5th International Conference on Industrial Artificial Intelligence (IAI), Aug. 21-24, 2023, Shenyang, China

Spatial-Temporal Graph Fusion Transformer for Long-term Water Quality Prediction

Ziqi Wang
Faculty of Information Technology
Beijing University of Technology
Beijing, China
ziqi_wang@emails.bjut.edu.cn

Xiangxi Wu
Faculty of Information Technology
Beijing University of Technology
Beijing, China
Wuxiangxi7@emails.bjut.edu.cn

Xingyang Chang
Faculty of Information Technology
Beijing University of Technology
Beijing, China
David205x@emails.bjut.edu.cn

Renren Wu
South China Institute of Environmental Sciences
Ministry of Ecology and Environment
Guangzhou, China
wurenren@scies.org

Jing Bi
Faculty of Information Technology
Beijing University of Technology
Beijing, China
bijing@bjut.edu.cn

Junfei Qiao
Faculty of Information Technology
Beijing University of Technology
Beijing, China
junfeiq@bjut.edu.cn

2024 6th International Conference on Industrial Artificial Intelligence (IAI) | 979-8-3503-5661-8/24/$31.00 ©2024 IEEE | DOI: 10.1109/IAI63275.2024.10730231

***Abstract*—Over the past decades of rapid development, the global water pollution problem has become prominent. Accurate water quality prediction can detect the trend and anomaly of water quality changes in advance, thereby taking timely measures to avoid the occurrence of water quality problems. Traditional statistical methods for water quality prediction make it difficult to capture the complex relationship between multiple variables and deep learning models make it difficult to capture temporal dependence and spatial correlation of the water quality simultaneously. 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 (SAE), a temporal attention encoder (TAE), an adaptive dynamic adjacency matrix generator (ADMG), and a multi-graph fusion layer. Among them, SAE and TAE are adopted to capture the spatial correlations and temporal characteristics among different water quality monitoring stations, respectively. ADMG generates adaptive and dynamic adjacency matrices to reflect potential spatial relationships in the river network. Experimental results with real-life water quality datasets demonstrate that STGFT outperforms current state-of-the-art models regarding prediction accuracy.**

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

I. INTRODUCTION

Nowadays, the deterioration of the water environment has become one of the most important factors constraining the sustainable development of society. To solve this problem, water quality prediction methods [1] are designed to forecast elemental values of the water environment in the future based on past monitoring data. In this case, 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 [2]. Moreover, these methods are often based on specific assumptions, *e.g.*, water quality trends are linear and time is steady. However, these assumptions may not be consistent with the actual situation, which biases the prediction results.

Moreover, deep learning models, *e.g.*, Back Propagation Neural Networks [3], Recurrent Neural Networks [4], and Convolutional Neural Networks [5] are suitable for water quality prediction through the limited water quality information. However, with the strengthening of socio-economic ties between regions, the water environment is gradually showing complex changes across regions. Moreover, multiple water quality monitoring stations interact with each other, and the data from them are not only affected by the historical values but also by the values from the upstream monitoring stations, which increases the complexity of the water quality prediction.

To solve the above problem, Graph Neural Networks (GNN) [6] have shown powerful capabilities in dealing with complex spatial sequence data [7]. Specifically, it can handle non-Euclidean [8] data and represent water quality data in spatial dimensions. Therefore, it can model the spatial relationship between the location of each water quality monitoring station as a graph structure [9]. However, due to the high complexity of river networks and the uncertainty of spatial relationships, inaccurate information used in defining the graph structure results in an inaccurate graph. Therefore, a predefined graph structure can only capture the local spatial information, and it is difficult to adequately capture the spatial dependencies, which affects the accuracy of the water quality prediction.

Based on the aforementioned analysis, this paper 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. Moreover, an adaptive dynamic adjacency matrix gen-

This work was supported by Beijing Natural Science Foundation (L233005), National Natural Science Foundation of China (NSFC) under Grants 62173013 and 62073005.

erator (ADMG) is designed to utilize water quality spatial and temporal characteristics to generate adaptive and dynamic graphs to better reflect potential spatial relationships in the river network, which allows STGFT not to be restricted by the predefined graph structure. Experimental results based on three real-world datasets show that the STGFT has high accuracy in long-term water quality predictions.

II. Proposed Methodology

A. 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 [10]. 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 the water quality, which in turn affects the subsequent changes in it. In this case, to capture the correlation of water quality elements between different time steps, this paper designs a TAE that learns the temporal features of each water quality monitoring station. The structure of the TAE is shown in Fig. 1. It is assumed that there are N water quality monitoring stations and C water quality elements. It incorporates multiple temporal attention layers and they are stacked together. 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 $X=\{X_{:,1}, X_{:,2}, X_{:,t}, ..., X_{:,T}\}\in\mathbb{R}^{N\times T\times C}$ of water quality monitoring stations, where $\mathbb{R}$ denotes a set of real numbers, T denotes the time steps and D denotes the embedding dimension. Then, according to (1) and (2), the positional embedding (PE) [11] is obtained, where pos denotes the position number and i denotes the current dimension number. Moreover, $PE_{(pos,2i)}$ and $PE_{(pos,2i+1)}$ occur alternately and they are obtained by the sin function ($sin(\cdot)$) and the cos function ($cos(\cdot)$), respectively.

$$PE_{(pos,2i)}=\sin(pos/10000^{2i/D}) \tag{1}$$

$$PE_{(pos,2i+1)}=\cos(pos/10000^{2i/D}) \tag{2}$$

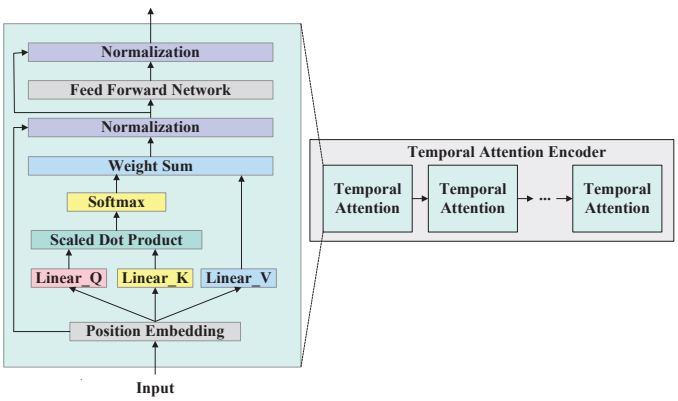

Fig. 1. Structure of the TAE.

Then, the input feature embedding is added to the positional embedding, obtaining the input of the temporal attention layer ($\hat{X}_T=(X'+PE)\in\mathbb{R}^{N\times T\times D}$). In the temporal attention layer, a self-attention mechanism [12] is adopted to extract the internal correlation of the historical sequence data for N sites in parallel. First, $\hat{X}_T$ is mapped to three different feature spaces, obtaining the query vector $Q_T\in\mathbb{R}^{N\times T\times D}$, key vector $K_T\in\mathbb{R}^{N\times T\times D}$, and the value vector $V_T\in\mathbb{R}^{N\times T\times D}$. 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, V_T, and the $\mathrm{Softmax}(\cdot)$ 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 ($Attention(\cdot)$). The specific calculation process is as follows:

$$Q_T=\hat{X}_T W_T^Q \tag{3}$$

$$K_T=\hat{X}_T W_T^K \tag{4}$$

$$V_T=\hat{X}_T W_T^V \tag{5}$$

$$Attention\left(Q_T, K_T, V_T\right)=\mathrm{Softmax}\left(\frac{Q_T K_T^T}{\sqrt{d_k}}\right)V_T \tag{6}$$

where W_T^Q, W_T^K, W_T^V represent trainable parameter matrices, d_k represents a scaling factor, and it is the size of the first dimension of the K_T. Moreover, to capture the complex features of the water quality, a multi-head attention mechanism [13] is adopted in the temporal attention layer, *i.e.*, training k groups of self-attention mechanisms while later concatenating the results and then remapping them back to the original dimensions. The specific calculation process is as follows:

$$\mathrm{head}_T(i)=Attention\left(W_T^Q(i)\hat{X}_T, W_T^K(i)\hat{X}_T, W_T^V(i)\hat{X}_T\right) \tag{7}$$

$$\mathrm{MultiHead}\left(\hat{X}_T\right)=\|_{i=1}^{k}\left(\mathrm{head}_T(i)\right)W_T^O \tag{8}$$

where $W_T^Q(i), W_T^K(i), W_T^V(i)$ 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 [14], the output of the multi-head attention mechanism ($\mathrm{MultiHead}(\hat{X}_T)$) is added to $\hat{X}$. Then it passes layer normalization [15] and a feed-forward neural network. Finally, after normalization, the output result of the TAE (O_{TAE}) is obtained, *i.e.*,

$$Z=\begin{cases}\hat{X}, i=0\\ O_{TAE}^{(i-1)}, \text{ otherwise}\end{cases} \tag{9}$$

$$\text{residual }^{(i)}=\mathrm{NL}(\mathrm{MultiHead}(Z)+Z) \tag{10}$$

$$O_{TAE}^{(i)}=\mathrm{NL}\left(W_{T_1}^{(i)}\,\mathrm{ReLU}\left(W_{T_0}^{(i)}\text{residual}^{(i)}\right)+\text{residual}^{(i)}\right) \tag{11}$$

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

B. ADMG based on SAE

1) SAE: 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. To effectively mine the potential spatial features of the water quality data, an SAE is proposed to capture the correlation between each water quality monitoring station. The structure of an SAE is shown in Fig. 2. Specifically, the predefined adjacency matrix A and X' are used as the input of the graph convolutional networks (GCN) [16] layer, resulting in a node embedding vector $\hat{X}_S$. Next, similar to the temporal attention layer, parameter matrices W_S^Q, W_S^K, 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. 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 [17], obtaining the spatial attention $Attention_S$, *i.e.*,

$$\hat{X}_S = GCN\left(X', A\right) \tag{12}$$

$$\mathrm{head}_S(i) = \mathrm{Attention}\left(W_S^Q(i)\hat{X}_S, W_S^K(i)\hat{X}_S, W_S^V(i)\hat{X}_S\right) \tag{13}$$

$$\mathrm{MultiHead}\left(\hat{X}_S\right) = \|_{i=1}^{k}\left(\mathrm{head}_S(i)\right) W_S^O \tag{14}$$

$$Attention_S = W_{S_1}\,\mathrm{ReLU}\left(W_{S_0}\left(\mathrm{MultiHead}\left(\hat{X}_S\right)\right)\right) \tag{15}$$

where $W_S^Q(i), W_S^K(i), W_S^V(i)$ represent trainable parameter matrices in the group i of self-attention mechanisms. 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. The specific calculation is shown in (16). Specifically, this paper defines the output of the last stacked spatial attention layer as O_{SAE}.

$$O_{SAE} = GCN(Attention_S, A) \tag{16}$$

2) ADMG: Due to the high complexity and uncertainty of spatial relationships in river networks, the predefined graph structure cannot reflect the real spatial relationships. Therefore, an ADMG is designed to generate adaptive and dynamic adjacency matrices to mine the potential spatial dependencies in river networks. As shown in Fig. 3, the ADMG first uses a randomly initialized vector $E \in \mathbb{R}^{N\times D}$ to construct a adaptive adjacency matrix $A_p \in \mathbb{R}^{N\times N}$, By constructing it, the deficiency of the A in representing node relationships is compensated. Moreover, the A_p is fixed after training. Furthermore, to construct a dynamic adjacency matrix, the

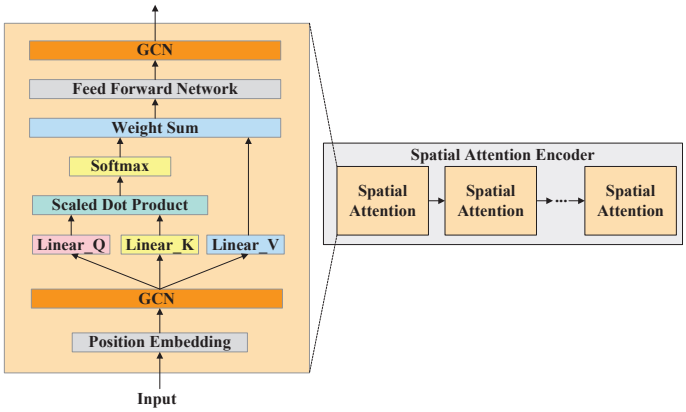

Fig. 2. Structure of the SAE.

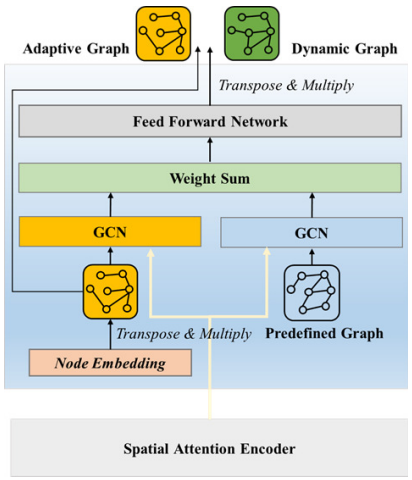

Fig. 3. Structure of the ADMG.

O_{SAE} is input into two parallel GCNs. They take the A_p and the A as parameters, obtaining the dynamic feature mapping $F_d \in \mathbb{R}^{N\times T\times D}$. This process is shown in (17) and (18), where α and β are trainable parameters and they are used to weight the output results of the two GCNs.

$$A_p = \mathrm{Softmax}(\mathrm{ReLU}(E \cdot E^T)) \tag{17}$$

$$F_d = \alpha GCN(O_{SAE}, A_p) + \beta GCN(O_{SAE}, A) \tag{18}$$

Then, the F_d is converted into a two-dimensional matrix ($F'_d \in \mathbb{R}^{(D\times T)\times N}$), and a linear transformation [18] 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. This process is shown in (19) and (20).

$$\tilde{F}_d = W_f F'_d \tag{19}$$

$$E_d = \mathrm{ReLU}(\mathrm{Tanh}(\tilde{F}_d \odot E)) \tag{20}$$

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, as shown in

(21), E_d is multiplied by its transpose matrix E_d^T to generate a dynamic adjacency matrix $A_d \in \mathbb{R}^{N \times N}$.

$$A_d = \mathrm{ReLU}(\mathrm{Tanh}(E_d \cdot E_d^T)) \tag{21}$$

C. STGFT

Sections II-A and II-B describe the three main components of the STGFT, *i.e.*, TAE, SAE, and ADMG. This section introduces the overall architecture of the STGFT.

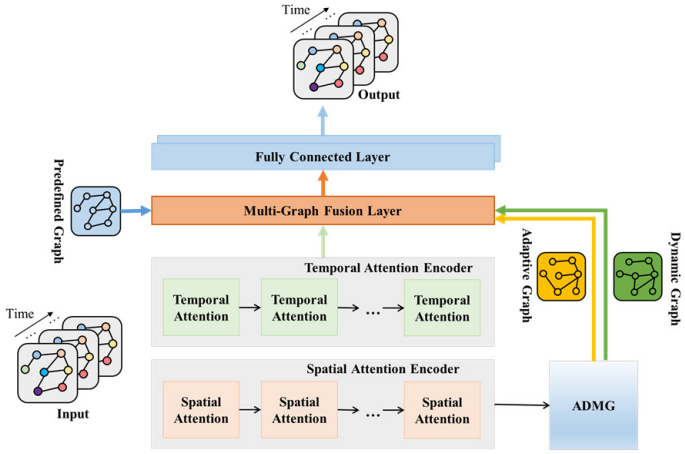

Fig. 4. Overall framework of the STGFT.

Fig. 4 shows the architecture of the STGFT. The original water quality sequence data X is input in parallel to TAE and SAE, obtaining the temporal features O_{TAE} that contains the correlation between different time steps and the spatial features O_{SAE} that contains the correlation between different nodes. Then O_{SAE} is used as the input of ADMG and then obtains the adaptive adjacency matrix A_p and the dynamic adjacency matrix A_d. It is worth noting that A_d contains the potential spatial features between nodes. Moreover, the multi-graph fusion layer adopts three parallel GCN to fuse the A_p, A_d, and A to generate node embeddings F_{latent}. The specific process is shown in (22), where μ, ν, ω are parameters that are adopted to weight the output results of the three GCN.

$$F_{latent} = \mu(GCN(A, X)) + \nu(GCN(A_p, X)) + \omega(GCN(A_d, X)) \tag{22}$$

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

$$Y' = \mathrm{FW}_t(F_{\text{latent}}) = W_t^1 \,\mathrm{ReLU}\left(W_t^0 F_{\text{latent}}\right) \tag{23}$$

$$Y = \mathrm{FW}_d(Y') = \mathrm{ReLU}\left(Y' W_d^0\right) W_d^1 \tag{24}$$

where FW_t and FW_d represent two feed forward networks, FW_t is used to transform the time dimension, converting F_{latent} into a vector Y' of the target prediction length, 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.

III. 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. Compared with the Alabama and Beijing datasets, the BTH dataset contains more complex spatial relationships. It includes 24 water quality monitoring stations in different administrative divisions of the Beijing-Tianjin-Hebei region in China. It is worth noting that this paper 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 and these hyperparameters include the number of heads of the multi-head attention mechanism (H), embedding dimension (E), 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 at the same time, thereby enhancing the model's generalization and representation abilities. An appropriate H can help to improve the overall predictive performance of the model. This paper sets H within [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, it is necessary to adjust the E during the training process. This paper sets E within [8,16,32]. Finally, the G is selected within [8,16,32,64]. Table I shows the root mean square error (RMSE), mean absolute error (MAE), and mean absolute percentage error (MAPE) 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

To verify the effectiveness of the STGFT, four baseline models are adopted for comparative experiments, *i.e.*, Attention Based Spatial-Temporal Graph Convolutional Networks (ASTGCN) [19], Graph WaveNet [20], Spatial-Temporal Synchronous Graph Convolutional Networks (STSGCN) [21], Graph Attention WaveNet (GATWNet) [22]. Figs. 5 and 6 show the RMSE and MAE of STGFT and comparative models on prediction steps from 1 to 10. Table II shows the prediction error of STGFT and comparative models on Alabama, Beijing, and BTH datasets. It is shown in Figs. 5 and 6 that STGFT

TABLE I
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)	0.0603	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

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 II that STGFT achieves the lowest prediction error on almost 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 models, indicating that STGFT has higher accuracy and stability on water quality predictions. Furthermore, compared with experimental results on Alabama and Beijing datasets that are on a smaller spatial scale, STGFT has a greater improvement in prediction accuracy on the BTH dataset. This shows that as spatial scale increases, STGFT can effectively capture time features and potential spatial features in spatiotemporal water quality data. Fig. 7 shows the prediction effect of 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 STGFT is closer to the true value, proving that STGFT has advantages in water quality spatial-temporal prediction.

In addition, this paper 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. 8 that the adaptive adjacency matrix learns the main river network spatial relationships, while the dynamic adjacency matrix generated based on input features provides some potential spatial relationship as an auxiliary. Therefore, the original adjacency matrix, adaptive adjacency matrix, and dynamic adjacency matrix 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.

IV. CONCLUSIONS

With the continuous growth of human activities and rapid economic development, water environment problems becoming increasingly prominent. The usage of water quality prediction techniques can help anticipate water quality problems and take timely action to avoid further deterioration. However, the water environment presents the characteristics of cross-

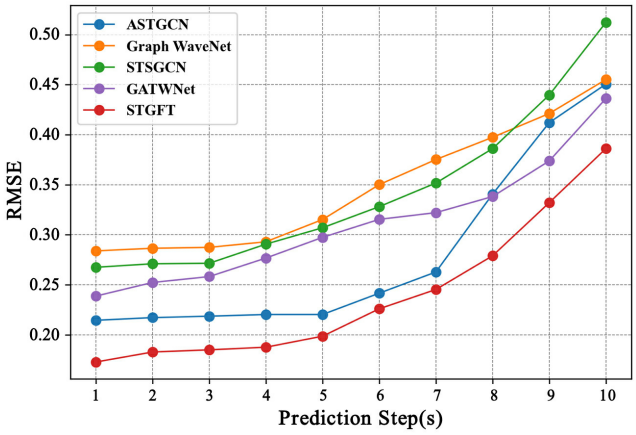

Fig. 5. Comparison of multi-step prediction RMSE on the BTH dataset.

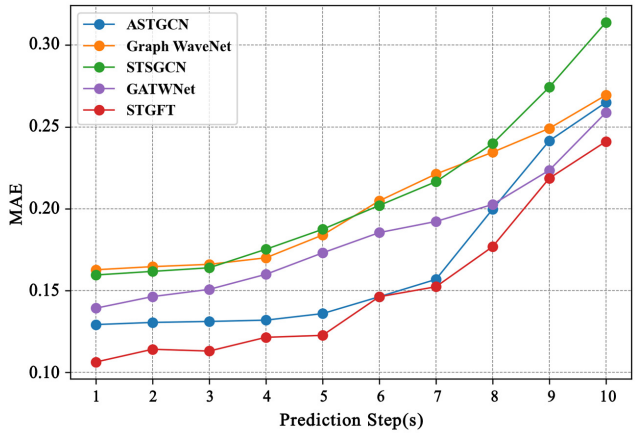

Fig. 6. Comparison of multi-step prediction MAE on the BTH dataset.

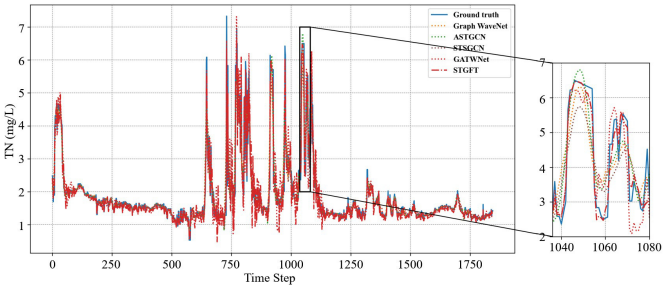

Fig. 7. Comparison of prediction results (Beiyang Bridge).

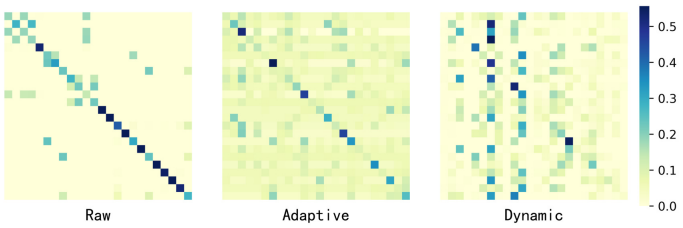

Fig. 8. Heat map of the adjacency matrices of the 24 nodes in the BTH dataset.

TABLE II
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**	**0.1160**	**0.0152**	**0.4169**	**0.2526**	0.0713	**0.2562**	**0.1512**	**0.0435**

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 prediction of water quality. 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 paper proposes a novel water quality prediction model named Spatial-Temporal Graph Fusion Transformer (STGFT). It incorporates a spatial attention encoder and a 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 mine potential spatial dependencies in the river network. Finally, the experimental results based on three real-world datasets show that STGFT achieves higher accuracy in long-term water quality prediction compared to its peers.

In our future work, we will further integrate meteorology [23] and geography [24] into our STGFT to enhance the robustness and reliability of the model. In addition, we intend to use intelligent optimization and distributed computing 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 Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–16, Nov. 2023.

[2] T. X. Bach, N. D. Anh, N. Van Linh and K. Than "Dynamic Transformation of Prior Knowledge Into Bayesian Models for Data Streams," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 4, pp. 3742–3750, Apr. 2023.

[3] J. Xu and Z. Liu, "A Back Propagation Neural Network-Based Algorithm for Retrieving All-Weather Precipitable Water Vapor From MODIS NIR Measurements," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–14, Nov. 2022.

[4] N. Mohajerin and S. L. Waslander, "Multistep Prediction of Dynamic Systems With Recurrent Neural Networks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 11, pp. 3370-3383, Nov. 2019.

[5] D. He, Y. Zhong, X. Wang and L. Zhang, "Deep Convolutional Neural Network Framework for Subpixel Mapping," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, pp. 9518–9539, Nov. 2021.

[6] G. Ciano, A. Rossi, M. Bianchini and F. Scarselli, "On Inductive–Transductive Learning With Graph Neural Networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 2, pp. 758–769, Feb. 2022.

[7] Y. Guo, S. Guo, Z. Jin, S. Kaul, D. Gotz and N. Cao, "Survey on Visual Analysis of Event Sequence Data," *IEEE Transactions on Visualization and Computer Graphics*, vol. 28, no. 12, pp. 5091–5112, Dec. 2022.

[8] L. Chu and H. Chen, "Sequential Change-Point Detection for High-Dimensional and Non-Euclidean Data," *IEEE Transactions on Signal Processing*, vol. 70, pp. 4498-4511, Sept. 2022.

[9] J. Liang, Z. Du, J. Liang, K. Yao and F. Cao, "Long and Short-Range Dependency Graph Structure Learning Framework on Point Cloud," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 12, pp. 14975-14989, Dec. 2023.

[10] J. Zheng, Z. Zhao, Y. Zeng, B. Shi and Z. Yu, "An Event-Driven Real-Time Simulation for Power Electronics Systems Based on Discrete Hybrid Time-Step Algorithm," *IEEE Transactions on Industrial Electronics*, vol. 70, no. 5, pp. 4809–4819, May. 2023.

[11] K. Wu, J. Fan, P. Ye and M. Zhu, "Hyperspectral Image Classification Using Spectral–Spatial Token Enhanced Transformer With Hash-Based Positional Embedding," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–16, Mar. 2023.

[12] 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 Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–9, Sept. 2021.

[13] H. Chen, D. Jiang and H. Sahli, "Transformer Encoder With Multi-Modal Multi-Head Attention for Continuous Affect Recognition," *IEEE Transactions on Multimedia*, vol. 23, pp. 4171-4183, Nov. 2021.

[14] 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 Transactions on Instrumentation and Measurement*, vol. 72, pp. 1-11, Feb. 2023.

[15] N. Passalis, A. Tefas, J. Kanniainen, M. Gabbouj and A. Iosifidis, "Deep Adaptive Input Normalization for Time Series Forecasting," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 9, pp. 3760-3765, Sept. 2020.

[16] M. Mesgaran and A. B. Hamza, "Anisotropic Graph Convolutional Network for Semi-Supervised Learning," *IEEE Transactions on Multimedia*, vol. 23, pp. 3931–3942, Oct. 2021.

[17] H. Li and L. Zhang, "A Bilevel Learning Model and Algorithm for Self-Organizing Feed-Forward Neural Networks for Pattern Classification," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 11, pp. 4901-4915, Nov. 2021.

[18] L. Mo, X. Lu, J. Yuan, C. Zhang, Z. Wang and P. Popovski, "Generalized Unitary Approximate Message Passing for Double Linear Transformation Model," *IEEE Transactions on Signal Processing*, vol. 71, pp. 1524-1538, Apr. 2023.

[19] X. Wan, Y. Peng, R. Hao and Y. Guo, "Capturing Spatial-Temporal Correlations with Attention Based Graph Convolutional Networks for Network Traffic Prediction," *2023 15th International Conference on Communication Software and Networks (ICCSN)*, Shenyang, China, 2023, pp. 95–99.

[20] N. Rathore, P. Rathore, A. Basak, S. H. Nistala and V. Runkana, "Multi Scale Graph Wavenet for Wind Speed Forecasting," *2021 IEEE International Conference on Big Data (Big Data)*, Orlando, FL, USA, 2021, pp. 4047–4053.

[21] D. Zhao, Q. Yang, X. Zhou, H. Li and S. Yan, "A Local Spatial–Temporal Synchronous Network to Dynamic Gesture Recognition," *IEEE Transactions on Computational Social Systems*, vol. 10, no. 5, pp. 2226–2233, Oct. 2023.

[22] S. Liu, J. Zhu, W. Lei and P. Zhang, "Spatial-Temporal Attention Graph WaveNet for Traffic Forecasting," *2023 5th International Conference on Data-driven Optimization of Complex Systems (DOCS)*, Tianjin, China, pp. 1–8, Oct. 2023.

[23] 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 Research*, vol. 251, Mar. 2024.

[24] Y. Liu, Y. Yao and Q. Zhao, "Real-Time Rainfall Nowcast Model by Combining CAPE and GNSS Observations," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–9, Sept. 2022.