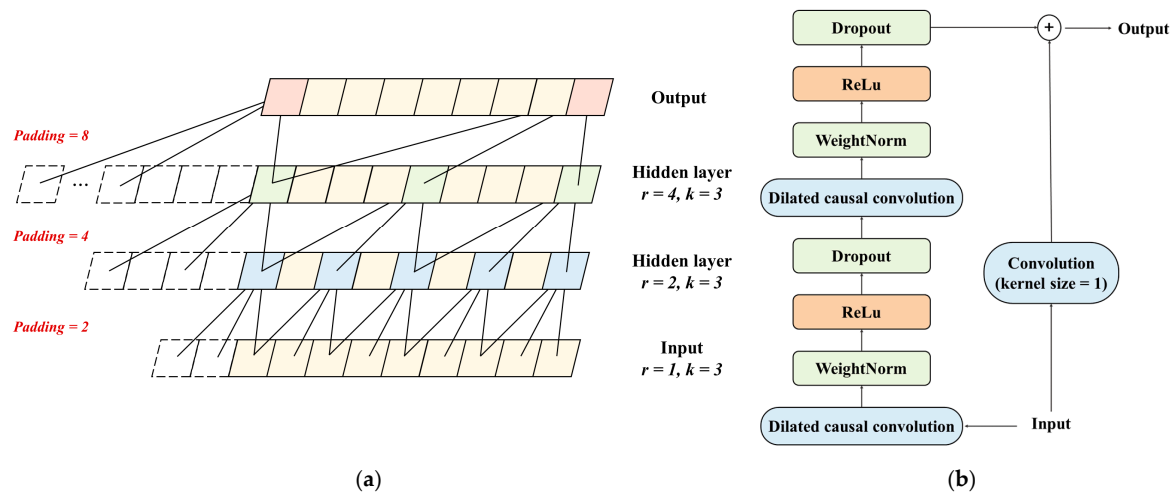


Figure 3. Detailed architecture of TCN: (a) dilated causal convolution and (b) residual block.

Causal convolution is a unidirectional structure, which implies that the outcome of d layer at t moment is only related to the factors before moment $t - 1$ in layer $d - 1$. Given an input sequence $X = (x_1, x_2, x_3, \dots, x_T)$ and a filter $F = (f_1, f_2, f_3, \dots, f_d)$, the convolution operation at x_s can be calculated via Equation (6) as follows [30]:

$$(F \times X)_{(s)} = \sum_{d=1}^D f_d x_{s-D+d} \quad (6)$$

Instead of utilizing pooling to increase the receptive field like CNN, TCN allows interval sampling of inputs. A standard convolutional layer can adjust the number of kernel intervals by a hyperparameter denoted dilation rate r to control the size of the convolution window, as shown in Equation (7).

$$m = (k - 1) \times r + 1 \quad (7)$$

where k is the filter size. Considering the dilated convolution injects $m - 1$ empty units into the one-dimensional convolutional layer, it is necessary to make padding with the same number of zeros at the end of each layer.

As presented in Figure 3b, each residual block in TCN consists of two dilated causal convolutional layers and ReLU nonlinear mapping layers. Moreover, the weight normalization and dropout layer are inserted after them to regularize the network, avoiding the gradient vanishing or explosion problems. To prevent the performance of TCN from degrading with an increasing number of layers, the input is directly added to the extracted features from the second convolutional layer. Since the two tensor shapes can be different, a common convolutional layer with a kernel size of 1×1 is applied for shape regularization before the residual summation.

2.4. Improvements

Due to the limitation of TCN in learning the interactions between variables, this work incorporates an adjacency matrix from graph algorithms to describe the information between nodes. This matrix stores the spearman correlation coefficients, serving as an additional input to help the model better understand and extract the dynamic relationships among variables.

GCN is specifically designed to deal with graph-structured data [31]. During the training process, GCN can achieve abundant feature representations by graph convolutions utilizing the constructed adjacency matrix. Thus, GCN is built to adapt the weights of the

input variables according to the correlations in the adjacency matrix and aggregate highly correlated features. Then, the output tensor is merged with time characteristics captured by the TCN module, which is composed of several residual blocks. By combining TCN and GCN, the hybrid model can obtain an output that contains both cross-sequence common patterns of features and the historical trends within the sequence.

As shown in Figure 4, extra calculations associated with GCN enable TCN to explicitly learn the data patterns and diverse relationships in multivariate time series, thereby improving the prediction accuracy and generalization capability.

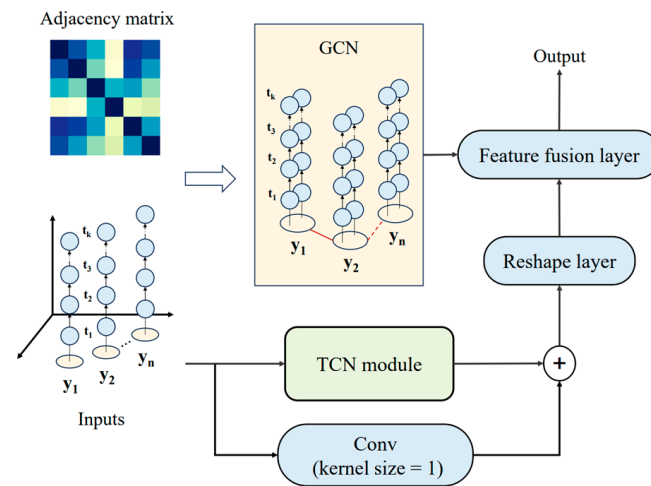


Figure 4. The architecture of TCN combined with GCN.

2.5. Modelling Methodology

As illustrated in Figure 5, modelling of the effluent COD by the multi-scale temporal convolutional network (MSTCN) can be carried out in the following three steps:

1. Fourier transform. Decompose the time series into the frequency domain and identify the k frequency components with the largest amplitudes. Then, reconstruct the original data into two-dimensional series at different time scales according to the top k periods.
2. Model training. For each time scale, construct an adjacency matrix to represent the correlations between variables and input it along with the time series data into the GCN layer to capture finer-grained relationships of features in the spatial dimension. Similarly, build a corresponding TCN layer to extract broad historical dependences within the sequences. Then, the representations learned at that time scale can be achieved by feature fusion of the results from GCN and TCN layers, and the final output of the hybrid model is aggregated adaptively from k scales using a fully connected layer.
3. Evaluation. Root mean squared error (RMSE), mean absolute error (MAE) [32], and the coefficient of determination (R^2) [33] are chosen as the metrics to evaluate the prediction accuracy and validate the proposed model performance statistically. The mathematical definitions of them are as follows:

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - y_{pred,i})^2} \quad (8)$$

$$MAE = \frac{1}{N} \sum_{i=1}^N |y_{pred,i} - y_i| \quad (9)$$

$$R^2 = 1 - \frac{\sum_{i=1}^N (y_i - y_{pred,i})^2}{\sum_{i=1}^N (y_i - \mu)^2} \quad (10)$$

When N is the sample size, μ denotes the mean value, and y_i and $y_{pred,i}$ are the measurements and the estimated values of the model, respectively. The R^2 value ranges from 0 to 1, with values closer to 1 indicating a higher proportion of the variance in the target variable that can be explained by the predictive variables, thus signifying a better fit of the model. Conversely, lower values of RMSE and MAE reflect a smaller error between the predictions and observations.

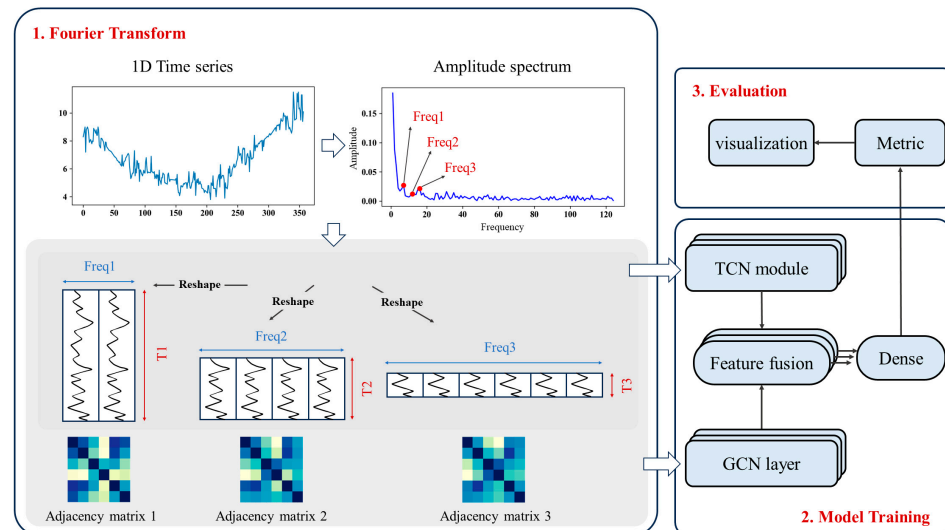


Figure 5. The modeling process of MSTCN.

3. Experimental Design

The experiments in this work all worked on a machine running on a Windows 11 operating system and equipped with an AMD Ryzen 5800H processor, which has a memory of 16 GB. All models were constructed under the Keras 2.3.1 framework by Python 3.7.0 and were trained with NVIDIA GeForce RTX 3060 GPU. Both equipment were manufactured in Santa Clara, CA, USA.

3.1. Data Pre-Processing

Missing values are removed directly as most of them exist in the first five records and the missing rate is less than 1%. The boxplots and violin plots for the remaining 358 samples are displayed in Figure 6.

After removing the outlier in Flow, the wastewater treatment dataset contains 357 samples, which are split into the training and test sets at a ratio of 7:3 in this work. For the integrity and continuity of the time series data, the first 250 samples are considered as the training set, and the remaining 107 measurements are used as the test set. Among all violin plots, SS, BOD, COD, and TN exhibit symmetrical shapes around their median lines, with high-density areas aligning closely with the median, indicating that these variables approximately follow a normal distribution. However, the data points of TP and COD-E are significantly clustered in the smaller value areas, and it can be inferred that these two variables are supposed to show a negative skew distribution. Thus, it is necessary to perform exponential transformation on them before training to eliminate the skewness.

Together with the results of the time lag analysis in Figure 7 and the spearman correlation analysis in Figure 8, there are strong interactions between variables of the experimental data, while each series has a complex temporal pattern within itself. As observed in Figure 7, there are still rather strong dependencies between the data point at time t and that before 30 timesteps in terms of BOD, COD, and COD-E. These factors mentioned above are both crucial for accuracy when forecasting with the multivariate time series. Therefore, a multi-scale model can better deal with the dynamic characteristics, and the introduction of the adjacency matrix allows the model to flexibly capture diverse relationships in data, enhancing its predictive performance and interpretability.

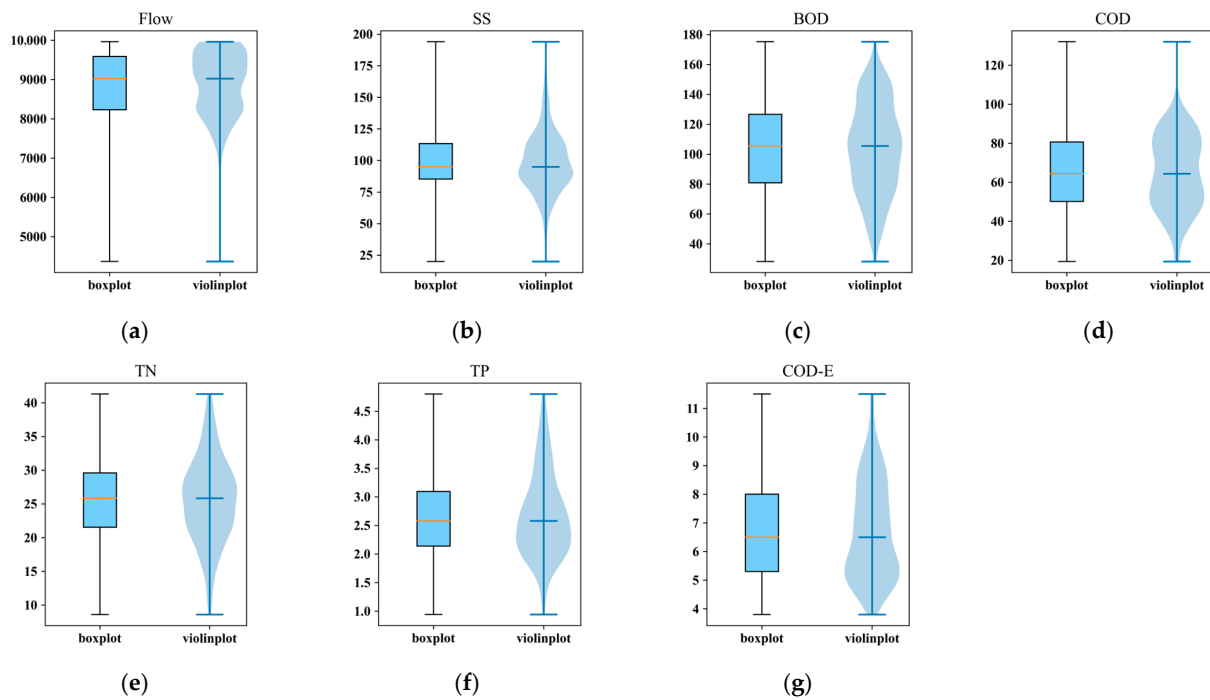


Figure 6. The boxplot and violin plot of each variable: (a) Flow; (b) SS; (c) BOD; (d) COD; (e) TN; (f) TP; and (g) COD-E.

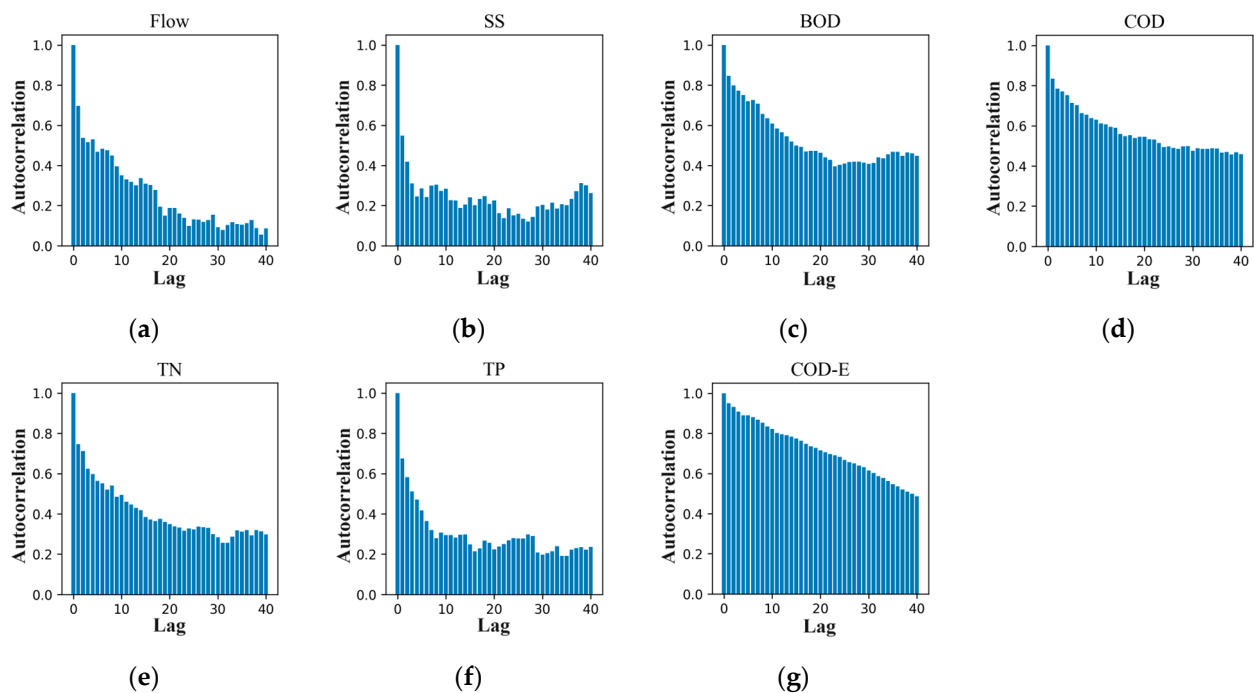


Figure 7. The temporal correlation of variables: (a) flow; (b) SS; (c) BOD; (d) COD; (e) TN; (f) TP; and (g) COD-E.

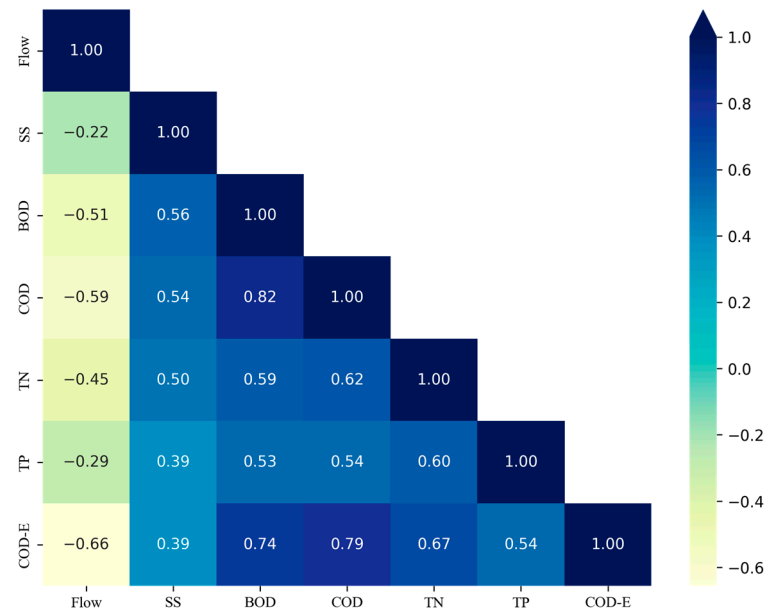


Figure 8. Spearman coefficient of correlation between variables.

3.2. Fourier Transform

To construct the input for MSTCN, the major periods of each variable in the original data are first identified using the Fourier transform approach. The amplitude spectrum is plotted in Figure 9, where the top three largest magnitudes are marked with red dots. According to the mode of main frequencies, the top three periods for the entire dataset can be determined to be 20, 25, and 41, respectively.

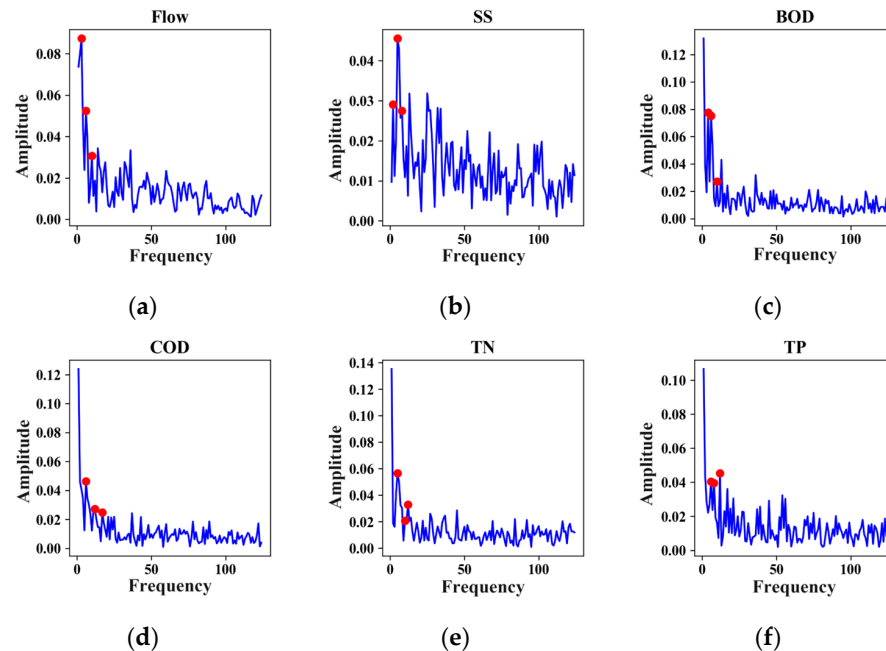


Figure 9. The amplitude spectrum for each input variable: (a) flow; (b) SS; (c) BOD; (d) COD; (e) TN; and (f) TP. Red dots are the top three largest magnitudes of each variable.

Then, the training set of an initial size of (250, 6) is reshaped according to these determined periods into matrices with sizes of (13, 20, 6), (10, 25, 6), and (7, 41, 6), respectively. The spearman correlation coefficients for the input variables at each time scale are calculated and stored in the corresponding adjacency matrices, as shown in Figure 10.

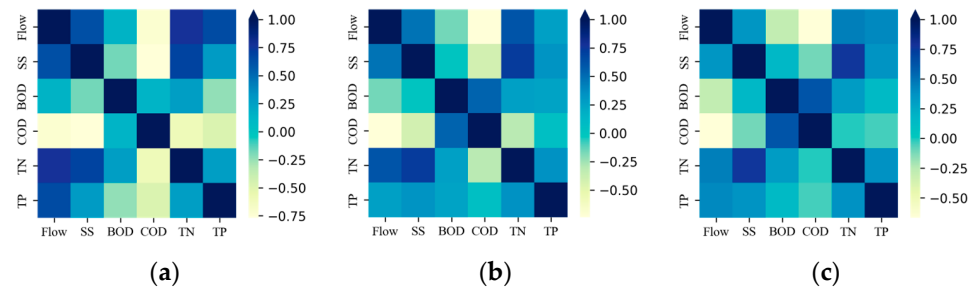


Figure 10. The correlations between variables at different time scales: (a) period = 20; (b) period = 25; and (c) period = 41.

3.3. Optimization

Considering the complex structures and training consumptions of deep learning models, the Bayesian optimization algorithm with the Gaussian process is used for hyperparameter tuning. It is a classic method for the global optimization of unknown functions, which adjusts the parameter selection for the next iteration within the specified range based on the results of previous iterations and updates the posterior distribution of the objective function until it closely aligns with the true distribution [34].

In this study, the metric *RMSE* is taken as the optimization target, and the global optimal solutions for the following hyperparameters are searched within 100 iterations: kernel size *ks*, dilation rate *r*, batch size, optimizer, and learning rate. To prevent overfitting and reduce the computational resources required, an early stopping mechanism is also performed in this stage. As a result, the MSTCN model with a kernel size of 7 and a dilation rate of 3 is proven to outperform other structures, which is trained at a learning rate of 0.01 and a batch size of 64 using the Adam optimizer.

Since the kernel size of the convolutional layer in GCN is generally the same size as the adjacency matrix, the critical factors that can significantly influence the final performance of MSTCN are the kernel size of the TCN module and its dilation rate. Thus, the grid search method is applied to further determine their values, and the R^2 of MSTCN on the test set for different parameter combinations is plotted in Figure 11. It is obvious that the model provides the most superior ability to fit future trends of the sequence when *ks* is 7 and *r* is 3, which is consistent with the results of Bayesian optimization. The detailed structure of the optimized MSTCN model is listed in Table 1.

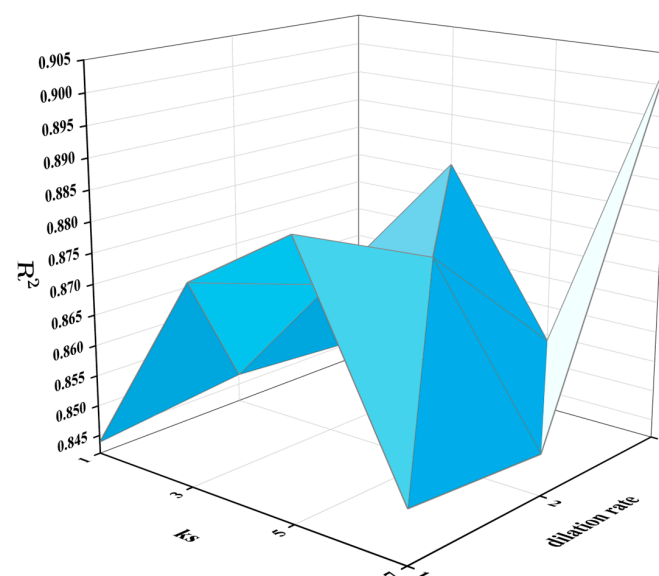


Figure 11. The results of grid search for MSTCN model.

Table 1. Structure of the MSTCN model with the best forecasting performance.

NO	Layer	Parameter	NO	Layer	Parameter
1	Input	Input1 = (13, 20, 6) Input2 = (10, 25, 6) Input3 = (7, 41, 6)	4	TCN	Dilation_rate = 3 Kernel_size = 7 Activation = ReLU
2	GCN	Kernal_size = 6	5	Dropout	Rate = 0.2
3	Dropout	Rate = 0.2	6	Dense	Units = 1

3.4. Comparison Methods

To validate the superiority of the proposed MSTCN framework, several classical deep learning models are trained in this study to predict the effluent COD concentrations for the following day, including CNN, LSTM, LSTM-based on the attention mechanism (ALSTM), and the common TCN. The hyperparameters involved in building and training these mentioned models are illustrated in Table 2.

Table 2. Hyperparameters of CNN, LSTM, ALSTM, and TCN.

Parameter	CNN	LSTM	ALSTM	TCN
Kernel size	3	/	/	5
filters	32	/	/	(128, 64, 32, 16, 16)
Dilation rate	/	/	/	(1, 2, 4, 8, 16)
Activation	ReLU	Tanh	Tanh	LeakyReLU
Dropout rate	0.2	0.1	0.2	0.2
Units	/	120	(120, 240)	/
Batch size	64	64	32	64
Optimizer	RMSprop	Adam	RMSprop	Adam
Max epochs	500	500	500	500
Learning rate	0.01	0.01	0.001	0.01

4. Results and Discussion

Table 3 presents the predictive performance of the proposed MSTCN and comparison models. The training duration required for each model to converge is also recorded in it for more comprehensive evaluations from different dimensions.

Table 3. The performance metrics of comparison models for predicting COD-E.

Model	Training/Test Set	R^2	RMSE	MAE	Time (s)
CNN	Training	0.9286	0.3353	0.2914	7.2
	Test	0.7688	0.6108	0.5606	
LSTM	Training	0.9768	0.1910	0.1432	14.3
	Test	0.8027	0.5643	0.4659	
ALSTM	Training	0.9196	0.3557	0.2855	10.2
	Test	0.8747	0.4496	0.3552	
TCN	Training	0.8834	0.4286	0.3969	7.6
	Test	0.8406	0.5071	0.4292	
MSTCN	Training	0.9786	0.1834	0.1480	8.0
	Test	0.9044	0.3927	0.2765	

It is obvious that MSTCN not only demonstrates remarkable improvements in predictive accuracy over other models but also exhibits superior training efficiency. MSTCN has a training R^2 of 0.9786, RMSE of 0.1834, and MAE of 0.1480, while it achieves the highest R^2 of 0.9044, lowest RMSE of 0.3927, and lowest MAE of 0.2765 on the test set. Figure 12 plots its result curves against the measured one. Although there are some larger deviations between predictions and observations at the last few samples, MSTCN continues to successfully capture the overall trends in the test data and maintain a high degree of accuracy. When

dealing with the untrained sequence that contains both long-term dependencies and short-term fluctuations, MSTCN can also generate predictions with slight biases, confirming its robustness and reliability in forecasting effluent COD concentrations. Additionally, it takes 8 s to reach convergence, which is quite efficient given the model complexity.

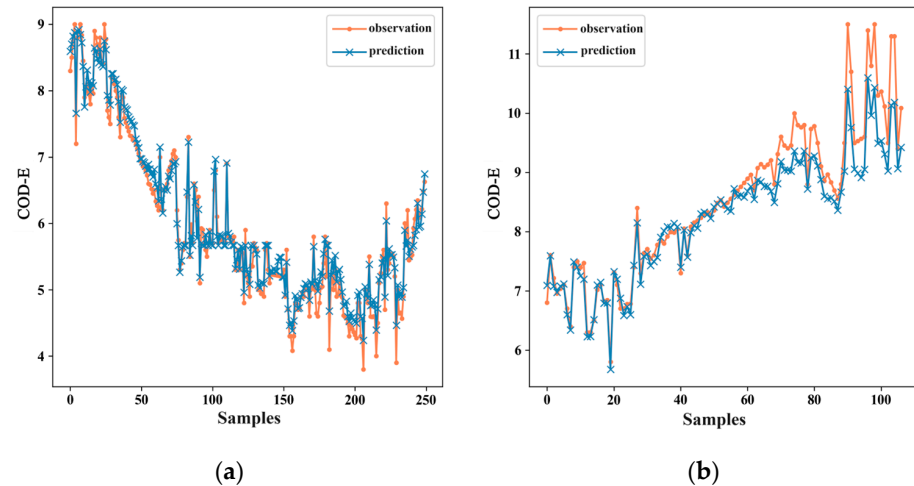


Figure 12. The forecasting results of the MSTCN model: (a) training set and (b) test set.

Figure 13 presents the predictions of CNN and LSTM on the test set, which suggests that LSTM can better fit the observation curve with smaller errors when there exist similar trends within a relatively long time period. However, both of them have poor generalization abilities on this wastewater treatment process dataset. Although these two models can achieve an explanation degree of over 90% for the variability in the COD-E data during training, there is a significant decrease in predictive performance on the test set. The test R^2 value of MSTCN is 17.63% and 12.67% higher than that of CNN and LSTM, respectively. According to the test RMSE, MSTCN has a 35.68% and 30.42% reduction compared to them, respectively.

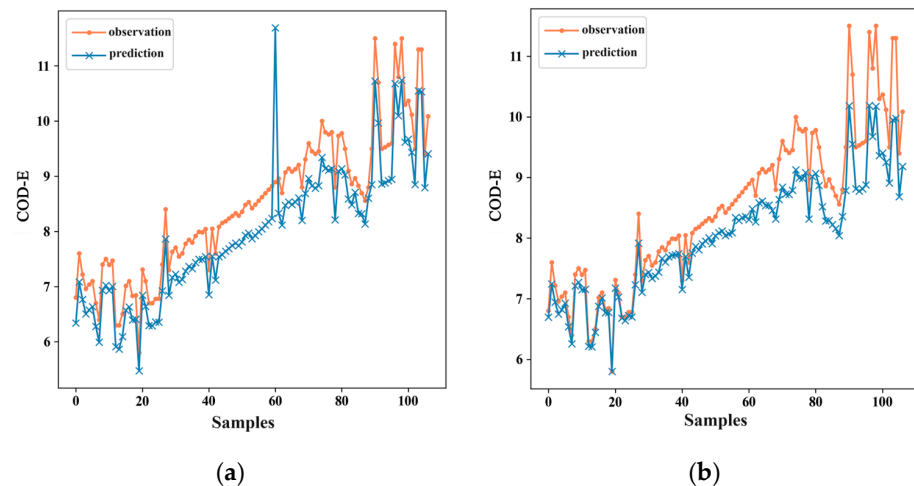


Figure 13. Comparison of CNN and LSTM on the test set: (a) CNN and (b) LSTM.

Different from the fixed receptive field utilized by CNN for capturing local characteristics, the causal convolution in TCN enables it to refer to the state of time $t - 1$ in the previous hidden layer when learning the changes at time t . Thus, TCN can provide more accurate predictions for overall trends in the time series but is sometimes less effective than CNN on short-term abrupt changes, which can also be confirmed by comparing Figures 13a and 14a. Considering that the dilation rate allows TCN to adjust the size of its receptive field window and makes its learning range more flexible, TCN outperforms LSTM on the dataset, which contains many fluctuations only existing in a short period. RMSE of the TCN model on the test

set is 0.5071, which is 10.14% lower than LSTM, while the R^2 value is 4.72% higher. However, MSTCN still shows significant improvements over TCN in both training and test performance metrics, as the multi-scale approach of MSTCN enables the TCN layers in it to be trained separately to learn the unique temporal patterns of the target variable for each scale. Thus, there is a 7.59% increase in test R^2 and a 22.56% decrease in RMSE compared with TCN.

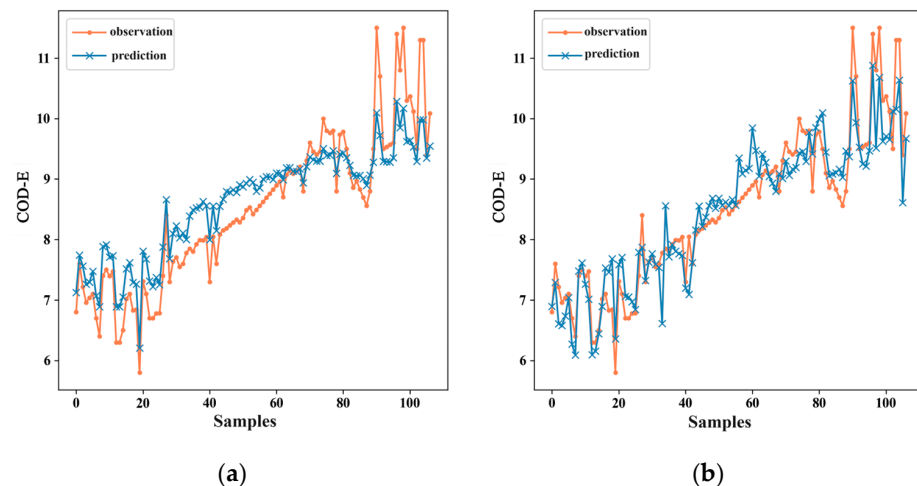


Figure 14. Comparison of TCN and ALSTM on the test set: (a) TCN and (b) ALSTM.

The attention mechanism is a method that can highlight important information by assigning higher weights to critical features in the input sequence [35]. By inserting an attention module between the LSTM and the input layer, the negative influence of redundant information on LSTM performance can be effectively reduced during the training process. Figure 14b plots the forecasting results of ALSTM on the test set, and it is obvious that there are smaller differences between the ALSTM curve and the observed one during the stage with a gradually increasing trend. Comparing Figures 14a and 14b, ALSTM has a superior explanation of dynamic changes at the end of the sequence than TCN. The test R^2 , RMSE, and MAE of it are 0.8747, 0.4496, and 0.3552, indicating a second-ranked performance among all comparative models. Due to the GCN layers, MSTCN can extract more diverse feature representations with the extracted relationships at each scale from the adjacency matrix. As a result, these metrics of MSTCN are 3.39%, 12.64%, and 22.14% improved than those of ALSTM.

Considering the convolution operations used in CNN, TCN, and MSTCN, their training time is much shorter than LSTM and ALSTM, which are based on serial computation. MSTCN takes 8 s to reach its optimal performance, which is 2.2 s shorter than ALSTM and 6.3 s shorter than LSTM. The ability of MSTCN to converge faster while maintaining high predictive performance suggests its scalability and applicability in complex and large-scale industrial settings [36].

Overall, the proposed MSTCN model shows high predictive accuracy in both training and test sets, making it a reliable tool for forecasting effluent COD concentrations after wastewater treatment. The model effectively identifies and simulates both long-term trends and short-term fluctuations, which is crucial for achieving accurate predictions from time series with complex data patterns [37]. These findings highlight the potential of MSTCN for improving accuracy and operational efficiency in wastewater quality prediction.

5. Conclusions

In this work, a hybrid model based on time scale decomposition and spatio-temporal feature extraction, which is denoted as MSTCN, is proposed to identify more complex patterns and variable interactions in the data and achieve more accurate predictions of the wastewater effluent indicator COD-E. The Fourier transform method is applied first to determine the top three main frequencies in the original data, according to which the inputs can

be reconstructed into series at three time scales. For each time scale, the spearman correlation coefficients between variables are stored in a separate adjacency matrix. Then, build TCN and GCN layers to learn temporal dependencies within the sequences and aggregate highly weighted feature representations, respectively. Finally, the output of MSTCN is the result of feature fusion from TCN and GCN layers across different time scales.

The dataset used to validate the model performance is sampled from the nutrient removal process of a wastewater treatment plant in South Korea. The evaluation indicator COD-E for effluent quality was used as the target variable, which included both abrupt short-term fluctuations and long-term overall trends in its series. The proposed MSTCN model achieved an R^2 of 0.9786, an RMSE of 0.1834, and an MAE of 0.1480 on the training set, while the values of R^2 , RMSE, and MAE were 0.9044, 0.3927, and 0.2765 for the test set, respectively. The experimental results demonstrate the effectiveness of MSTCN in predicting future COD-E concentrations with multivariate and multi-pattern time series data.

In future work, the effluent biochemical oxygen demand (BOD) is supposed to be considered as another critical indicator for evaluating the wastewater quality after treatment, and the relationship between BOD and COD can be focused on to achieve a more comprehensive monitoring and more accurate predictions of the effluent quality.

Author Contributions: Methodology, Y.G.; data collection, F.Z.; supervision, H.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by Shandong Provincial Natural Science Foundation, China (ZR2021MF135) and Natural Science Foundation of Jiangsu Provincial Universities, China (22KJA530003).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Restrictions apply to the availability of these data. Data was obtained from a third party and are available from the authors with the permission of the third party. All data applied or analyzed during this study is included in this published article: <https://doi.org/10.2166/wst.2009.346>.

Conflicts of Interest: Fengshan Zhang and Hongbin Liu were employed by the company Shandong Huatai Paper Co., Ltd. and Shandong Yellow Triangle Biotechnology Industry Research Institute Co., Ltd. The remaining author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Liu, M.Y.; Lv, J.P.; Qin, C.H.; Zhang, H.; Wu, L.L.; Guo, W.; Guo, C.S.; Xu, J. Chemical fingerprinting of organic micropollutants in different industrial treated wastewater effluents and their effluent-receiving river. *Sci. Total Environ.* **2022**, *838*, 156399. [CrossRef] [PubMed]
2. Methneni, N.; Morales-González, J.A.; Jaziri, A.; Ben Mansour, H.; Fernandez-Serrano, M. Persistent organic and inorganic pollutants in the effluents from the textile dyeing industries: Ecotoxicology appraisal via a battery of biotests. *Environ. Res.* **2021**, *196*, 110956. [CrossRef] [PubMed]
3. Li, L.L.; Shi, Y.B.; Huang, Y.; Xing, A.L.; Xue, H. The Effect of Governance on Industrial Wastewater Pollution in China. *Int. J. Environ. Res. Public Health* **2022**, *19*, 9316. [CrossRef] [PubMed]
4. Sathasivam, M.; Shanmugapriya, S.; Yogeshwaran, V.; Priya, A. Industrial waste water treatment using advanced oxidation process—A review. *Int. J. Eng. Adv. Technol.* **2019**, *8*, 485–488.
5. Yu, X.; Zuo, J.; Tang, X.; Li, R.; Li, Z.; Zhang, F. Toxicity evaluation of pharmaceutical wastewaters using the alga *Scenedesmus obliquus* and the bacterium *Vibrio fischeri*. *J. Hazard. Mater.* **2014**, *266*, 68–74. [CrossRef] [PubMed]
6. Newhart, K.B.; Holloway, R.W.; Hering, A.S.; Cath, T.Y. Data-driven performance analyses of wastewater treatment plants: A review. *Water Res.* **2019**, *157*, 498–513. [CrossRef] [PubMed]
7. Liu, S.; Chong, Y.; Hui, Z.; Hong-Bin, L. Soft-sensor modeling of papermaking wastewater treatment process based on Gaussian process. *China Environ. Sci.* **2018**, *38*, 2564–2571.
8. Geerdink, R.B.; van den Hurk, R.S.; Epema, O.J. Chemical oxygen demand: Historical perspectives and future challenges. *Anal. Chim. Acta* **2017**, *961*, 1–11. [CrossRef] [PubMed]
9. Ching, P.M.L.; Zou, X.; Wu, D.; So, R.H.Y.; Chen, G.H. Development of a wide-range soft sensor for predicting wastewater BOD5 using an eXtreme gradient boosting (XGBoost) machine. *Environ. Res.* **2022**, *210*, 112953. [CrossRef]
10. Singh, N.K.; Yadav, M.; Singh, V.; Padhiyar, H.; Kumar, V.; Bhatia, S.K.; Show, P.L. Artificial intelligence and machine learning-based monitoring and design of biological wastewater treatment systems. *Bioresour. Technol.* **2023**, *369*, 128486. [CrossRef]

11. Alavi, J.; Ewees, A.A.; Ansari, S.; Shahid, S.; Yaseen, Z.M. A new insight for real-time wastewater quality prediction using hybridized kernel-based extreme learning machines with advanced optimization algorithms. *Environ. Sci. Pollut. Res.* **2022**, *29*, 20496–20516. [\[CrossRef\]](#)
12. Lotfi, K.; Bonakdari, H.; Ebtehaj, I.; Mjalli, F.S.; Zeynoddin, M.; Delatolla, R.; Gharabaghi, B. Predicting wastewater treatment plant quality parameters using a novel hybrid linear-nonlinear methodology. *J. Environ. Manag.* **2019**, *240*, 463–474. [\[CrossRef\]](#)
13. Dairi, A.; Cheng, T.Y.; Harrou, F.; Sun, Y.; Leiknes, T. Deep learning approach for sustainable WWTP operation: A case study on data-driven influent conditions monitoring. *Sustain. Cities Soc.* **2019**, *50*, 101670. [\[CrossRef\]](#)
14. Wongburi, P.; Park, J.K. Prediction of Wastewater Treatment Plant Effluent Water Quality Using Recurrent Neural Network (RNN) Models. *Water* **2023**, *15*, 3325. [\[CrossRef\]](#)
15. Kohen, E.; Farhi, N.; Shavitt, Y.; Mamane, H. Prediction of a full scale WWTP activated sludge SVI test using an LSTM neural network. *Environ. Sci. Water Res. Technol.* **2022**, *8*, 2786–2795. [\[CrossRef\]](#)
16. Wang, Z.F.; Man, Y.; Hu, Y.S.; Li, J.G.; Hong, M.N.; Cui, P.Z. A deep learning based dynamic COD prediction model for urban sewage. *Environ. Sci. Water Res. Technol.* **2019**, *5*, 2210–2218. [\[CrossRef\]](#)
17. Zhu, H.Q.; Wang, Q.L.; Zhang, F.X.; Yang, C.H.; Li, Y.G. A prediction method of electrocoagulation reactor removal rate based on Long Term and Short Term Memory-Autoregressive Integrated Moving Average Model. *Process Saf. Environ. Prot.* **2021**, *152*, 462–470. [\[CrossRef\]](#)
18. Wan, K.Y.; Du, B.X.; Wang, J.H.; Guo, Z.W.; Feng, D.; Gao, X.; Shen, Y.; Yu, K.P. Deep learning-based intelligent management for sewage treatment plants. *J. Cent. South Univ.* **2022**, *29*, 1537–1552. [\[CrossRef\]](#)
19. Jiang, Y.Q.; Li, C.L.; Song, H.X.; Wang, W.H. Deep learning model based on urban multi-source data for predicting heavy metals (Cu, Zn, Ni, Cr) in industrial sewer networks. *J. Hazard. Mater.* **2022**, *432*, 128732. [\[CrossRef\]](#)
20. Yan, K.F.; Li, C.L.; Zhao, R.B.; Zhang, Y.T.; Duan, H.P.; Wang, W.H. Predicting the ammonia nitrogen of wastewater treatment plant influent via integrated model based on rolling decomposition method and deep learning algorithm. *Sustain. Cities Soc.* **2023**, *94*, 104541. [\[CrossRef\]](#)
21. Liu, X.; Wang, W. Deep Time Series Forecasting Models: A Comprehensive Survey. *Mathematics* **2024**, *12*, 1504. [\[CrossRef\]](#)
22. Haq, K.R.A.; Harigovindan, V. Water quality prediction for smart aquaculture using hybrid deep learning models. *IEEE Access* **2022**, *10*, 60078–60098.
23. Xie, Y.F.; Chen, Y.Q.; Wei, Q.; Yin, H.L. A hybrid deep learning approach to improve real-time effluent quality prediction in wastewater treatment plant. *Water Res.* **2024**, *250*, 121092. [\[CrossRef\]](#)
24. Ali, A.; Zhu, Y.; Zakarya, M. Exploiting dynamic spatio-temporal graph convolutional neural networks for citywide traffic flows prediction. *Neural Netw.* **2022**, *145*, 233–247. [\[CrossRef\]](#)
25. Kim, M.; Kim, Y.; Prabu, A.; Yoo, C. A systematic approach to data-driven modeling and soft sensing in a full-scale plant. *Water Sci. Technol.* **2009**, *60*, 363–370. [\[CrossRef\]](#)
26. Duhamel, P.; Vetterli, M. Fast Fourier transforms: A tutorial review and a state of the art. *Signal Process.* **1990**, *19*, 259–299. [\[CrossRef\]](#)
27. Weisstein, E.W. Fourier Transform. 2004. Available online: <https://mathworld.wolfram.com> (accessed on 3 February 2004).
28. Heckbert, P. Fourier transforms and the fast Fourier transform (FFT) algorithm. *Comput. Graph.* **1995**, *2*, 15–463.
29. Bai, S.; Kolter, J.Z.; Koltun, V. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv* **2018**, arXiv:1803.01271.
30. Bi, J.; Zhang, X.; Yuan, H.; Zhang, J.; Zhou, M. A hybrid prediction method for realistic network traffic with temporal convolutional network and LSTM. *IEEE Trans. Autom. Sci. Eng.* **2021**, *19*, 1869–1879. [\[CrossRef\]](#)
31. Zhang, S.; Tong, H.; Xu, J.; Maciejewski, R. Graph convolutional networks: A comprehensive review. *Comput. Soc. Netw.* **2019**, *6*, 11. [\[CrossRef\]](#)
32. Hodson, T.O. Root-mean-square error (RMSE) or mean absolute error (MAE): When to use them or not. *Geosci. Model Dev.* **2022**, *15*, 5481–5487. [\[CrossRef\]](#)
33. Di Bucchianico, A. Coefficient of determination (R^2). In *Encyclopedia of Statistics in Quality and Reliability*; Wiley: Hoboken, NJ, USA, 2008.
34. Laumanns, M.; Ocenasek, J. Bayesian optimization algorithms for multi-objective optimization. In Proceedings of the International Conference on Parallel Problem Solving from Nature, Granada, Spain, 7–11 September 2002; pp. 298–307.
35. Niu, Z.; Zhong, G.; Yu, H. A review on the attention mechanism of deep learning. *Neurocomputing* **2021**, *452*, 48–62. [\[CrossRef\]](#)
36. Ly, Q.V.; Truong, V.H.; Ji, B.; Nguyen, X.C.; Cho, K.H.; Ngo, H.H.; Zhang, Z. Exploring potential machine learning application based on big data for prediction of wastewater quality from different full-scale wastewater treatment plants. *Sci. Total Environ.* **2022**, *832*, 154930. [\[CrossRef\]](#)
37. Wu, J.; Wang, Z. A hybrid model for water quality prediction based on an artificial neural network, wavelet transform, and long short-term memory. *Water* **2022**, *14*, 610. [\[CrossRef\]](#)

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.