# Long-Term Water Quality Prediction With Transformer-Based Spatial-Temporal Graph Fusion

Jing Bi, *Senior Member, IEEE*, Ziqi Wang, *Student Member, IEEE*, Haitao Yuan, *Senior Member, IEEE*, Xiangxi Wu, Renren Wu, Jia Zhang, *Senior Member, IEEE*, and MengChu Zhou, *Fellow, IEEE*

***Abstract*— Over the past decades of rapid development, the global water pollution problem became prominent. Accurate water quality prediction can detect the trend and anomaly of water quality changes in advance, thereby taking timely measures to avoid water quality problems. Traditional statistical methods for water quality prediction tend to fail to capture the complex relationship among multiple water quality variables. Deep learning models face a challenge to capture both temporal dependence and spatial correlation of the water quality series data. To solve the above problems, this work proposes an adaptive and dynamic graph fusion water quality prediction model based on a spatiotemporal attention mechanism named Spatial-Temporal Graph Fusion Transformer (STGFT). It integrates a spatial attention encoder, a temporal attention encoder, an adaptive dynamic adjacency matrix generator, and a multi-graph fusion layer. Among them, the first two are adopted to capture the spatial correlations and temporal characteristics among different water quality monitoring stations, respectively. The generator can produce adaptive and dynamic adjacency matrices to reflect potential spatial relationships in a river network. Experimental results with real-life water quality datasets reveal that the prediction accuracy of STGFT outperforms the existing state-of-the-art models.**

***Note to Practitioners*—This paper is motivated by the problem of long-term water quality prediction. The highly volatile water quality data and the nonlinear characteristics of the time series greatly affect the accuracy of the forecasting task. Existing approaches fail to simultaneously capture spatial correlations and temporal characteristics among different water quality monitoring stations, affecting the accuracy of water quality predictions. This work proposes a water quality prediction method that captures the spatial correlations and temporal characteristics among different water quality monitoring stations. Moreover, it produces adaptive and dynamic adjacency matrices to reflect potential spatial relationships in a river network. Experimental results from three real-world datasets show that this approach is feasible and obtains more accurate prediction results. Furthermore, this method can also be applied to other areas of time series prediction, including finance, traffic, and smart manufacturing.**

***Index Terms*— Spatiotemporal prediction, water environment, graph neural networks, attention mechanism.**

## I. Introduction

NOWADAYS, the deterioration of water environment has become one of the most important issues constraining the sustainable development of our society. To solve this problem, water quality prediction methods [1] are proposed to forecast elemental values of the water environment in the future based on past monitoring data. Hence, people can take timely steps to address water pollution by accurately predicting future water quality. There are two common methods of predicting water quality, *i.e.*, mechanism models and deep learning ones. The former needs to select proper model parameters and requires prior knowledge and professional experience. They are often based on specific assumptions, *e.g.*, water quality trends are linear, and time series are stationary. However, these assumptions may not be true in an actual situation, thus biasing the prediction results.

Deep learning models, *e.g.*, back propagation neural networks, recurrent neural networks, and convolutional neural networks [2] are suitable for water quality prediction through the limited water quality information. However, with socio-economic ties among regions strengthening, the water environment shows complex changes. Multiple water quality monitoring stations interact, and their data are affected by historical values and values from upstream monitoring stations, increasing the difficulty of making accurate water quality predictions. Graph Neural Networks (GNNs) have shown powerful capabilities in dealing with complex spatial sequence data, and they can thus be adopted to solve the above problem. Specifically, they can handle non-Euclidean data and represent water quality data in spatial dimensions. Therefore, they can model the spatial relationship among water quality monitoring stations at different locations [3].

Received 9 November 2024; accepted 24 January 2025. Date of publication 27 January 2025; date of current version 11 April 2025. This article was recommended for publication by Associate Editor C. Wang and Editor K. Liu upon evaluation of the reviewers' comments. This work was supported in part by Beijing Natural Science Foundation under Grant L233005 and Grant 4232049, in part by the National Natural Science Foundation of China under Grant 62173013 and Grant 62473014, and in part by Beihang World TOP University Cooperation Program. *(Corresponding author: Haitao Yuan.)*

Jing Bi, Ziqi Wang, and Xiangxi Wu are with the College of Computer Science, Beijing University of Technology, Beijing 100124, China (e-mail: bijing@bjut.edu.cn; ziqi_wang@emails.bjut.edu.cn; Wuxiangxi7@emails.bjut.edu.cn).

Haitao Yuan is with the School of Automation Science and Electrical Engineering, Beihang University, Beijing 100191, China (e-mail: yuan@buaa.edu.cn).

Renren Wu is with the State Environmental Protection Key Laboratory of Water Environmental Simulation and Pollution Control, South China Institute of Environmental Sciences, Ministry of Ecology and Environment of the People's Republic of China, Guangzhou 510530, China (e-mail: wurenren@scies.org).

Jia Zhang is with the Department of Computer Science, Lyle School of Engineering, Southern Methodist University, Dallas, TX 75205 USA (e-mail: jiazhang@smu.edu).

MengChu Zhou is with the Department of Electrical and Computer Engineering, New Jersey Institute of Technology, Newark, NJ 07102 USA (e-mail: zhou@njit.edu).

Digital Object Identifier 10.1109/TASE.2025.3535415

However, due to the high complexity of river networks and the uncertainty of spatial relationships, inaccurate information in defining their graph structure may result in an inaccurate graph. Therefore, a predefined graph structure can only capture the local spatial information, and it is difficult to capture the spatial dependencies adequately.

Based on the aforementioned analysis, this work proposes a water quality prediction model named Spatial-Temporal Graph Fusion Transformer (STGFT). It integrates spatial attention encoder (SAE) and temporal attention encoder (TAE) to capture the spatial correlations and temporal characteristics among different water quality monitoring stations, respectively. An adaptive dynamic adjacency matrix generator (ADMG) is designed to utilize spatial and temporal characteristics to generate adaptive and dynamic graphs, better reflecting potential spatial relationships in a river network. Specifically, ADMG utilizes spatial features output from SAE, a randomly initialized adjacency matrix, and a predefined graph to generate these two graphs. The adaptive graph is completely deprived of prior knowledge of the river network, and the main spatial features are extracted based on learning. The dynamic graph further incorporates prior knowledge of the predefined graph to provide potential spatial relationships as an auxiliary. It captures spatial relationships over time, reflecting dynamic dependency structures in river networks. The usage of ADMG prevents STGFT from being restricted by a predefined graph structure of the fixed river network information. Experimental results on three practical datasets show that STGFT has high accuracy in long-term water quality predictions. Furthermore, the effectiveness of each module in STGFT is verified by our ablation studies, which proves that the proposed ADMG can help to excavate the potential spatial dependence and play a significant role in raising the model's prediction performance. The main contributions of this work are summarized as:

1) TAE is proposed to learn temporal features of the time series data in each water quality monitoring station. It captures the correlation of water quality elements among different time steps. Moreover, ADMG based on SAE is designed to capture the spatial correlations among water quality monitoring stations. Despite the high complexity and uncertainty of spatial relationships in river networks, ADMG generates dynamic and adaptive graphs to mine the potential spatial dependencies in river networks.
2) TAE, SAE, and ADMG are integrated into STGFT for long-term water quality prediction. It can utilize water quality data to capture the spatial correlations and temporal characteristics for accurate water quality prediction.
3) STGFT is compared with four typical peers under three real-world water quality datasets. Experimental results show that STGFT is superior to its peers in the long-term water quality prediction.

The remainder of this work is structured as follows. Section II discusses the related work of different water quality prediction methods. Section III describes each component of the proposed STGFT and gives its overall architecture. Section IV introduces the experimental datasets and conducts the comparative experiments. Section VI concludes this work.

## II. Related Work

Water quality plays a vital role in aquatic ecosystems because it can affect the growth of aquatic organisms and reflect the extent of water pollution [4]. Accurate water quality predictions are essential for environmental monitoring, sustainable ecosystem development, and human health. The main purpose of water quality prediction is to predict water bodies' key water quality elements in the future, *e.g.*, dissolved oxygen, total phosphorus, and total nitrogen. It is worth noting that water quality prediction belongs to the field of time series forecasting. The methods of water quality prediction can also be applied to other fields like traffic flow forecasting. Scholars have studied statistical, machine learning, and deep learning methods for time series forecasting in recent years.

### A. Statistical Methods on Time Series Forecasting

Multiple Linear Regression (MLR) and Auto-Regressive Integrated Moving Average (ARIMA) are two common statistical methods used in time series forecasting. MLR uses past data on independent and dependent variables to model linear relationships, while ARIMA uses past data on dependent variables to model time series. Francis et al. [5] describe the mathematical relationships between several physicochemical parameters in water quality to determine these parameters with minimal equipment. Specifically, the authors analyze seven physicochemical parameters weekly for two drinking water sources (tap and well water) stored in containers for six weeks. MLR is utilized to investigate the statistical relationships between these factors. However, the model cannot simulate the situation as time changes, and the model cannot be easily adjusted according to different situations. Wang et al. [6] introduce a Holt-Winters seasonal model that builds on the ARIMA time series framework. The authors develop a comprehensive water quality forecast model incorporating eutrophication indicators, including total phosphorus and nitrogen as key parameters. However, it is a static model and cannot be applied to predict other water quality indicators, thus limiting its scalability.

In practical applications, changes in water quality data often present complex nonlinear and nonsmooth characteristics. Traditional linear statistical models have limited ability to fit this kind of data, and they are difficult to capture the complex relationships between multiple variables. Moreover, the mechanism water quality prediction models are modeled through a physical approach, and they cannot be adjusted promptly in facing sudden water quality problems such as heavy rainfall, resulting in its incapability of accurately predicting subsequent changes in water quality. In addition, the mechanism model cannot make real-time predictions, which limits its applicability.

### B. Machine Learning Methods on Time Series Forecasting

To solve the problems of statistical methods on water quality prediction, some researchers have applied machine learning

methods, *e.g.*, Principal Component Analysis (PCA), Support Vector Regression (SVR), and decision trees. Batur et al. [7] investigate evaluating surface water quality indicators through applying PCA for data integration and extraction techniques. It focuses on several key quality metrics of water quality. By merging satellite images' spectral and spatial resolution capabilities, this study leverages data mining methodologies to generate enhanced images with refined spatial detail and temporal frequency. The prediction of surface water quality metrics employs a PCA-driven response surface regression approach. Nevertheless, it is important to note that the dataset underpinning this analysis is exclusively derived from summertime observations. Consequently, the observational timeframe needs to be extended to ensure the development of a more comprehensive model that encapsulates variability across different seasons. Su et al. [8] integrate an enhanced version of the sparrow search algorithm called improved sparrow search algorithm (ISSA), with the SVR model to predict future water quality. Specifically, ISSA is deployed to select the SVR's penalty factor and kernel function parameters, thereby improving the model's predictive precision and generalization ability. However, the assessment of water quality is influenced by various intricate factors. In this research, only five variables impacting water quality classification are chosen from the collected data without considering the potential interrelationships among these influencing factors. Lu amd Ma [9] propose hybrid models that integrate decision trees within the machine learning framework to enhance the precision of time series prediction. The core algorithms of these hybrid models include extreme gradient boosting and random forest, each enhanced with an advanced data-denoising approach. Their work analyzes water samples collected from the Tualatin River Basin. The proposed model aims to predict key water quality metrics, *e.g.*, water temperature, and dissolved oxygen. Nonetheless, this method should incorporate additional variables influencing water quality, achieving long-term water quality prediction. Furthermore, the computational complexity of this approach is relatively high, posing challenges for real-time prediction. Zhan et al. [10] integrate meteorological and hydrological data, and perform correlation analysis between meteorological and hydrological data features. Then, the correlation matrix between these features is derived. Finally, an SVR model using the radial basis function as a kernel function is used for prediction. Since the method combines meteorological and hydrological data, it can better adapt to contingency situations. However, it does not consider the spatial characteristics of water quality data, which leads to difficulties in dealing with complex river network structures.

However, there are challenges when using machine learning methods to predict water quality. The results obtained from them may contain errors due to statistical inferences. Most machine learning methods generate results based on previously fixed patterns. Thus, any new experience or data may not have accurate predictions. Moreover, the initial investment of time in training machine learning algorithms is huge, and the demand for data is enormous. Therefore, the training and maintenance of these models have high complexity [11]. Even minor logical inaccuracies can give rise to significant defects in the machine learning workflow, leading to erroneous outputs. Furthermore, traditional machine learning models usually use shallow models with limited ability to model complex patterns and relationships. They are also prone to overfitting problems when the number of features is large, compromising their performance and generalizability.

### C. Deep Learning Methods on Time Series Forecasting

Deep learning introduces deep neural networks that can learn more abstract features and deal with complex problems. In addition, they have better generalization ability than machine learning methods. Therefore, deep learning is widely applied in the time series prediction. Jiang et al. [12] present a dynamic temporal dependency model, which modifies the transformer architecture. With its encoder-decoder framework, the model allows for flexible adjustments to historical data and forecast ranges, making it easier to learn multi-step temporal dependencies and minimize error accumulation. Xu et al. [13] introduce a long short-term memory (LSTM) variant incorporating a spatial autocorrelation and the nonlocal attention module for predicting vegetation indices. This approach utilizes the nonlocal attention mechanism to capture long-range dependencies, while the spatial autocorrelation modeling leverages the local Moran index to understand spatial relationships. However, the above studies only focus on temporal or spatial features and do not integrate them, making the model more effective in short-term prediction but not applicable to long-term time series prediction. To avoid this drawback, spatiotemporal prediction methods are widely applied. Qiao et al. [14] design a novel spatiotemporal prediction model named fusion spatiotemporal GCN network. This model employs a temporal attention mechanism to address the nonlinear nature of water quality time series. In addition, the model utilizes graph convolution to capture spatial dependencies within river networks. The integration of spatiotemporal fusion facilitates the capture of comprehensive spatiotemporal features. Furthermore, a temporal convolution residual mechanism is incorporated, enhancing the accuracy of long-term series prediction. However, this method is not well fused with spatial and temporal features. It cannot adaptively select the most relevant input features and appropriately capture long-term temporal dependence of the water quality data.

In addition, several studies have introduced spatial encoders and GCNs to better reflect the spatial correlations. Chen et al. [15] focus on predicting spatial-temporal air pollutants by developing an adaptive adjacency matrix-based graph convolutional recurrent network. This model integrates points of interest and meteorological data into a fully connected neural network to determine the weights of the adjacency matrix. The pollutant data and the adjacency matrix are then fed into the GCN unit, which is integrated with LSTM units to capture spatiotemporal dependencies. However, when extracting the spatial relationship, the model does not incorporate a graph completely detached from prior knowledge, and some potential spatial features may be lost. Shen and Yoon [16] apply spatiotemporal fusion prediction to traffic flow prediction. The authors note that traditional graph structures are trained in the training phase and do not reflect the data used in the

testing phase. This shortcoming is particularly prominent in traffic prediction due to the frequent unexpected changes in traffic data and irregularities in the time series. Therefore. The authors propose a novel traffic prediction framework named progressive GCN. It constructs a set of graphs using the online input data during the training and testing. However, this work does not consider the predefined road network structure, and the summation of the representations extracted from different subgraphs to obtain the final graph structure loses some vital information in the final graph structure. Yang et al. [17] focus on spatiotemporal gas turbine energy consumption prediction. The authors adopt a graph structure with a fusion strategy to solve the irregularity problem of multi-source sensor data and then propose a diffusion graph network with an adaptive neighbor matrix. The adaptive neighbor matrix module captures the temporal trend of spatial information based on the input data. The diffusion graph cycle module can memorize historical sequences. However, this work is entirely detached from the prior experience in constructing the adaptive graph, which leads to the poor robustness of the model.

Unlike the above studies, this work proposes a spatiotemporal water quality prediction model based on graph attention networks to predict the trends of key water quality indicators in aquatic environments. The performance of the graph attention network is highly dependent on the quality of the employed graph structure. To better reflect the actual spatial characteristics of the river network, the proposed ADMG generates three graphs for mutual fusion. The first one is a predefined graph generated by manually measuring the distance of each monitoring station. It can reflect the actual geospatial characteristics of the monitoring stations. However, due to the complexity of the spatial structure of river networks, the relationship between upstream and downstream may even change due to the influence of rainfall and other meteorological factors. Thus, the randomly initialized matrix introduces and trains the adaptive graph. It is entirely detached from the prior knowledge of the predefined map. The potential spatial characteristics and the change pattern are explored. The dynamic graph combines the prior knowledge to extract the spatial features more comprehensively. It captures spatial relationships over time, reflecting dynamic dependency structures in river networks. These three graphs and temporal features cooperate to predict water quality accurately.

## III. Proposed Methodology

For water quality prediction, it is shown in Fig. 1 that due to the complex upstream and downstream spatial influence of the river network, the future water quality of a region is affected by its historical water quality and the historical water quality of other monitoring stations. Therefore, combining spatial and temporal features is necessary to predict future water quality accurately. This section introduces the overall framework of the proposed STGFT. It then introduces the structure of TAE that captures the temporal features of each water quality monitoring station. It next introduces the structure of ADMG based on SAE that learns the correlation among water quality monitoring stations and generates adaptive and dynamic adjacency matrices to mine the potential spatial dependencies in river networks. It finally discusses the combination of the above modules to construct STGFT. Its notations are summarized in Table I.

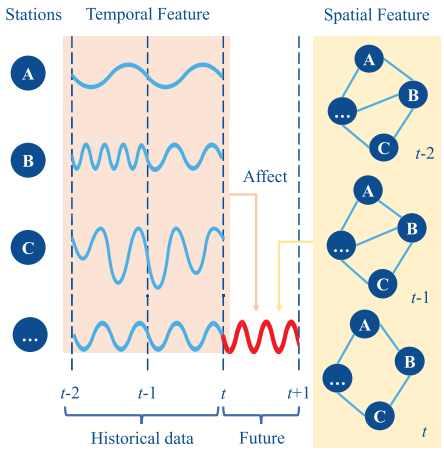



Fig. 1. Water quality prediction.

TABLE I
Main Notations

| Notation | Definition |
|---|---|
| $X$ | Historical temporal feature data |
| $X'$ | Feature embedding vector |
| $Q_T, K_T, V_T$ | Query, key and value vectors of TAE |
| $\mathrm{Softmax}(\cdot)$ | Softmax function |
| $\mathbb{A}(\cdot)$ | Self-attention mechanism |
| $W_T^Q, W_T^K, W_T^V$ | Trainable parameter matrices in TAE |
| $d_k$ | Scaling factor |
| $\mathrm{h}_T(i), \mathrm{h}_S(i)$ | Output of the self-attention mechanism of group $i$ in TAE and SAE, respectively |
| $W_T^Q(i), W_T^K(i), W_T^V(i)$ | Trainable parameter matrices in the group $i$ of self-attention mechanisms of TAE |
| $O_T$ | Output result of the TAE |
| $\mathrm{r}^{(i)}$ | Residual result of group $i$ |
| $\mathrm{N_L}(\cdot)$ | Layer normalization |
| $W_{T_0}, W_{T_1}$ | Trainable parameter matrices in the feed-forward neural network |
| $\mathrm{ReLU}(\cdot)$ | Activation function |
| $\hat{X_S}$ | Node embedding vector |
| $W_S^Q, W_S^K, W_S^V$ | Trainable parameter matrices in SAE |
| $W_S^Q(i), W_S^K(i), W_S^V(i)$ | Trainable parameter matrices in the group $i$ of self-attention mechanisms of SAE |
| $O_S$ | Output result of the SAE |
| $A, A_P, A_D$ | Predefined adjacency matrix, adaptive adjacency matrix and dynamic adjacency matrix |
| $\mathrm{FW}_t$ and $\mathrm{FW}_d$ | Two feed forward networks |
| $H$ | Number of heads of attention mechanism |
| $E$ | Embedding dimension |
| $G$ | GCN output dimension |

### A. Temporal Attention Encoder (TAE)

In water quality prediction tasks, historical data can affect the future trend of change, and the monitoring values at different time steps also have different impacts on the future water quality change. For example, when the rainfall is excessive during the flood season, some pollutants enter the river with the rainwater, leading to a significant deterioration of water quality, and it affects the subsequent changes. In this case,

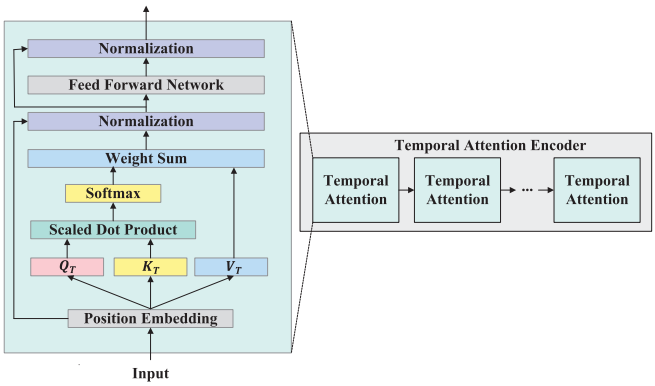



Fig. 2. Structure of the TAE.

to capture the correlation of water quality elements between different time steps, this work designs TAE that learns the temporal features of each water quality monitoring station. TAE's structure is shown in Fig. 2. It includes multiple temporal attention layers, and they are stacked together. Each layer mainly includes position embedding, linear transformation, calculation of attention weights, normalization, and a feed-forward network. It is assumed that there are $N$ water quality monitoring stations and $C$ water quality elements. Before entering the first temporal attention layer, a feature embedding vector $X' \in \mathbb{R}^{N\times T\times D}$ is generated based on the historical temporal feature data $X=\{X_{:,1}, X_{:,2}, X_{:,t}, \ldots, X_{:,T}\} \in \mathbb{R}^{N\times T\times C}$ of water quality monitoring stations, where $\mathbb{R}$ denotes a set of real numbers, $T$ denotes the number of time steps, and $D$ denotes the embedding dimension. Then, the positional embedding ($P$) [18] is obtained as $P_{(p,2m)}=\sin(p/10000^{2m/D})$ and $P_{(p,2m+1)}=\cos(p/10000^{2m/D})$, where $p$ denotes the position number and $m$ denotes the current dimension number. Moreover, $P_{(p,2m)}$ and $P_{(p,2m+1)}$ occur alternately, and they are obtained by the sin function ($sin(\cdot)$) and the cos function ($cos(\cdot)$), respectively.

Then, the input feature embedding is added to the positional embedding, obtaining the input of the temporal attention layer ($\hat{X}_T=(X'+P)\in\mathbb{R}^{N\times T\times D}$). In the temporal attention layer, a self-attention mechanism [19] is adopted to extract the internal correlation of the historical sequence data for $N$ sites in parallel. It calculates the similarities between different timestamps, capturing the dynamic temporal dependencies among neighboring timestamps. It allows the model to be more flexible with inputs from different time steps and better adapt to the requirements of other tasks. First, $\hat{X}_T$ is mapped to three different feature spaces, obtaining a query vector $Q_T\in\mathbb{R}^{N\times T\times D}$, a key vector $K_T\in\mathbb{R}^{N\times T\times D}$, and a value vector $V_T\in\mathbb{R}^{N\times T\times D}$. $Q_T$ denotes the current focus of attention, indicating the value to be predicted for the time step. $K_T$ denotes the information about the historical timesteps, which is used to match with $Q_T$. $V_T$ denotes a vector containing the actual information corresponding to $K_T$. Then, the scaled dot product is used to calculate the attention intensity of each time step for other time steps on $Q_T$, $K_T$, and $V_T$. Softmax$(\cdot)$ is adopted for normalization, obtaining the attention coefficient. Finally, the attention coefficient is multiplied by $V_T$, resulting in the output of the self-attention mechanism ($\mathbb{A}(\cdot)$). The specific calculation process is as follows:

$$Q_T = \hat{X}_T W_T^Q \tag{1}$$

$$K_T = \hat{X}_T W_T^K \tag{2}$$

$$V_T = \hat{X}_T W_T^V \tag{3}$$

$$\mathbb{A}(Q_T, K_T, V_T) = \text{Softmax}\left(\frac{Q_T K_T^{\mathsf{T}}}{\sqrt{d_k}}\right) V_T \tag{4}$$

where $W_T^Q$, $W_T^K$, and $W_T^V$ represent trainable parameter matrices, $d_k$ represents a scaling factor, and it is the size of the first dimension of $K_T$. Moreover, to capture the complex features of water quality, a multi-head attention mechanism [20] is adopted in the temporal attention layer, *i.e.*, training $I$ groups of self-attention mechanisms while later concatenating the results and then remapping them back to the original dimensions. $\text{h}_T(i)$ represents the output of the self-attention mechanism of group $i$. The specific calculation process is as follows:

$$\text{h}_T(i) = \mathbb{A}\left(W_T^Q(i)\hat{X}_T, W_T^K(i)\hat{X}_T, W_T^V(i)\hat{X}_T\right) \tag{5}$$

$$\mathbb{H}\left(\hat{X}_T\right) = \|_{i=1}^{I}(\text{h}_T(i))W_T^O \tag{6}$$

where $W_T^Q(i)$, $W_T^K(i)$, and $W_T^V(i)$ represents trainable parameter matrices in the group $i$ of self-attention mechanisms, and $W_T^O$ is a trainable parameter matrix. Based on the idea of residual connection [21], the output of the multi-head attention mechanism ($\mathbb{H}(\hat{X}_T)$) is added to $\hat{X}$. Then, it passes layer normalization [22] and a feed-forward neural network. Finally, the output result of TAE ($O_T$) is obtained after normalization, *i.e.*,

$$Z = \begin{cases} \hat{X}, & i=0 \\ O_T^{(i-1)}, & \text{otherwise} \end{cases} \tag{7}$$

$$\text{r}^{(i)} = \text{N}_\text{L}(\mathbb{H}(Z)+Z) \tag{8}$$

$$O_T^{(i)} = \text{N}_\text{L}\left(W_{T_1}^{(i)}\,\text{ReLU}\left(W_{T_0}^{(i)}\text{r}^{(i)}\right)+\text{r}^{(i)}\right) \tag{9}$$

where $\text{r}^{(i)}$ denotes the residual result of group $i$. $\text{N}_\text{L}(\cdot)$ represents layer normalization. $W_{T_0}$ and $W_{T_1}$ represent trainable parameter matrices in the feed-forward neural network. ReLU$(\cdot)$ means the activation function ReLU.

### *B. Adaptive Dynamic Adjacency Matrix Generator Based on Spatial Attention Encoder*

*1) Spatial Attention Encoder:* Water quality monitoring sensors are widely distributed in rivers and lakes, and the water quality conditions of the downstream are often affected by the upstream water quality. An SAE is proposed to capture the correlation between each water quality monitoring station to effectively mine the potential spatial features of the water quality data. The structure of the SAE is shown in Fig. 3. It includes multiple spatial attention layers, and they are stacked together. Each layer mainly includes position embedding, GCN, linear transformation, calculation of attention weights, and a feed-forward network. Specifically, the predefined adjacency matrix $A$ and $X'$ are used as the input of the GCN [23] layer, resulting in a node embedding vector $\hat{X}_S$. Fig. 4 shows an example of the construction process of $A$

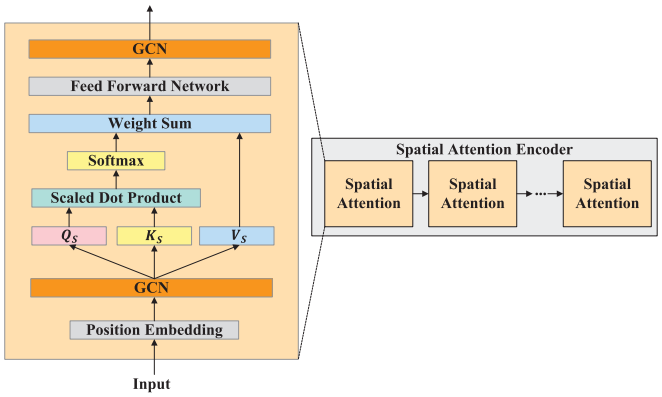



Fig. 3. Structure of the SAE.

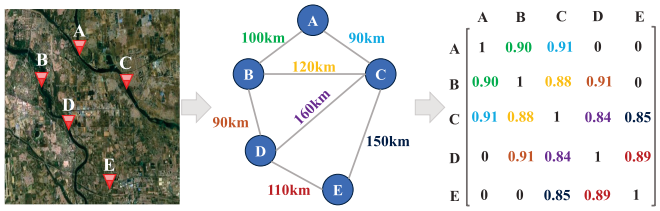



Fig. 4. Construction of the predefined graph.

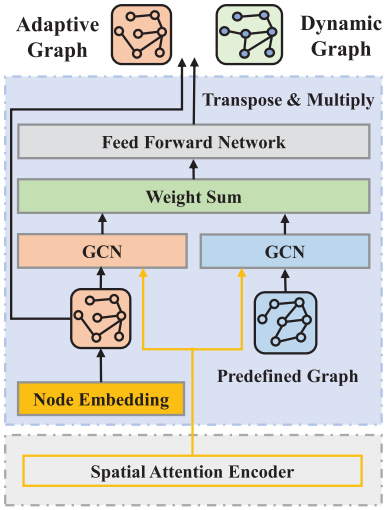



Fig. 5. Structure of the ADMG.

with five monitoring stations. It constructs a graph structure by weighing the distances between each monitoring station after measuring them manually. $A$ shows the geospatial structure of river networks. Specifically, each node denotes a monitoring station, and the value denotes the impact of the two monitoring stations. It is normalized to [0,1], where the larger value indicates that the relationship between these two stations is greater. Next, similar to the temporal attention layer, parameter matrices $W_S^Q$, $W_S^K$, and $W_S^V$ are adopted to map the $\hat{X}_S$ to three different feature spaces, resulting in query vector, key vector, and value vector. For SAE, $Q_S$ denotes the current focus of attention, indicating the spatial features that need to be predicted for the time step. $K_S$ denotes the spatial information of the historical time step, which is the corresponding spatial feature obtained from the input data. $V_S$ denotes a vector of values containing the actual spatial information corresponding to $K_S$. Then, the scaled dot product is used to calculate the attention coefficient. After that, it performs a weighted summation on the value vector, resulting in the output result of the self-attention mechanism. Finally, the results pass through a feed-forward neural network [24], obtaining the spatial attention $\mathbb{A}_S$, *i.e.*,

$$\hat{X}_S = \phi\left(X', A\right) \tag{10}$$

$$\mathrm{h}_S(i) = \mathbb{A}\left(W_S^Q(i)\hat{X}_S, W_S^K(i)\hat{X}_S, W_S^V(i)\hat{X}_S\right) \tag{11}$$

$$\mathbb{H}\left(\hat{X}_S\right) = \|_{i=1}^{I}(\mathrm{h}_S(i))W_S^O \tag{12}$$

$$\mathbb{A}_S = W_{S_1}\,\mathrm{ReLU}\left(W_{S_0}\left(\mathbb{H}\left(\hat{X}_S\right)\right)\right) \tag{13}$$

where $\phi$ denotes a GCN, $\mathrm{h}_S(i)$ represents the output of the self-attention mechanism of group $i$. $W_S^Q(i)$, $W_S^K(i)$ and $W_S^V(i)$ represent its trainable parameter matrices. $W_S^O$, $W_{S_1}$ and $W_{S_0}$ represent parameter matrices. The above result is used as an input feature of the GCN layer, thus extracting spatial features. This process is shown as $O_S{=}\phi(\mathbb{A}_S, A)$. Specifically, this work defines the output of the last stacked spatial attention layer as $O_S$.

*2) Adaptive Dynamic Adjacency Matrix Generator:* Due to the high complexity and uncertainty of spatial relationships in river networks, the predefined graph structure cannot reflect the real spatial relationships. Therefore, ADMG is designed to generate adaptive and dynamic adjacency matrices to mine the potential spatial dependencies in river networks. As shown in Fig. 5, ADMG first uses a randomly initialized vector $E{\in}\mathbb{R}^{N\times D}$ to construct adaptive adjacency matrix $A_P{\in}\mathbb{R}^{N\times N}$, the deficiency of $A$ in representing node relationships is compensated by constructing it. Moreover, $A_P$ is fixed after training. Furthermore, the $O_S$ is input into two parallel GCNs to construct a dynamic adjacency matrix. They take $A_P$ and $A$ as parameters, obtaining the dynamic feature mapping $F_d{\in}\mathbb{R}^{N\times T\times D}$. This process is shown as follows, where $\alpha$ and $\beta$ are trainable parameters. They are used to weigh the output results of the two GCNs.

$$A_P = \mathrm{Softmax}(\mathrm{ReLU}(E\cdot E^{\mathsf{T}})) \tag{14}$$

$$F_d = \alpha\phi(O_S, A_P){+}\beta\phi(O_S, A) \tag{15}$$

Then, $F_d$ is converted into a two-dimensional matrix ($F_d'{\in}\mathbb{R}^{(D\times T)\times N}$), and a linear transformation [25] is performed on $F_d'$ to obtain a dynamic feature of a specific dimension ($\tilde{F}_d{\in}\mathbb{R}^{N\times f}$), where $f$ denotes the number of linear layers. Then, a dynamic embedded $E_d{\in}\mathbb{R}^{N\times f}$ is generated by $\tilde{F}_d$ and $E$. We have:

$$\tilde{F}_d = W_f F_d' \tag{16}$$

$$E_d = \mathrm{ReLU}(\mathrm{Tanh}(\tilde{F}_d \odot E)) \tag{17}$$

where $W_f{\in}\mathbb{R}^{f\times(D\times T)}$ represents trainable parameters in the linear layer, $\mathrm{Tanh}(\cdot)$ denotes the hyperbolic tangent function, and $\odot$ represents the Hadamard product. Finally, $E_d$ is multiplied by its transpose matrix $E_d^{\mathsf{T}}$ to generate a dynamic adjacency matrix $A_D{\in}\mathbb{R}^{N\times N}$, *i.e.*,

$$A_D{=}\mathrm{ReLU}(\mathrm{Tanh}(E_d\cdot E_d^{\mathsf{T}})) \tag{18}$$

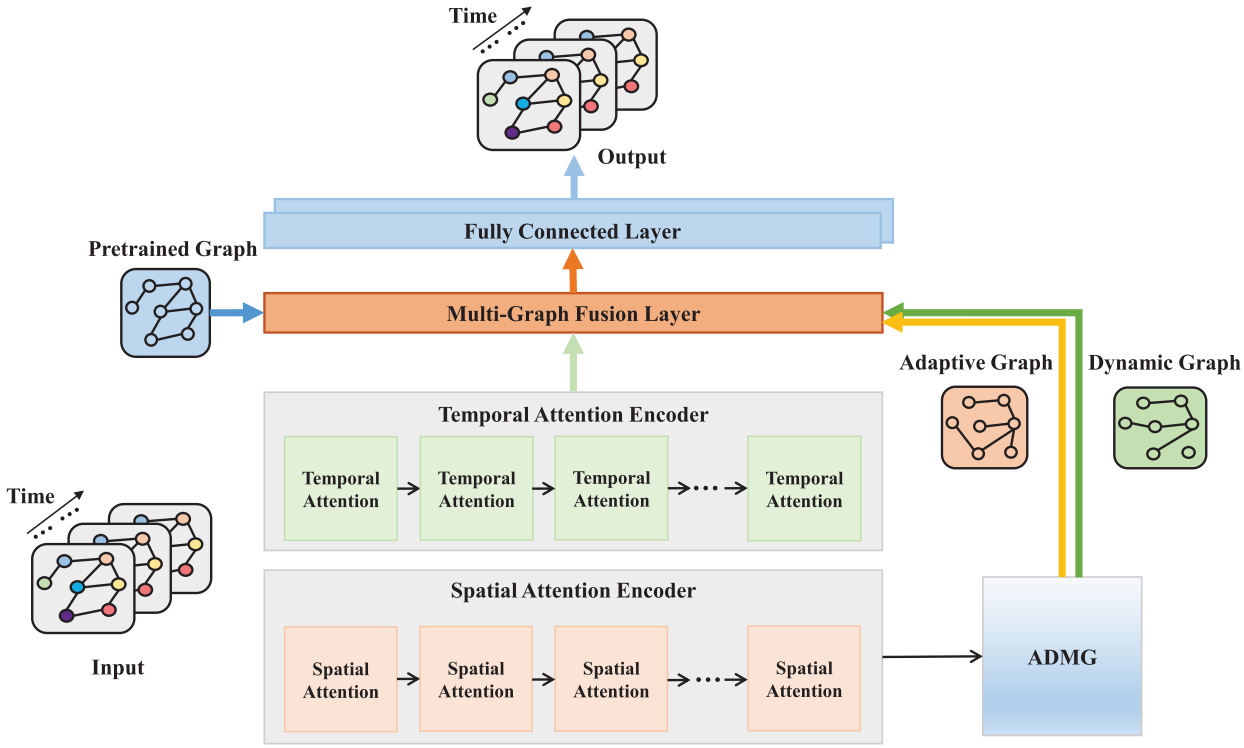



Fig. 6. Overall framework of STGFT.

### C. Spatial-Temporal Graph Fusion Transformer (STGFT)

Sections III-A and III-B have described the three main components of STGFT, *i.e.*, TAE, SAE, and ADMG. This section introduces the overall architecture of STGFT. Fig. 6 shows its architecture. The original water quality sequence data $X$ is the input in parallel to TAE and SAE, obtaining the temporal features $O_T$ that contain the correlations among different time steps and the spatial features $O_S$ that contains the spatial correlation among different nodes. Then, $O_S$ is used as the input of ADMG and obtains the adaptive adjacency matrix $A_P$ and the dynamic adjacency matrix $A_D$. Moreover, the multi-graph fusion layer adopts three parallel GCNs to fuse $A_P$, $A_D$, and $A$ and generate node embeddings $F_l$. Then, we have $F_l = \mu(\phi(A, O_T)) + \nu(\phi(A_P, O_T)) + \omega(\phi(A_D, O_T))$, where $\mu$, $\nu$, and $\omega$ are parameters adopted to weight the output results of three GCNs and they are obtained through the training.

After that, $F_l$ is decoded using feed forward networks in the fully connected layer, predicting the future water quality sequence data $Y$. The specific process is shown as follows:

$$Y' = \mathrm{FW}_t(F_l) = W_t^1 \,\mathrm{ReLU}\left(W_t^0 F_l\right) \tag{19}$$

$$Y = \mathrm{FW}_d\left(Y'\right) = \mathrm{ReLU}\left(Y' W_d^0\right) W_d^1 \tag{20}$$

where $\mathrm{FW}_t$ and $\mathrm{FW}_d$ represent two feed forward networks, $\mathrm{FW}_t$ is used to transform the time dimension, converting $F_l$ into a vector $Y'$ of the target prediction length, $\mathrm{FW}_d$ is used to transform the water quality feature dimension, converting $Y'$ into a vector $Y$ of the target feature dimension. $W_t^0$, $W_t^1$, $W_d^0$, and $W_d^1$ represent training parameter matrices.

TABLE II
OVERVIEW OF WATER QUALITY PREDICTION DATASETS

| Dataset | Station Numbers | Sample Size | Frequency | Indices |
|---|---|---|---|---|
| BTH | 24 | 233,440 | 4 hours | TN, TP, DO |
| Beijing | 6 | 58,350 | 4 hours | TN, TP, DO |
| Alabama | 5 | 99,315 | 1 hours | DO |

## IV. EXPERIMENTS AND RESULTS ANALYSIS

### A. Dataset Selection and Parameter Tuning

*1) Dataset Description:* Three real-world water quality datasets are selected to verify the effectiveness of the STGFT, *i.e.*, Alabama, Beijing, and Beijing-Tianjin-Hebei (BTH) datasets. Table II shows the overview of water quality prediction datasets. Specifically, Alabama datasets comprise the historical dissolved oxygen data from May 2017 to Aug. 2019 at the Cahaba River station with 1-hour sampling intervals. The Beijing dataset includes the total nitrogen data of six water quality monitoring stations in Beijing from Oct. 2018 to Aug. 2022. The BTH dataset contains more complex spatial relationships than the Alabama and Beijing datasets. It includes 24 water quality monitoring stations in different administrative divisions of the Beijing-Tianjin-Hebei region in China. Moreover, it contains 48 edges and 9,275 sampling points, covering the total nitrogen historical time series data sampled every 4 hours from Oct. 2018 to Dec. 2022. In addition, Table III shows an example of the detailed data structure of the three datasets on one day. The satellite map of water quality monitoring stations in the Beijing-Tianjin-Hebei region is shown in Fig. 7. Moreover, Table IV lists their geographical information, including the regions, longitude and latitude, and

TABLE III
DATASET DESCRIPTION

| Timestamp | BTH (TN) | | | | Beijing (TN) | | | | Alabama (DO) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | S1 | S2 | ... | S24 | S1 | S2 | ... | S6 | S1 | S2 | ... | S5 |
| 2018/10/8 0:00 | 6.45 | 6.27 | ... | 9.25 | 2.65 | 7.27 | ... | 1.2 | 7.6 | 9.3 | ... | 8.6 |
| 2018/10/8 4:00 | 5.74 | 6.36 | ... | 9.21 | 2.33 | 6.49 | ... | 1.2 | 7.5 | 8.9 | ... | 8.4 |
| 2018/10/8 8:00 | 6.21 | 6.09 | ... | 9.26 | 1.78 | 6.24 | ... | 1.16 | 7.4 | 8.8 | ... | 8.6 |
| 2018/10/8 12:00 | 6.76 | 5.86 | ... | 9.12 | 2.04 | 6.87 | ... | 1.16 | 7.4 | 8.9 | ... | 8.5 |
| 2018/10/8 16:00 | 5.82 | 5.93 | ... | 9.18 | 1.7 | 6.65 | ... | 1.16 | 8.2 | 9.1 | ... | 18.4 |
| 2018/10/8 20:00 | 5.95 | 6.28 | ... | 9.33 | 2.25 | 6.68 | ... | 1.16 | 8.5 | 8.9 | ... | 8.6 |

TABLE IV
GEOGRAPHIC INFORMATION OF WATER QUALITY MONITORING STATIONS IN THE BTH DATASET

| Serial Number | Name of Monitoring Station | Belonging Region | Longitude | Latitude | Belonging basin |
|---|---|---|---|---|---|
| 1 | Beiyang Bridge | Hebei District, Tianjin | 117°10’7.68” East | 39°9’48.96”N | Haihe River basin |
| 2 | Caozhuangzi pumping station | Xiqing District, Tianjin | 117°5’27.60” East | 39°8’27.96”N | Haihe River basin |
| 3 | Dahongmen Gate | Fengtai District, Beijing | 116°25’5.52” East | 39°49’58.44” N | Haihe River basin |
| 4 | Dahongqiao | Hongqiao District, Tianjin | 117°8’25.44” East | 39°9’53.64” N | Haihe River basin |
| 5 | Dawangwu | Langfang, Hebei Province | 116°49’17.76” East | 39°23’59.28” N | Haihe River basin |
| 6 | Gujiaying | Yanqing District, Beijing | 115°48’49.99” East | 40°24’43.37” N | Haihe River basin |
| 7 | Drumtower outer street | Dongcheng District, Beijing | 116°24’8.82” East | 39°57’20.27” N | Haihe River basin |
| 8 | Guangbei Riverfront Road (bridge) | Xicheng District, Beijing | 116°20’31.92” East | 39°53’57.84” N | Haihe River basin |
| 9 | Guohe Bridge | Jizhou District, Tianjin | 117°43’34.32” East | 40°1’15.96” N | Haihe River basin |
| 10 | Haihe River sluice gate | Binhai District, Tianjin | 117°42’24.12” East | 38°59’14.64” N | Haihe River basin |
| 11 | HouCheng | Fengtai District, Beijing | 116°9’10.80” East | 40°38’51.36” N | Haihe River basin |
| 12 | Garden Road | Haidian District, Beijing | 116°22’14.16” East | 39°58’55.20” N | Haihe River basin |
| 13 | Huairoushuiku | Huairou District, Beijing | 116°36’3.60” East | 40°18’14.40” N | Haihe River basin |
| 14 | Jinggang Mountains Bridge | Hongqiai District, Tianjin | 117°9’7.20” East | 39°9’2.52” N | Haihe River basin |
| 15 | Miyunshuiku | Miyun District, Beijing | 116°54’51.12” East | 40°29’44.52” N | Haihe River basin |
| 16 | Nandahuang Bridge | Shijingshan District, Beijing | 116°10’23.32” East | 39°53’58.43” N | Haihe River basin |
| 17 | Qing Jing Yellow Moisture Barrier | Binhai District, Tianjin | 117°31’53.76” East | 38°39’46.80” N | Haihe River basin |
| 18 | Qinghe Gate | Haidian District, Beijing | 116°21’18.36” East | 40°1’56.96” N | Haihe River basin |
| 19 | Sanchakou | Haidian District, Beijing | 117°10’58.80” East | 39°8’59.28” N | Haihe River basin |
| 20 | Sanxiaoying Gate | Langfang, Hebei Province | 116°31’5.16” East | 39°35’20.04” N | Haihe River basin |
| 21 | Shahe Bridge | Jizhou District, Tianjin | 117°46’5.88” East | 40°3’34.20” N | Haihe River basin |
| 22 | Shawo | Chaoyang District, Beijing | 116°37’23.88” East | 39°56’44.16” N | Haihe River basin |
| 23 | Wucun | Langfang, Hebei Province | 116°56’9.60” East | 39°48’7.20” N | Haihe River basin |
| 24 | Yuqiao Reservoir Outlet | Jizhou District, Tianjin | 117°26’3.84” East | 40°1’54.12” N | Haihe River basin |

basins of each water quality monitoring station. It is worth noting that this work adopts the same data preprocessing method for each dataset, and each dataset is divided into training, validation, and testing sets in the ratio of 70%, 10%, and 20%. The input length of each sample is 40, and the output length is 10, *i.e.*, 40 historical time steps of data are used to predict 10 future time steps of data.

*2) Parameter Tuning:* To optimize the prediction performance of the STGFT, some hyperparameters need to be manually adjusted. These hyperparameters include the number of heads of the multi-head attention mechanism ($H$), embedding dimension ($E$), and GCN output dimension of the multi-graph fusion layer ($G$). Therefore, this section selects the optimal combination of parameters for STGFT through experiments.

The multi-head attention mechanism allows STGFT to perform attention calculation in multiple subspaces in parallel, allowing the model to concentrate on different subspace information simultaneously, thereby enhancing the model’s generalization and representation abilities. An appropriate $H$ can help improve the model’s overall predictive performance. This work sets $H \in \{1, 2, 4\}$. Moreover, the embedding dimension has an important impact on the model’s representation ability and computational efficiency. A small embedding dimension may lose information and reduce the accuracy of predictions, while a large one may cause the model to fall into local minima. Therefore, adjusting the $E$ during the training process is necessary. This work lets $E \in \{8, 16, 32\}$. Finally, $G$ is selected from $\{8, 16, 32, 64\}$. Table V shows the Root Mean Square Error (RMSE) [26], Mean Absolute Error (MAE) [27], and Mean Absolute Percentage Error (MAPE) [28] for the predicted values of STGFT compared to the true values. It is shown that STGFT achieves the best prediction accuracy when $H$, $E$, and $G$ are set to 2, 16, and 16, respectively.

### B. Comparative Experiments

This experiment is conducted on a server with an Intel Xeon 6248R CPU and a GTX3090 GPU. The code for the

Fig. 7. The satellite map of water quality monitoring stations in the Beijing-Tianjin-Hebei region.

TABLE V
PREDICTED EFFECTS OF STGFT WITH DIFFERENT SETS OF HYPERPARAMETERS

| $(H, E, G)$ | RMSE | MAE | MAPE |
|---|---|---|---|
| (1, 8, 8) | 0.3249 | 0.2055 | 0.0595 |
| (1, 16, 16) | 0.3029 | 0.1856 | 0.0544 |
| (1, 16, 32) | 0.3091 | 0.1949 | 0.0635 |
| (2, 8, 16) | 0.3053 | 0.1835 | 0.0526 |
| (2, 8, 16) | 0.2732 | 0.1725 | 0.0553 |
| **(2, 16, 16)** | **0.2562** | **0.1512** | **0.0435** |
| (2, 16, 32) | 0.2865 | 0.1871 | 0.0603 |
| (2, 16, 64) | 0.2603 | 0.1861 | 0.0524 |
| (2, 32, 64) | 0.3085 | 0.1802 | 0.0505 |
| (4, 16, 16) | 0.2828 | 0.1740 | 0.0523 |
| (4, 16, 32) | 0.2958 | 0.1944 | 0.0603 |
| (4, 32, 64) | 0.3253 | 0.2054 | 0.2054 |

model is written in Pytorch framework. The batch size is set to 64. Moreover, the dropout of the model is set to 0.3 to prevent the overfitting problem. During the training process, the model is trained by using the Adam optimizer with the learning rate initialized to 0.01 and the weight decay to $1\times10^4$. To verify the effectiveness of STGFT, four baseline models, which are recent and represent the state-of-the-art are adopted for comparative experiments, *i.e.*, Attention-based Spatial-Temporal Graph Convolutional Networks (ASTGCN) [29], Graph WaveNet [30], Spatial-Temporal Synchronous Graph Convolutional Networks (STSGCN) [31], and Graph Attention WaveNet (GATWNet) [32]. Figs. 8 and 9 show the RMSE and MAE of STGFT and comparative models on prediction steps from 1 to 10.

Table VI shows the prediction error of STGFT and comparative models on Alabama, Beijing, and BTH datasets on one prediction step. It is shown in Figs. 8 and 9 that STGFT achieves the lowest RMSE and MAE on all prediction steps, which proves the predictions obtained by the STGFT are closer to the real values. Moreover, it is shown in Table VI that STGFT achieves the lowest prediction error on all datasets compared with the baseline models. Its RMSE on three datasets is reduced by an average of 10.63–19.74%, 1.69–23.97%, and 14.28–30.01% compared to the baseline

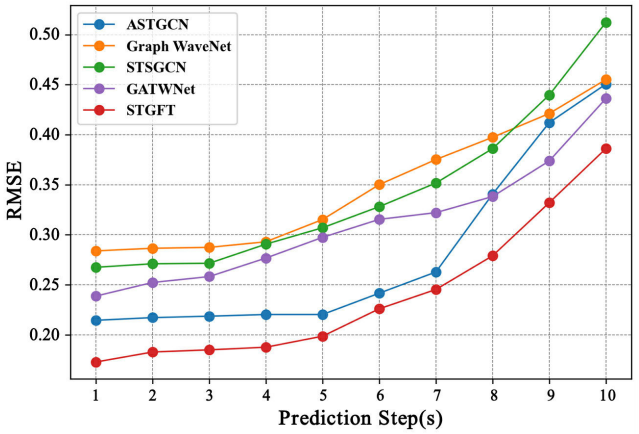



Fig. 8. Comparison of multi-step (1-10) prediction RMSE on BTH dataset.

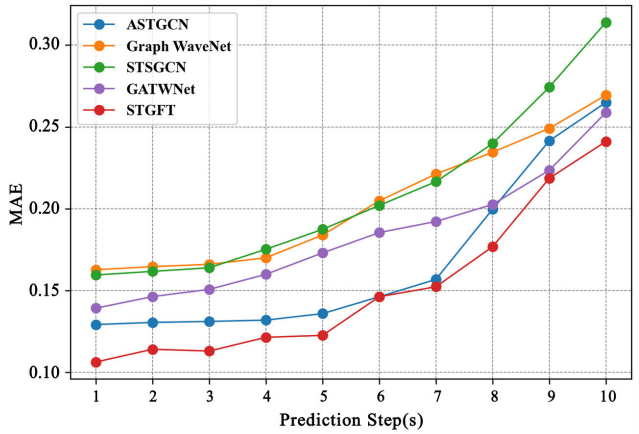



Fig. 9. Comparison of multi-step (1-10) prediction MAE on BTH dataset.

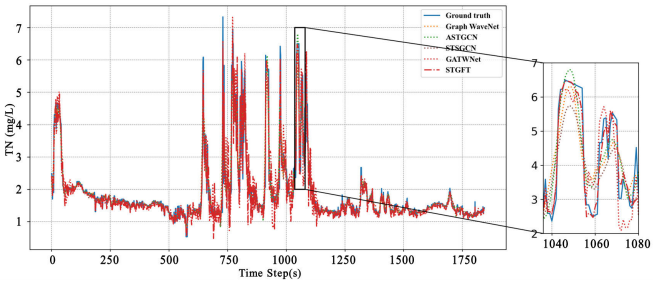



Fig. 10. Comparison of prediction results (Beiyang Bridge).

models, indicating that STGFT has higher accuracy and stability on water quality predictions. Furthermore, compared with experimental results on Alabama and Beijing datasets on a smaller spatial scale, STGFT has a greater improvement in prediction accuracy on the BTH dataset. This shows that STGFT can effectively capture time and potential spatial features in spatiotemporal water quality data as spatial scale increases. Fig. 10 shows the prediction effect of the STGFT and comparative models by drawing the prediction curve of one water quality monitoring station (Beiyang Bridge) in the BTH dataset. It is shown that the prediction result of the STGFT is closer to the true value, proving that STGFT has advantages in water quality spatial-temporal prediction.

TABLE VI
COMPARISON OF PREDICTIVE METRICS OF STGFT WITH OTHER BASELINE MODELS

| Model | Alabama | | | Beijing | | | BTH | | |
|---|---|---|---|---|---|---|---|---|---|
| | RMSE | MAE | MAPE | RMSE | MAE | MAPE | RMSE | MAE | MAPE |
| ASTGCN | 0.2010 | 0.1424 | 0.0185 | 0.5484 | 0.3525 | 0.0997 | 0.3302 | 0.1832 | 0.0482 |
| Graph WaveNet | 0.2072 | 0.1418 | 0.0187 | 0.5230 | 0.3176 | 0.0964 | 0.3661 | 0.2026 | 0.0537 |
| STSGCN | 0.2137 | 0.1430 | 0.0188 | 0.4376 | 0.2721 | 0.0781 | 0.3645 | 0.2094 | 0.0558 |
| GATWNet | 0.1919 | 0.1310 | 0.0164 | 0.4241 | 0.2565 | **0.0679** | 0.2989 | 0.1638 | 0.0436 |
| **STGFT** | **0.1715** | **0.116** | **0.0152** | **0.4169** | **0.2526** | 0.0713 | **0.2562** | **0.1512** | **0.0435** |

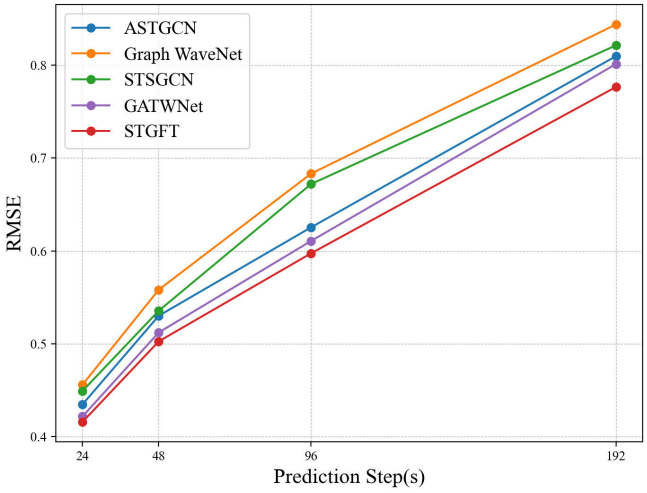



Fig. 11. Comparison of multi-step (24-192) prediction RMSE on BTH dataset.

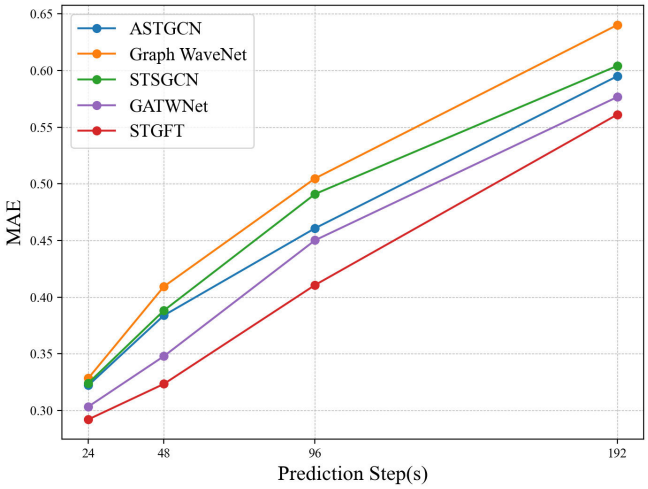



Fig. 12. Comparison of multi-step (24-192) prediction MAE on BTH dataset.

Moreover, we also increase the step size to validate the long-term prediction accuracy of STGFT. It is shown in Figs. 11 and 12 that with the increase of prediction steps, the prediction accuracy of all models decreases. This is because the models need to understand the relationship between more distant time points, leading to information loss and increased uncertainty. However, STGFT achieves the lowest RMSE and MAE on all prediction steps, proving the model's superiority.

In addition, this work adopts the heat map to show the original adjacency matrix, ADMG-generated adaptive adjacency matrix, and ADMG-generated dynamic adjacency matrix composed of 24 nodes in the BTH dataset to show the effectiveness of ADMG. It is shown in Fig. 13 that the adaptive adjacency matrix learns the main river network spatial relationships. In contrast, the dynamic adjacency matrix generated based on input features provides some potential spatial relationship as an auxiliary. Therefore, the predefined, adaptive, and dynamic adjacency matrixes complement each other in the spatial relationship. Finally, they are fused at the multi-graph fusion layer, providing a spatially dependent basis for aggregating spatiotemporal relationships.

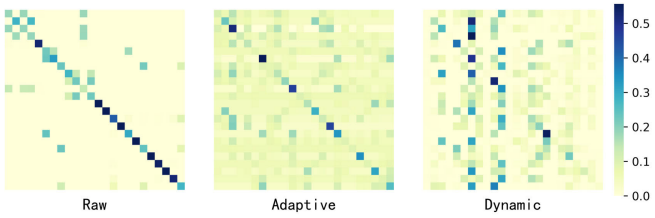



Fig. 13. Heat map of the adjacency matrices in the BTH dataset.

### C. K-Fold Cross-Validation

$K$-fold cross-validation is employed to avoid model overfitting and improve generalization ability. It divides the dataset into $K$ subsets of the same size and gradually uses different subsets as the validation set. This can ensure that each data point appears in the validation set and that the division of training and validation sets is more balanced. The $K$ is selected as 10 in the experiment. Taking the BTH dataset as an example, the dataset is randomly divided into 10 equal-sized subsets. Nine pieces of data are used as the training set and one as the validation set in each iteration, ensuring that each subset can appear once as the validation set. In that case, the model's generalization performance is evaluated at different prediction step sizes (1-10, 24, 48, 96, 192), which leads to more robust and reliable experimental results and effectively avoids the overfitting problem of the model. The result of the $K$-fold cross-validation of the single-step prediction of the BTH dataset is shown as an example in Table VII. The results demonstrate that the model's RMSE, MAE, and MAPE do not change significantly and perform well at different validation rounds, proving the model's robustness.

### D. Ablation Studies

The ablation experiment aims to analyze the effectiveness of each module in the STGFT. The vertical coordinate represents

TABLE VII

*K*-FOLD CROSS-VALIDATION FOR THE BTH DATASET

| Validation Rounds | RMSE | MAE | MAPE |
|---|---|---|---|
| 1 | 0.3072 | 0.2164 | 0.0755 |
| 2 | 0.3216 | 0.2025 | 0.0506 |
| 3 | 0.3288 | 0.2326 | 0.0790 |
| 4 | 0.2954 | 0.2011 | 0.0735 |
| 5 | 0.3123 | 0.2157 | 0.0760 |
| 6 | 0.3199 | 0.2108 | 0.0742 |
| 7 | 0.3085 | 0.2206 | 0.0782 |
| 8 | 0.3267 | 0.2293 | 0.0766 |
| 9 | 0.3151 | 0.2232 | 0.0777 |
| 10 | 0.3298 | 0.2185 | 0.0728 |
| **Average** | **0.3165** | **0.2171** | **0.0736** |

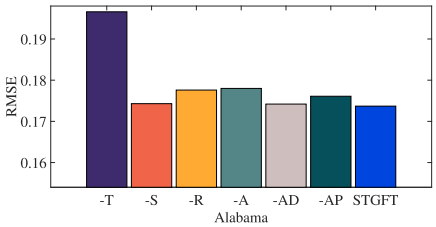



Fig. 14. Ablation on alabama dataset.

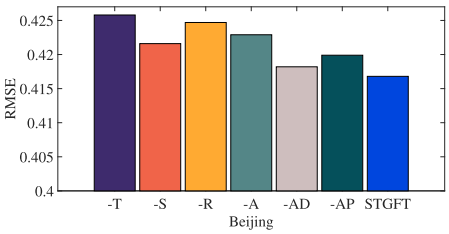



Fig. 15. Ablation on beijing dataset.

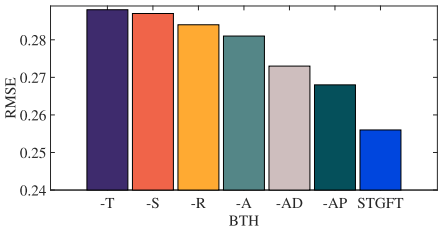



Fig. 16. Ablation on BTH dataset.

the RMSE, with smaller values indicating more accurate predictions. It compares the prediction accuracy of the STGFT after removing certain modules from it. Specifically, this subsection verifies the effectiveness of ADMG, TAE, and SAE.

Figs. 14-16 show the prediction accuracy after removing a certain part of STGFT when the input length is 40, and the prediction length is 10. *X* means removing $X \in \{\text{T, S, R, A, AD, AP}\}$, where T and S mean the benchmarks without TAE and SAE, respectively. R and A mean the benchmarks without removing the predefined adjacency matrix and ADMG, respectively. AD and AP mean benchmarks without removing dynamic and adaptive adjacency matrices generated by ADMG, respectively. It is shown in Figs. 14-16 that removing any module in STGFT negatively impacts its prediction accuracy. After removing ADMG or its generated adaptive adjacency matrix and dynamic adjacency matrix, STGFT's prediction accuracy on the BTH dataset compared with the other two datasets has a more obvious reduction. Therefore, the ablation experiments verify the effectiveness of each module in STGFT, proving that ADMG based on SAE plays a vital role in mining potential spatial features of the water environment. ADMG assists STGFT in capturing richer spatial dependencies.

## V. DISCUSSION

The computational complexity of each model is discussed. Table VIII shows the computational complexity among STGFT and other baseline models. It illustrates the number of parameters and FLOPs of each model. FLOPs measure how many floating-point operations are performed during the model training and inference. It determines the complexity of the model and affects the speed of the model training and inference. It is shown that STGFT has the least number of parameters but relatively high FLOPs. This is mainly due to the graph training. However, the adaptive and dynamic graphs significantly contribute to the accuracy of prediction results.

TABLE VIII

COMPUTATIONAL COMPLEXITY AMONG STGFT AND OTHER BASELINE MODELS

| Models | Number of parameters | FLOPs |
|---|---|---|
| ASTGCN | $3.18 \times 10^4$ | $1.80 \times 10^{10}$ |
| Graph WaveNet | $1.80 \times 10^{10}$ | $7.00 \times 10^4$ |
| STSGCN | $2.34 \times 10^{11}$ | $1.09 \times 10^8$ |
| GATWNet | $5.54 \times 10^{10}$ | $3.29 \times 10^6$ |
| STGFT | $2.80 \times 10^3$ | $1.13 \times 10^{10}$ |

Furthermore, STGFT combines spatiotemporal features for predicting water environment data, essentially a time serial prediction technique. Thus, it can be applied to other time series prediction tasks, especially ones with simultaneous spatial features. **First**, it can be used in traffic prediction [16], where traffic congestion varies periodically. In addition, spatial characteristics of the road network also exist because the traffic condition of a road is affected by its surrounding roads. Thus, our method is suitable for traffic prediction. **Second**, it can also be used for photovoltaic output prediction in grid-connected power plants [37]. This problem aims to predict the future photovoltaic sequence using historical photovoltaic sequence data, which is essentially a time series prediction problem. The problem also has spatial characteristics, *i.e.*, photovoltaic power generation from neighboring sites affects the power generation at that site, so spatial attention can automatically capture matching photovoltaic power generation sequences from neighboring sites. Therefore, our method is also applicable to this problem. **Third**, it can also be used for air quality prediction [15]. Air quality has strong spatial characteristics, and the air quality in one location can be affected by other areas. Therefore, reasonably incorporating spatial information can predict future air quality more accurately. Our method is suitable for other time series prediction tasks, especially when the problem has spatial characteristics.

## VI. CONCLUSION

With the continuous growth of human activities and rapid economic development, water environment problems become increasingly prominent. The usage of water quality prediction

techniques can help anticipate water quality problems and avoid further quality deterioration of water by taking some timely actions. However, water environment presents the characteristics of cross-regional and multi-site interactions. In that case, traditional water quality prediction methods ignore the spatial correlation of water quality changes, making it difficult to meet the demand for accurate water quality prediction. Moreover, they focus on predefined graph structures to reflect the spatial features that cannot capture potential spatial dependencies when dealing with complex water quality data. To solve the above problems, this work proposes a novel water quality prediction model named Spatial-Temporal Graph Fusion Transformer (STGFT). It incorporates a spatial and temporal attention encoder to capture the spatial correlations and temporal characteristics among different water quality monitoring stations. Moreover, an adaptive dynamic adjacency matrix generator is designed to generate adaptive and dynamic graphs to effectively mine potential spatial dependencies in a river network. Finally, the experimental results based on three real-world datasets show that STGFT can achieve higher accuracy in long-term water quality prediction than its state-of-the-art peers.

Our future work aims to further integrate meteorology [33] and geography [34] into our STGFT to enhance the robustness and reliability of the model. In addition, due to the high requirement for real-time water quality predictions, we intend to use intelligent optimization [35] and distributed computing [36] to accelerate the training and inference process of the model.

## REFERENCES


[1] L. Jia, N. Yen, and Y. Pei, "Spatial and temporal water quality data prediction of transboundary watershed using multiview neural network coupling," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 3000816.

[2] D. He, Y. Zhong, X. Wang, and L. Zhang, "Deep convolutional neural network framework for subpixel mapping," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 11, pp. 9518–9539, Nov. 2021.

[3] J. Liang, Z. Du, J. Liang, K. Yao, and F. Cao, "Long and short-range dependency graph structure learning framework on point cloud," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 12, pp. 14975–14989, Dec. 2023.

[4] J. Yin, J. Wei, Q. Li, and O. O. Ayantobo, "Regional characteristics and impact factors of change in terrestrial water storage in Northwestern China from 2002 to 2020," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 386–398, 2023.

[5] O. Francis, T. Michael, and A. Charles, "Multiple linear regression (MLR) model: A tool for water quality interpretation," *Momona Ethiopian J. Sci.*, vol. 12, no. 1, pp. 123–134, Apr. 2020.

[6] J. Wang, L. Zhang, W. Zhang, and X. Wang, "Reliable model of reservoir water quality prediction based on improved ARIMA method," *Environ. Eng. Sci.*, vol. 36, no. 9, pp. 1041–1048, Sep. 2019.

[7] E. Batur and D. Maktav, "Assessment of surface water quality by using satellite images fusion based on PCA method in the Lake Gala, Turkey," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 5, pp. 2983–2989, May 2019.

[8] X. Su, X. He, G. Zhang, Y. Chen, and K. Li, "Research on SVR water quality prediction model based on improved sparrow search algorithm," *Comput. Intell. Neurosci.*, vol. 2022, pp. 1–23, Apr. 2022.

[9] H. Lu and X. Ma, "Hybrid decision tree-based machine learning models for short-term water quality prediction," *Chemosphere*, vol. 249, Jun. 2020, Art. no. 126169.

[10] Y. Zhan, H. Zhang, and Y. Liu, "Forecast of meteorological and hydrological features based on SVR model," in *Proc. 4th Int. Conf. Adv. Electron. Mater., Comput. Softw. Eng. (AEMCSE)*, Changsha, China, Mar. 2021, pp. 579–583.

[11] T. Jahani-Nezhad and M. A. Maddah-Ali, "Berrut approximated coded computing: Straggler resistance beyond polynomial computing," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 111–122, Jan. 2023.

[12] B. Jiang et al., "Dynamic temporal dependency model for multiple steps ahead short-term load forecasting of power system," *IEEE Trans. Ind. Appl.*, vol. 60, no. 4, pp. 5244–5254, Jul. 2024.

[13] L. Xu, R. Cai, H. Yu, W. Du, Z. Chen, and N. Chen, "Monthly NDVI prediction using spatial autocorrelation and nonlocal attention networks," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 3425–3437, 2024.

[14] J. Qiao, Y. Lin, J. Bi, H. Yuan, G. Wang, and M. Zhou, "Attention-based spatiotemporal graph fusion convolution networks for water quality prediction," *IEEE Trans. Autom. Sci. Eng.*, vol. 22, pp. 1–10, 2025, doi: 10.1109/TASE.2023.3285253.

[15] Q. Chen, R. Ding, X. Mo, H. Li, L. Xie, and J. Yang, "An adaptive adjacency matrix-based graph convolutional recurrent network for air quality prediction," *Sci. Rep.*, vol. 14, no. 1, pp. 1–17, Feb. 2024.

[16] Y. Shin and Y. Yoon, "PGCN: Progressive graph convolutional networks for spatial–temporal traffic forecasting," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 7, pp. 7633–7644, Jul. 2024.

[17] X. Yang, Q. Zhao, Y. Wang, and K. Cheng, "Multistate prediction for in-service gas turbine via adaptive diffusion graph network," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–12, 2023.

[18] K. Wu, J. Fan, P. Ye, and M. Zhu, "Hyperspectral image classification using spectral–spatial token enhanced transformer with hash-based positional embedding," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5507016.

[19] R. Chen, D. Cai, X. Hu, Z. Zhan, and S. Wang, "Defect detection method of aluminum profile surface using deep self-attention mechanism under hybrid noise conditions," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–9, 2021.

[20] H. Chen, D. Jiang, and H. Sahli, "Transformer encoder with multi-modal multi-head attention for continuous affect recognition," *IEEE Trans. Multimedia*, vol. 23, pp. 4171–4183, 2021.

[21] J. P. Sahoo, S. P. Sahoo, S. Ari, and S. K. Patra, "Hand gesture recognition using densely connected deep residual network and channel attention module for mobile robot control," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–11, 2023.

[22] N. Passalis, A. Tefas, J. Kanniainen, M. Gabbouj, and A. Iosifidis, "Deep adaptive input normalization for time series forecasting," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 9, pp. 3760–3765, Sep. 2020.

[23] M. Mesgaran and A. B. Hamza, "Anisotropic graph convolutional network for semi-supervised learning," *IEEE Trans. Multimedia*, vol. 23, pp. 3931–3942, 2021.

[24] H. Li and L. Zhang, "A bilevel learning model and algorithm for self-organizing feed-forward neural networks for pattern classification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 11, pp. 4901–4915, Nov. 2021.

[25] L. Mo, X. Lu, J. Yuan, C. Zhang, Z. Wang, and P. Popovski, "Generalized unitary approximate message passing for double linear transformation model," *IEEE Trans. Signal Process.*, vol. 71, pp. 1524–1538, 2023.

[26] W. Zhang, Z. Zhu, and Y. Geng, "Simultaneous conductivity and permeability reconstructions for electromagnetic tomography using deep learning," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–11, 2023.

[27] Y. Zhang, L. Jiang, and H. T. Ewe, "A novel data-driven modeling method for the spatial–temporal correlated complex sea clutter," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5104211.

[28] Y. Liu, Q. Zhang, and Z. Lv, "Real-time intelligent automatic transportation safety based on big data management," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 7, pp. 9702–9711, Jul. 2022.

[29] X. Wan, Y. Peng, R. Hao, and Y. Guo, "Capturing spatial–temporal correlations with attention based graph convolutional networks for network traffic prediction," in *Proc. 15th Int. Conf. Commun. Softw. Netw. (ICCSN)*, Shenyang, China, Jul. 2023, pp. 95–99.

[30] N. Rathore, P. Rathore, A. Basak, S. H. Nistala, and V. Runkana, "Multi scale graph wavenet for wind speed forecasting," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Orlando, FL, USA, Dec. 2021, pp. 4047–4053.

[31] D. Zhao, Q. Yang, X. Zhou, H. Li, and S. Yan, "A local spatial–temporal synchronous network to dynamic gesture recognition," *IEEE Trans. Computat. Social Syst.*, vol. 10, no. 5, pp. 2226–2233, Oct. 2023.

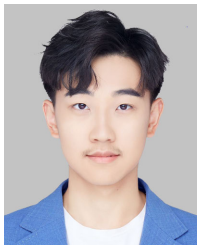
[32] S. Liu, J. Zhu, W. Lei, and P. Zhang, "Spatial–temporal attention graph WaveNet for traffic forecasting," in *Proc. 5th Int. Conf. Data-driven Optim. Complex Syst. (DOCS)*, Tianjin, China, Sep. 2023, pp. 1–8.

[33] J. Seo, J. Won, H. Lee, and S. Kim, "Probabilistic monitoring of meteorological drought impacts on water quality of major rivers in South Korea using copula models," *Water Res.*, vol. 251, Mar. 2024, Art. no. 121175.
[34] Y. Liu, Y. Yao, and Q. Zhao, "Real-time rainfall nowcast model by combining CAPE and GNSS observations," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4109909.
[35] J. Bi, Z. Wang, H. Yuan, J. Zhang, and M. Zhou, "Self-adaptive teaching-learning-based optimizer with improved RBF and sparse autoencoder for high-dimensional problems," *Inf. Sci.*, vol. 630, pp. 463–481, Jun. 2023.
[36] J. Bi, Z. Wang, H. Yuan, J. Zhang, and M. Zhou, "Cost-minimized computation offloading and user association in hybrid cloud and edge computing," *IEEE Internet Things J.*, vol. 11, no. 9, pp. 16672–16683, May 2024.
[37] J. Yang et al., "Day-ahead PV power forecasting model based on fine-grained temporal attention and cloud-coverage spatial attention," *IEEE Trans. Sustain. Energy*, vol. 15, no. 2, pp. 1062–1073, Apr. 2024.

**Jing Bi** (Senior Member, IEEE) received the B.S. and Ph.D. degrees in computer science from Northeastern University, Shenyang, China, in 2003 and 2011, respectively. She is currently a Professor with the College of Computer Science, Beijing University of Technology, Beijing, China. She has over 170 publications in international journals and conference proceedings. Her research interests include distributed computing, cloud and edge computing, large-scale data analytics, machine learning, industrial internet, and performance optimization. She is an Associate Editor of IEEE TRANSACTIONS ON SYSTEMS MAN AND CYBERNETICS: SYSTEMS.

**Ziqi Wang** (Student Member, IEEE) received the B.E. degree in Internet of Things from Beijing University of Technology, Beijing, China, in 2022, where he is currently pursuing the master's degree with the College of Computer Science. His research interests include edge-cloud computing, task scheduling, intelligent optimization algorithms, deep learning, and machine learning. He received the Best Paper Award from the 2024 Joint International Conference on Automation-Intelligence-Safety and the International Symposium on Autonomous Systems.

**Haitao Yuan** (Senior Member, IEEE) received the Ph.D. degree in computer engineering from New Jersey Institute of Technology (NJIT), Newark, NJ, USA, in 2020. He is currently the Deputy Director with the Department of Science and Technology Innovation, Wenchang International Aerospace City, Hainan, China. He is currently an Associate Professor at the School of Automation Science and Electrical Engineering, Beihang University, Beijing, China. His research interests include the Internet of Things, edge computing, deep learning, data-driven optimization, and computational intelligence algorithms. He received the Chinese Government Award for Outstanding Self-Financed Students Abroad, the 2021 Hashimoto Prize from NJIT, the Best Paper Award in the 17th ICNSC, and the Best Student Paper Award Nominees in 2024 IEEE SMC. He is an Associate Editor of IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS: SYSTEMS, IEEE INTERNET OF THINGS JOURNAL, and *Expert Systems with Applications*. He is named in the world's top 2% of Scientists List.



**Xiangxi Wu** received the B.E. degree in software engineering from Beijing University of Technology, Beijing, China, in 2022, where he is currently pursuing the master's degree with the College of Computer Science. His research interests include big data analysis and processing, deep learning, machine learning, and data mining.

**Renren Wu** received the Ph.D. degree in environmental engineering from South China University of Technology in 2011. He is currently a Professor with South China Institute of Environmental Sciences (SCIES), Ministry of Ecology and Environment of the People's Republic of China. He is also the Deputy Director of the State Environmental Protection Key Laboratory of Water Environmental Simulation and Pollution Control. He has over 30 publications in international journals. His research interests include water pollution source tracking, pollution control in coastal waters, and river basins.

**Jia Zhang** (Senior Member, IEEE) received the Ph.D. degree in computer science from the University of Illinois at Chicago. She is currently the Cruse C. and Marjorie F. Calahan Centennial Chair in Engineering and a Professor with the Department of Computer Science, Lyle School of Engineering, Southern Methodist University. Her research interests emphasize the application of machine learning and information retrieval methods to tackle data science infrastructure problems, with a recent focus on scientific workflows, provenance mining, software discovery, knowledge graph, and interdisciplinary applications of all of these interests in earth science.

**MengChu Zhou** (Fellow, IEEE) received the Ph.D. degree from the Rensselaer Polytechnic Institute, Troy, NY, USA, in 1990. Then, he joined New Jersey Institute of Technology, Newark, NJ, USA, where he is currently a Distinguished Professor. He has over 900 publications, including 12 books, more than 600 journal articles (more than 450 in IEEE TRANSACTIONS), 28 patents, and 29 book-chapters. His research interests include Petri nets, automation, the Internet of Things, and big data. He is a fellow of IFAC, AAAS, CAA, and NAI.