STMF: A Spatiotemporal Multimodal Fusion Model for Long-Term Water Quality Forecasting

Jing Bi, *Senior Member, IEEE*, Xiangxi Wu, *Student Member, IEEE*, Haitao Yuan, *Senior Member, IEEE*, Ziqi Wang, *Student Member, IEEE*, Daming Wei, Renren Wu, Jia Zhang, *Senior Member, IEEE*, Junfei Qiao, *Senior Member, IEEE*, and Rajkumar Buyya, *Fellow, IEEE*

***Abstract*—Water quality forecasting is a time-series analysis task involving estimating future water conditions, vital in environmental management and pollution control. However, existing time-series analysis methods focus only on historical observational data, neglecting information from other modalities, leading to incomplete feature extraction and affecting forecasting accuracy and robustness. In addition, the complex spatial dependencies between water quality monitoring stations and the nonlinear fluctuations in water quality indicators caused by meteorological factors present additional challenges. This work proposes a spatiotemporal multimodal fusion architecture for long-term water quality forecasting, named STMF, to address these issues. It first captures spatiotemporal dependencies by integrating temporal features with upstream–downstream relationships among monitoring stations. Then, STMF further designs a low-rank cross-modal interaction fusion (LRCIF) method, which fuses spatiotemporal features with precipitation features from the remote-sensing image, as an additional modality, effectively leveraging complementary information from multiple data sources to enhance the accuracy and stability of water quality forecasting. Experimental results on real-world water quality datasets demonstrate that the proposed STMF significantly outperforms existing state-of-the-art methods in prediction accuracy. In particular, for long-term forecasting tasks with a 192-step horizon, STMF improves mean-squared error and mean absolute error by 14% and 12%, respectively, compared to unimodal models. It further validates the effectiveness of the multimodal fusion strategy. Overall, STMF offers an effective solution for water quality monitoring and management.**

***Index Terms*—Deep learning, multimodal fusion, smart cities, spatiotemporal modeling, water quality forecasting.**

I. Introduction

Water quality forecasting is a time-series analysis problem that predicts future trends based on historical water quality indicators. Accurate water quality forecasting can reveal the future changes in water quality over a given period, providing essential decision-making support for water resource management, pollution control, and the protection of aquatic ecosystems, aiding timely responses to environmental changes. However, water quality forecasting relies on historical monitoring data and is influenced by various factors. Fig. 1 illustrates the scenarios of water quality forecasting. First, due to the spatial layout of monitoring stations, water quality observations often exhibit strong spatial dependencies. For instance, the monitoring values at the downstream station five may be influenced by the upstream stations two, three, and four. Second, meteorological factors, such as precipitation, can cause abrupt changes in water quality indicators, leading to nonlinear [1] variations. Therefore, effectively integrating spatiotemporal features and addressing sudden meteorological changes is a key challenge in improving the adaptability of water quality forecasting models.

Conventional statistical models, such as multiple linear regression (MLR) [2], prioritize linear relationships among variables and are inadequate in capturing correlations or nonlinear dynamics among water quality indicators. These models exhibit limited adaptability to the fluid nature of the aquatic environment, thereby constraining their accuracy and effectiveness in forecasting water quality. Additionally, mechanical models necessitate extensive theoretical knowledge in biology and environmental science, which renders them impractical for real-time water quality forecasting. Similarly, while random forest (RF) [3] can capture certain nonlinear relationships, it struggles to adapt to highly variable environmental conditions.

In contrast, deep learning methodologies have emerged as robust alternatives to traditional statistical and machine learning models, adeptly capturing the complex, nonlinear, and dynamic patterns in water quality data. Techniques, e.g., convolutional neural networks (CNNs) [4], long short-term

Received 21 April 2025; accepted 16 June 2025. Date of publication 19 June 2025; date of current version 8 September 2025. This work was supported in part by the National Natural Science Foundation of China under Grant 62173013 and Grant 62473014; in part by the Beijing Natural Science Foundation under Grant L233005 and Grant 4232049; and in part by the Beihang World TOP University Cooperation Program. *(Corresponding author: Haitao Yuan.)*

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

Daming Wei is with the Chinese Academy of Environmental Planning, Beijing 100043, China (e-mail: weidm@caep.org.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).

Junfei Qiao is with the College of Information Science and Technology, Beijing University of Technology, Beijing 100124, China (e-mail: junfeiq@bjut.edu.cn).

Rajkumar Buyya is with the Cloud Computing and Distributed Systems (CLOUDS) Laboratory, School of Computing and Information Systems, University of Melbourne, Melbourne, VIC 3010, Australia (e-mail: rbuyya@unimelb.edu.au).

Digital Object Identifier 10.1109/JIOT.2025.3581282

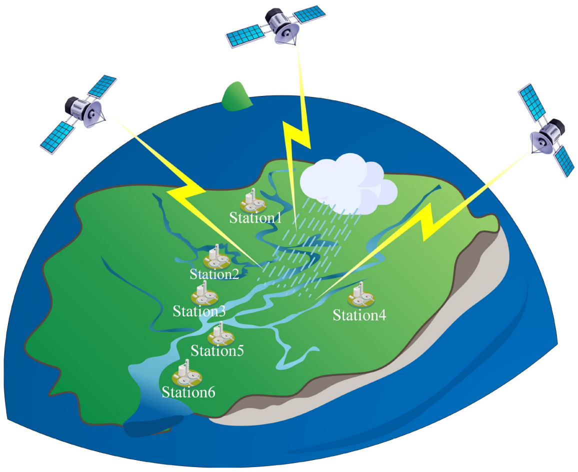

Fig. 1. Real-world water quality forecasting scenarios.

memory (LSTM) networks [5], and Transformer [6] models, demonstrate excellence in time-series forecasting by elucidating intricate relationships and exhibiting strong generalization capabilities. These techniques display significant potential in water quality forecasting [7], [8], [9], yielding results surpassing conventional models. Despite their success, deep learning approaches face limitations as they rely only on time-series data, neglecting other modalities, such as meteorological and pollutant emissions, leading to incomplete environmental context and constrained forecasting accuracy.

Based on the above analysis, to predict future water quality changes more accurately, it is essential to incorporate the spatiotemporal dependencies and meteorological factors. This work proposes a novel water quality forecasting architecture, named spatiotemporal multimodal fusion (STMF). It incorporates two key components: 1) a spatiotemporal modeling module that captures upstream–downstream dependencies among monitoring stations and 2) a low-rank cross-modal interaction fusion (LRCIF) module that fuses spatiotemporal features with precipitation data from remote-sensing images. These components jointly enhance the ability of the model to capture dynamic variations in water quality and improve forecasting accuracy. Main contributions of this work are summarized as follows.

1) STMF is a novel spatiotemporal multimodal architecture for water quality forecasting. It captures spatiotemporal dependencies by extracting temporal dynamics and spatial correlations from hydrological time series. Unlike conventional spatiotemporal forecasting models that rely solely on time-series data, STMF incorporates precipitation information from remote-sensing images as an additional modality. This multimodal design allows the model to capture complex environmental influences more effectively, improving forecasting accuracy.
2) STMF incorporates LRCIF, which is designed to integrate spatiotemporal features with precipitation features from remote-sensing image modality. This fusion method enables effective interaction between the two modalities by capturing both intramodal and cross-modal dependencies. As a result, the complementary information from time series and remote-sensing data is fully leveraged.
3) STMF is compared with six typical models across three real-world water quality datasets, and the experimental results demonstrate its superiority in long-term water quality forecasting. The average prediction accuracy of STMF improves by 22% compared to models relying solely on time-series data and by 21% compared to spatiotemporal forecasting models.

The remaining sections of this work are organized as follows. Section II reviews the related work on time-series forecasting, spatiotemporal modeling, and multimodal fusion. Section III describes the STMF components and the overall architecture. Section IV introduces experimental datasets and discusses comparative experiments. Finally, Section V concludes the work and outlines future research directions.

II. Related Work

A. Time-Series Forecasting

Time-series forecasting methods [10], [11], [12] have become a prominent research area in deep learning. Studies have explored statistical, machine learning, and deep learning methods for time-series forecasting. Traditional statistical methods, such as autoregressive integrated moving averages (ARIMA) [13], are widely applied due to their ability to model linear trends and seasonal patterns. However, the focus of ARIMA on linear relationships limits its ability to address complex and nonlinear dynamics often encountered in water quality forecasting. Machine learning techniques, such as support vector regression (SVR) [14], extend prediction capabilities by handling certain nonlinear relationships through kernel functions. However, SVR often struggles with the adaptability required for highly dynamic environmental conditions, limiting its suitability for real-time water quality applications. Deep learning models are better equipped to capture complex nonlinear relationships and long-term dependencies than statistical and traditional machine learning methods. Transformer-based models, in particular, have become mainstream in time-series forecasting. For instance, Gao et al. [15] proposed Di-Informer, an enhanced Informer-based GAN for missing-data imputation in mechanical bearing signals, leveraging ProbSparse self-attention and a binary mask to improve accuracy and fault diagnosis under varying missing data rates. Similarly, Assidiqie et al. [16] employed iTransformer for sea level forecasting in Bali, demonstrating its effectiveness over TCN and Transformer models in handling univariate time-series data.

Despite their strengths, they rely solely on historical time-series data. This work introduces an innovative approach by incorporating multimodal fusion into time-series prediction. Precipitation remote-sensing image features are utilized to supplement time-series data, enabling a more comprehensive understanding of potential factors influencing water quality changes. This multimodal approach facilitates multidimensional analysis and significantly enhances the accuracy of water quality forecasting. It effectively addresses the limitations of single time-series models in managing complex environmental factors, ultimately improving prediction reliability.

B. Spatiotemporal Forecasting

Spatiotemporal modeling is crucial in capturing spatial and temporal dependencies, making it indispensable for dynamic applications, e.g., traffic flow forecasting [17], air quality monitoring, and water quality forecasting. Spatiotemporal models enhance adaptability and prediction accuracy in complex real-world scenarios by integrating spatial relationships with temporal trends. For instance, Zheng et al. [18] proposed a spatiotemporal joint graph convolutional network (STJGCN) for traffic forecasting, which utilizes both predefined and adaptive spatiotemporal joint graphs to model dynamic correlations, thereby improving prediction accuracy in complex road networks. Likewise, Wu et al. [19] introduced a hierarchical spatiotemporal attention (HSTA) model, combining graph attention networks for spatial interactions and multihead attention for temporal dependencies, achieving state-of-the-art results in trajectory prediction tasks. Li et al. [20] proposed a Bayesian spatiotemporal graph convolutional network (DB-STGCN) for railway train delay prediction, integrating a dynamic Bayesian network with an attention-based spatiotemporal graph convolutional network. It identifies delay patterns, constructs dynamic causality graphs, and models spatiotemporal dependencies to enhance prediction accuracy. Du et al. [21] proposed a hybrid spatiotemporal response prediction model that combines CNNs for spatial feature extraction and Bi-LSTMs for temporal modeling, enabling accurate prediction of structural responses from excitations at multiple points.

Unlike the aforementioned models, this work builds upon upstream and downstream monitoring stations' dependencies and spatial distribution characteristics. In addition to traditional spatiotemporal forecasting, STMF also incorporates the influence of precipitation meteorological factors on water quality changes, further integrating spatiotemporal information with remote-sensing precipitation data, thereby providing a more comprehensive understanding of spatiotemporal dynamics. This innovation enables the model to significantly improve the accuracy and adaptability of water quality predictions under complex environmental conditions.

C. Multimodal Fusion

Multimodal fusion methods have received significant attention in various fields due to their ability to integrate complementary data from different modalities. These methods are widely applied in areas that include the integration of vision languages, sentiment analysis, and forecasting tasks. Specifically, in forecasting, multimodal approaches are effective in improving prediction accuracy by combining data from multiple sources. Yang et al. [22] proposed the multiscale inverted transform network for online oil monitoring. It integrates multimodal sensor data and uses a multiscale module for enhanced feature extraction. It outperforms traditional models in forecasting accuracy, especially in handling unknown variables. Jiang et al. [23] proposed a multimodal CNN-GNN hybrid framework for mobile traffic prediction, integrating SMS, call, and Internet data. Using ConvLSTM and Adaptive GCN, the model captures spatiotemporal dependencies and outperforms several baseline methods in real-world experiments. Lv et al. [24] proposed a learning autoencoder diffusion model for multimodal pedestrian trajectory prediction, combining pedestrian-group relationships with variational autoencoders and diffusion models. It outperforms several state-of-the-art models, enhancing prediction accuracy and real-time performance on public datasets. Guan et al. [25] proposed a multimodal Transformer-based model for egocentric early action prediction, which integrates visual data with sensor data and motion data. The model employs a two-stage optimization process to enhance the correlation between observed and unobserved video segments, improving prediction accuracy.

Unlike the above studies, this work introduces the LRCIF method in the multimodal fusion module of STMF, which integrates spatiotemporal dependencies with precipitation remote-sensing images. It first captures the cross-modal interactions between the two modalities, effectively combining their information. The interacted features are then decomposed using low-rank decomposition, which reduces computational complexity. LRCIF effectively combines multiple data sources, enabling the model to capture both the temporal evolution of water quality and the influence of precipitation environmental factors. This unique fusion strategy significantly enhances the accuracy and adaptability of long-term water quality forecasting by providing a more comprehensive understanding of water quality dynamics.

III. Proposed Methodology

This section presents the overall framework of the proposed STMF, highlighting its core components and their interactions. The modal feature extraction module is explained, which independently processes the time-series and remote-sensing images to extract relevant features from each modality. The spatiotemporal modeling module is then discussed. It focuses on how it captures complex spatial relationships among water quality monitoring stations, thereby enhancing the model's ability to understand spatiotemporal dependencies of water quality changes. Finally, the multimodal fusion module introduces LRCIF, which integrates spatiotemporal features with the precipitation features from remote-sensing images.

A. Overall Framework

Fig. 2 shows the overall architecture of STMF. STMF receives parallel inputs from the remote-sensing image modality X_r and the hydrological time-series modality X_t. In the feature extraction module, the hydrological time series undergoes batch normalization [26] and embedding [27] before being input into TimesNet to extract temporal features F_t. The remote-sensing image is processed with ResNet101 to extract image features F_r. In the spatiotemporal modeling module, GCNs take the temporal features F_t, and the adjacency matrix representing spatial information from water quality monitoring stations as inputs, producing spatiotemporal features F_{st}, that capture spatial relationships. Then, in the multimodal

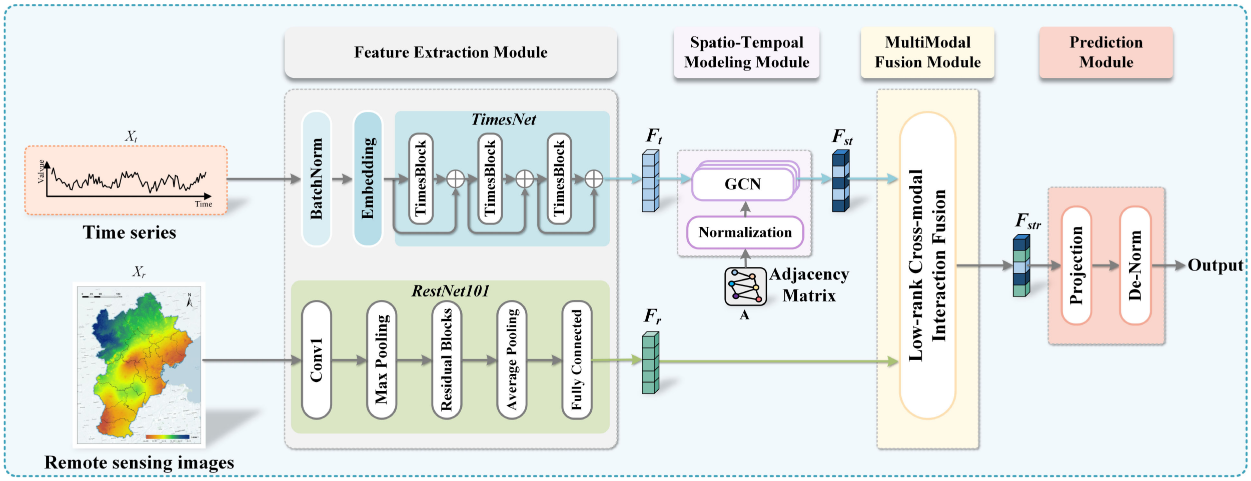

Fig. 2. Overall framework of STMF.

fusion module, LRCIF is applied to perform deep cross-modal interaction between the spatiotemporal features F_{st} and the precipitation features F_r, resulting in the fused representation F_{str}. Finally, in the prediction module, the model generates the final forecasting results based on F_{str} through projection and de-normalization.

B. Feature Extraction Module

1) Temporal Feature Extraction: Hydrological time series exhibit overlapping periodicities (e.g., daily, monthly, and annual), influenced by short-term intraperiod variations and long-term interperiod trends. To better capture these complexities, the 1-D time series is transformed into a 2-D representation, integrating both variations and overcoming the limitations of the original 1-D space.

The original 1-D arrangement for time-series data is denoted as $\mathbf{X}_{\mathrm{1D}} \in \mathbb{R}^{T\times C}$, where T represents the length and C represents recorded variables. The time series is analyzed in the frequency domain using the fast Fourier transform (FFT) [28] to identify trends and patterns in the interperiod variation. This process is given as

$$\mathbf{A} = \mathrm{Avg}(\mathrm{Amp}(\mathrm{FFT}(\mathbf{X}_{\mathrm{1D}}))) \tag{1}$$

$$\{f_1, \ldots, f_k\} = \mathrm{N}(\mathbf{A}),\ f_* \in \left\{1, \ldots, \left[\frac{T}{2}\right]\right\} \tag{2}$$

$$p_i = \left\lceil \frac{T}{f_i} \right\rceil,\ i \in \{1, \ldots, k\} \tag{3}$$

where $\mathrm{FFT}(\cdot)$ and $\mathrm{Amp}(\cdot)$ represent the FFT and the calculation of amplitude values, respectively. $\mathbf{A} \in \mathbb{R}^T$ denotes the amplitude calculated at each frequency, which is obtained by the average value $\mathrm{Avg}(\cdot)$ from C dimensions. $\mathrm{N}(\cdot)$ indicates the process of selecting k periods. In addition, due to the sparsity of the frequency domain and to reduce noise introduced by insignificant high frequencies, where the top k frequencies are chosen to minimize the noise impact.

Based on the selected top k frequencies $\{f_1, \ldots, f_k\}$ and their corresponding period lengths $\{p_1, \ldots, p_k\}$, the 1-D time series $\mathbf{X}_{\mathrm{1D}} \in \mathbb{R}^{T\times C}$ can be transformed into multiple 2-D tensors, i.e.,

$$\mathbf{X}_{\mathrm{2D}}^i = \mathrm{S}_{p_i, f_i}(\mathrm{P}(\mathbf{X}_{\mathrm{1D}})),\ i \in \{1, \ldots, k\} \tag{4}$$

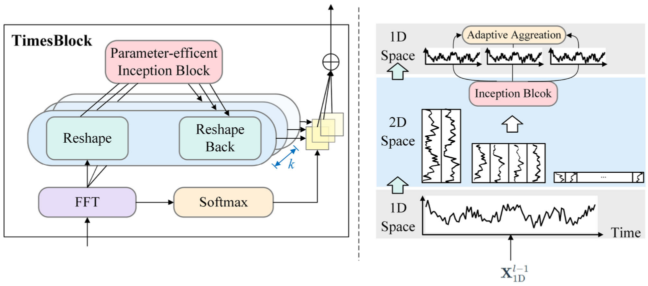

Fig. 3. Structure of TimesBlock.

where P represents padding, and $P(\cdot)$ expands the time series along the temporal dimension by padding with zeros, ensuring uniformity and compatibility with S_{p_i, f_i}. S represents the reshape operation that fills time-series data into a 2-D tensor. p_i and f_i represent the numbers of rows and columns of the 2-D tensor, respectively.

Finally, by leveraging the selected frequencies and estimated periods, a set of tensors $\mathbf{X}_{\mathrm{2D}}^1, \ldots, \mathbf{X}_{\mathrm{2D}}^k$are obtained through the fusion of remote-sensing images and hydrological time series. These tensors represent k distinct temporal 2-D variations generated across different periods.

Fig. 3 shows the structure of TimesBlock. It is constructed as a residual connection [29]. For layer l of TimesNet [30] with the input $\mathbf{X}_{\mathrm{1D}}^{l-1}$, the connection process is represented as

$$\mathbf{X}_{\mathrm{1D}}^l = O\left(\mathbf{X}_{\mathrm{1D}}^{l-1}\right) + \mathbf{X}_{\mathrm{1D}}^{l-1} \tag{5}$$

where $\mathrm{O}(\cdot)$ denotes the TimesBlock module.

After passing through all TimesBlock layers, the resulting 2-D representations are aggregated back into a 1-D representation. This is achieved by amplitude-based weights

$$\mathbf{X}_{\mathrm{1D}} = \sum_{i=1}^{k} \widehat{\mathbf{Y}}_{f_i} \times \widehat{\mathbf{X}}_{\mathrm{1D}}^i \tag{6}$$

where $\widehat{\mathbf{Y}}_{f_i}$ represents the normalized importance of each frequency, computed through Softmax based on the amplitude values, and $\widehat{\mathbf{X}}_{\mathrm{1D}}^i$ is the 1-D representation derived from the corresponding 2-D tensor.

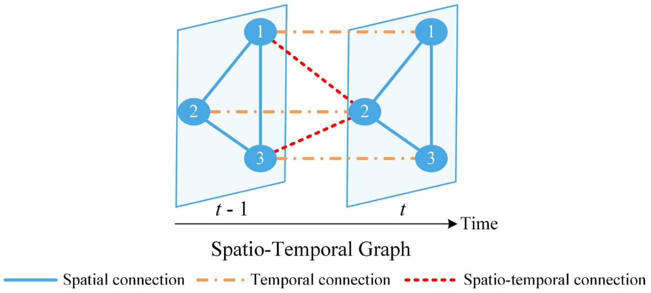

Fig. 4. Spatiotemporal modeling.

2) Precipitation Feature Extraction: This work adopts ResNet101 [31] to extract features from remote-sensing precipitation images. Compared to other shallow networks, ResNet101 is both deeper and more computationally efficient. In particular, the bottleneck structure of ResNet101 reduces the computational complexity by reducing the number of channels while maintaining the feature expression capability. In addition, the hierarchical feature extraction capability of ResNet 101 can comprehensively capture complex characteristics of remote-sensing precipitation images from the low-level edge information to the high-level semantic information, and it is especially good at handling multiscale and multilevel data. Therefore, ResNet 101 is chosen as the backbone network for feature extraction to fully use its deep and highly robust feature representation capability, thus providing a feature base with diversity and discriminative properties for subsequent tasks.

C. Spatiotemporal Modeling Module

Many studies formulate traffic forecasting and other tasks as spatiotemporal graph modeling problems. The basic assumption is that the state of each node is influenced by information from its neighboring nodes. This work applies the spatiotemporal modeling method to the water environment domain to explore the spatiotemporal characteristics of water quality monitoring data. First, since the water quality measurement of each monitoring station is affected by other upstream and downstream monitoring stations, this work constructs an adjacency matrix based on spatial geographic relationships of water quality monitoring stations to represent spatial correlations among monitoring stations. The adjacency matrix is calculated as

$$A_{i,j} = \exp\left(-\frac{\operatorname{dist}\left(s_i, s_j\right)^2}{\sigma^2}\right) \tag{7}$$

where $\operatorname{dist}(s_i, s_j)$ represents the geographic distance between station s_i and s_j, σ is the standard deviation of distances, and $A_{i,j}$ denotes the spatial edge weight between stations. This adjacency matrix, based on a Gaussian kernel function [32], effectively captures spatial interactions between water quality monitoring stations.

To further leverage information from temporal neighboring nodes, this work connects each node with its state in adjacent time steps through temporal edges, thereby constructing a spatiotemporal graph. As shown in Fig. 4, taking three water quality monitoring stations as an example, the constructed spatiotemporal relationship diagram visually demonstrates the combination of spatial and temporal edges. In the spatiotemporal graph, the spatial dependencies of water quality monitoring stations are reflected, and the temporal dependencies across different time steps are reflected.

After constructing the spatiotemporal graph, a GCN [33] is adopted to model the spatiotemporal graph, achieving a compelling fusion of spatial and temporal features. Through the feature propagation process of multiple GCN layers, the model can capture spatial correlations among stations and learn the dynamic variations of stations by incorporating temporal features. This spatiotemporal fusion mechanism enables the model better to capture the complex characteristics of hydrological time series, providing more robust support for water quality forecasting.

D. Multimodal Alignment and Fusion

1) Multimodal Alignment: Temporal and spatial alignment is essential for effective multimodal data fusion. Temporal alignment is achieved by synchronizing the timestamps of remote-sensing precipitation images and water quality time series, enabling pointwise comparability. Spatial alignment involves mapping both datasets to a unified geographic coordinate system, ensuring that precipitation data accurately correspond to water quality observations. During feature fusion, the proposed LRCIF module dynamically adjusts fusion weights based on intermodal interactions, promoting alignment in the shared feature space and enhancing the integration of features from different modalities.

2) Low-Rank Cross-Modal Interaction Fusion: Single-modality data often provides limited and context-specific information representations. Therefore, multimodal data fusion has become essential for achieving more comprehensive and reliable representations in environmental monitoring and analysis. In this context, LRCIF is introduced to capture water quality variations by fusing spatiotemporal features from hydrological time series (X_{st}) with precipitation features from remote-sensing images (X_r). Fig. 5 demonstrates the process of fusing two modalities through LRCIF. Specifically, the module first captures cross-modal dependencies between the spatiotemporal and precipitation features, and then applies a low-rank decomposition to obtain the final fused representation. This strategy improves forecasting accuracy and offers a solid foundation for real-world water environment management.

Previous studies employ the cross-attention (CA) mechanism to model cross-modal dependencies. However, directly computing pairwise similarities in a shared feature space may introduce instability due to inherent differences between modalities. To address this, the cross-diffusion attention (CDA) mechanism [34] is proposed to better capture intermodal dependencies and enable bidirectional information propagation. By integrating complementary information from both modalities, CDA enhances fusion stability and facilitates the construction of more robust multimodal representations. Building on this idea, LRCIF first calculates intramodal

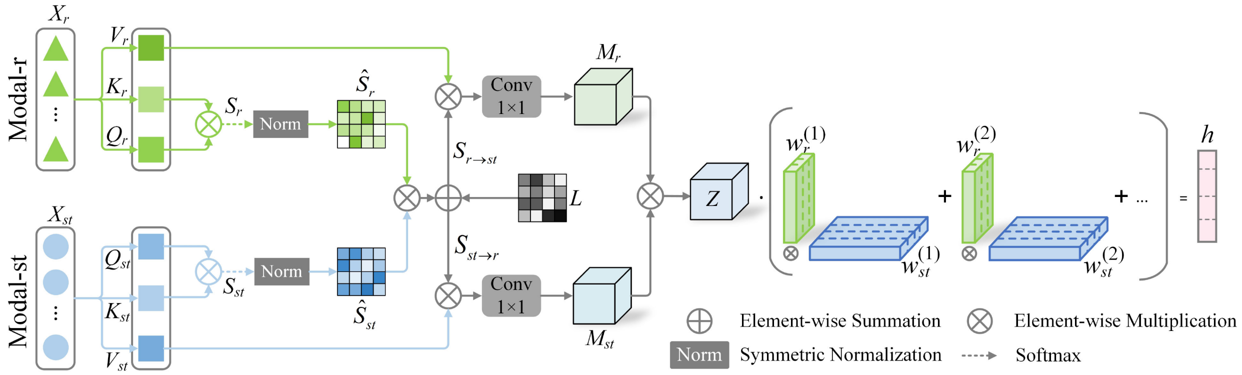

Fig. 5. Process of fusing two modalities through LRCIF.

similarity matrices S_r and S_{st} through self-attention mechanisms [35], which are then normalized as

$$\hat{S}_r = D_r^{-\frac{1}{2}} S_r D_r^{-\frac{1}{2}} \tag{8}$$

$$\hat{S}_{st} = D_{st}^{-\frac{1}{2}} S_{st} D_{st}^{-\frac{1}{2}} \tag{9}$$

where D_r and D_{st} are degree matrices. Cross-modal similarity matrices $S_{r\to st}$ is defined as

$$S_{r\to st} = \epsilon \cdot \hat{S}_r \hat{S}_{st}^T + (1-\epsilon)\cdot L \tag{10}$$

$$U_{r\to st} = S_{r\to st} V_{st} \tag{11}$$

where $L = S_r + S_{st}$, and $\epsilon \in (0,1)$ represent the balancing hyperparameter. V_{st} denotes the value matrix from the QKV decomposition of the spatiotemporal modality X_{st} and $U_{r\to st}$ is the resulting cross-modal representation of modality X_r.

Similarly, $S_{st\to r}$ and $U_{st\to r}$ can be obtained. By performing bidirectional interactions between X_{st} and X_r, we obtain the sets of representations $U_r, U_{r\to st}$ for modality X_r and $U_{st}, U_{st\to r}$ for modality X_{st}. These representations capture the mutual influence and dependencies between the spatiotemporal features and precipitation features. By combining these representations, more robust features denoted as M_r and M_{st} are obtained, e.g.,

$$M_r = f_r(U_r \| U_{r\to st}) \tag{12}$$

$$M_{st} = f_{st}(U_{st} \| U_{st\to r}) \tag{13}$$

where $\|$ represents the concatenation along the channel dimension, while $f_r(\cdot)$ and $f_{st}(\cdot)$ are two distinct 1×1 convolutional layers with separate parameters.

After obtaining the refined features M_r and M_{st}, these features are restructured into a high-dimensional tensor $\mathcal{Z}$ to integrate complementary information from both modalities while preserving their individual characteristics. The process is illustrated by the following formula:

$$\mathcal{Z} = M_r \bigotimes M_{st} \tag{14}$$

where $\bigotimes$ represents the tensor outer product. This tensor forms the foundation for efficient modality fusion. The input tensor $\mathcal{Z}$ is then passed through a linear layer $g(\cdot)$ to produce a vector representation

$$h = g(\mathcal{Z}; \mathcal{W}, b) = \mathcal{W} \cdot \mathcal{Z} + b \tag{15}$$

where $\mathcal{W}$ is the weight of the layer and b is the bias.

To avoid the high computational cost of directly generating $\mathcal{Z}$, the original tensor $\mathcal{W}$ is decomposed into m rank decomposition factors $\{\mathbf{w}_r^{(i)}\}_{i=1}^m$ and $\{\mathbf{w}_{st}^{(i)}\}_{i=1}^m$. The decomposition for these two input modalities is expressed as

$$\mathcal{W} = \sum_{i=1}^{m} \left(\mathbf{w}_r^{(i)} \bigotimes \mathbf{w}_{st}^{(i)} \right). \tag{16}$$

Then, by substituting (16) into (15), the fused representation is computed as

$$\begin{aligned} h &= \left(\sum_{i=1}^{m} w_r^{(i)} \bigotimes w_{st}^{(i)} \right) \cdot \mathcal{Z} \\ &= \left(\sum_{i=1}^{m} w_r^{(i)} \cdot M_r \right) \circ \left(\sum_{i=1}^{m} w_{st}^{(i)} \cdot M_{st} \right) \end{aligned} \tag{17}$$

where $\circ$ represents the element-wise product. As shown in (14), $\mathcal{Z}$ is constructed from M_r and M_{st}, following the same structural pattern as the low-rank decomposition of $\mathcal{W}$ in (16), and thus can also be decomposed in parallel. This operation integrates features from both modalities to produce the output vector h.

E. Training Process

Algorithm 1 shows the training process of STMF. Specifically, lines 2–4 extract multisource features from the input data, generating the spatiotemporal feature F_{st} and the precipitation feature F_r. Lines 5–9 perform cross-modal interaction between F_{st} and F_r, capturing both intramodal and cross-modal dependencies to enrich the feature representations. Line 10 fuses the interacted features M_{st} and M_r with low-rank decomposition to obtain the unified feature representation F_{str}. Lines 11–14 complete the forward computation and parameter optimization of STMF, obtaining the final prediction result Y.

IV. Performance Evaluation

A. Dataset Description and Preprocessing

1) Dataset Description: Three real-world water quality datasets are selected to verify the effectiveness of STMF, i.e., Beijing–Tianjin–Hebei (BTH), Beijing, and Alabama. Table I

Algorithm 1 Training Process of STMF

```
Input: Water quality time series ($X_t$), remote sensing precipitation images ($X_r$), Adjacency Matrix ($A$)
Output: Prediction result $Y$
for each epoch do
    Generate temporal feature $F_t$ via 2D-time variations.
    Generate spatiotemporal feature $F_{st}$ by feeding $F_t$ and $A$ into GCN.
    Generate precipitation feature $F_r$ from remote sensing images with ResNet101.
    Feed $F_{st}$ and $F_r$ into the multimodal fusion module LRCIF.
    Compute normalized intra-modal similarity matrices $\hat{S}_r$ and $\hat{S}_{st}$ using self-attention mechanism in (8) and (9).
    Compute cross-modal similarity matrices $S_{r\to st}$ and $S_{st\to r}$ in (10).
    Generate cross-modal representations $U_{r\to st}$ and $U_{st\to r}$ with $S_{r\to st}$ and $S_{st\to r}$ in (11).
    Generate precipitation interaction Feature $M_r$ and spatiotemporal interaction feature $M_{st}$ by concatenation and convolutional layer in (12) and (13), respectively.
    Generate the final fused feature $F_{str}$ by combining $M_r$ and $M_{st}$ with low-rank factor decomposition in (14)-(17).
    Generate the final prediction result $Y$ with projection and denormalize
    Compute MSE loss.
    Apply BPTT to backpropagate gradient.
    Train STMF for minimizing the loss with the Adam optimizer.
end for
```

provides an overview of these datasets. The dataset is split into training, validation, and testing sets with a ratio of 7:1:2. Specifically, the BTH and Beijing datasets are derived from publicly available data released by China's National Automatic Surface Water Quality Monitoring Stations, covering the period from 1 January 2019 to 31 December 2022. Each dataset consists of 8766 samples collected at 4-h intervals. The BTH dataset contains total nitrogen (TN) data from 24 monitoring stations, while the Beijing dataset includes dissolved oxygen (DO) data from six stations. Fig. 6 illustrates the spatial distribution of water quality monitoring stations in the BTH region. The Alabama water quality dataset consists of 19 863 samples collected from five stations in Alabama, USA, from 1 January 2021 to 31 December 2022, with data recorded hourly. Additionally, we introduce two additional precipitation remote-sensing image datasets to capture the spatial and temporal dynamics of water quality. The first dataset covers the BTH region and corresponds to both the BTH and Beijing datasets. The second dataset covers the Alabama region and corresponds to the Alabama water quality dataset. The precipitation remote-sensing images are obtained from the Goddard Center for Earth Science Data and Information Services (NASA). These images have 30-min temporal and 0.1 × 0.1° spatial resolution, providing high-resolution global precipitation data. Fig. 7 shows the precipitation remote-sensing images in the BTH region.

2) Dataset Preprocessing: The raw data from water quality monitoring stations often contains missing values, which may arise due to unpredictable weather or equipment malfunctions. To handle these missing values, linear interpolation [36] is applied by estimating the missing values based on a linear relationship between adjacent known data points. Since the occurrence of missing values is rare and sparsely distributed across the dataset, this simple yet efficient method is sufficient for accurate imputation without causing significant computational overhead. In addition, the precipitation remote-sensing images undergo several preprocessing steps to ensure their suitability for subsequent analysis. First, the images are denoised with the median filtering, which removes noise while preserving important details. Then, the pixel values are normalized to the range of [0, 1] for consistency. To meet model input requirements, all images are resized to 224 × 224 pixels. Finally, the spatial coordinates of the images are aligned with those of the water quality data to maintain spatial consistency.

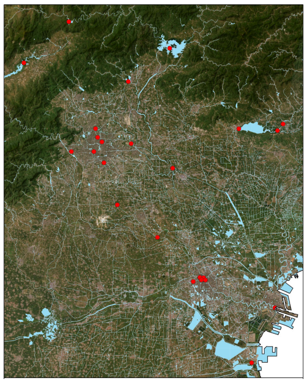

Fig. 6. Spatial distribution of water quality monitoring stations in the BTH region.

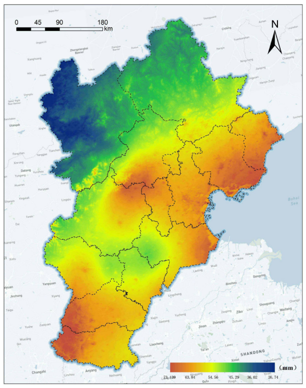

Fig. 7. Precipitation remote-sensing images in the BTH region.

TABLE I
DATASET PARAMETERS

Parameter	Datasets		
	BTH	Beijing	Alabama
Station number	24	6	5
Sampling frequency	4 hours	4 hours	1 hour
Data length	8,766	8,766	17,520
Water indicator	TN	DO	DO

B. Evaluation Metrics

To comprehensively evaluate the performance of the proposed STMF, mean-squared error (MSE) and mean absolute error (MAE) are adopted as evaluation metrics. MSE is more sensitive to large errors, which makes it effective in capturing significant deviations between predicted and true values, while emphasizing larger prediction errors. In contrast, MAE provides a straightforward interpretation of the average error magnitude by measuring the absolute differences between predicted and true values. By combining these two metrics, the evaluation offers a thorough understanding of the model's accuracy and robustness.

C. Hyperparameter Settings

The performance of STMF is influenced by several key hyperparameters, including batch size, input sequence length (S), the number of GCN layers (G), the number of attention heads (H), and the low-rank decomposition factor (R) in LRCIF. These parameters are tuned through systematic experiments. STMF is trained using the Adam optimizer with an initial learning rate of 0.0001 for a total of ten epochs, with early stopping applied using a patience of seven epochs.

The input sequence length S plays a critical role in capturing temporal dependencies. Short sequences may overlook long-term patterns, while overly long sequences may introduce noise or lead to overfitting. Table II presents the prediction results under different input lengths, and the results show that $S = 48$ yields the best performance.

The batch size also affects both convergence behavior and generalization. Smaller batch sizes tend to provide faster updates and potentially better generalization but may slow down convergence. Larger batch sizes offer more stable gradients and faster training but consume more memory. Table III shows that a batch size of 32 strikes an effective balance between efficiency and stability.

The number of attention heads (H), low-rank decomposition factor (R) in the LRCIF module, and the number of GCN layers (G) are tuned jointly to balance model expressiveness and computational efficiency. Specifically, increasing the number of attention heads H allows the model to attend to interactions from multiple subspaces, thereby capturing more diverse cross-modal relationships. However, an excessive number of heads can lead to redundant computation and higher resource demands. Likewise, the low-rank factor R controls the expressiveness of the fusion tensor. A larger R enables modeling more complex intermodal dependencies but comes at the cost of increased computation. The number of GCN layers G influences the receptive field for spatial relationships. While deeper GCNs help capture broader spatial dependencies across monitoring stations, too many layers may cause oversmoothing and degrade performance. The optimal values of H, R, and G are selected based on experimental evaluations over multiple candidate settings, specifically $H \in \{2, 4, 8\}$, $R \in \{4, 8, 16\}$, and $G \in \{1, 2, 3\}$. Table IV shows that STMF achieves the best prediction accuracy when H, R, and G are set to 4, 4, and 2, respectively.

TABLE II
MSE With Different Input Sequence Length

Input sequence length (S)	Datasets		
	BTH	Beijing	Alabama
24	0.415	0.617	0.156
32	0.414	0.613	0.152
48	**0.412**	**0.610**	**0.150**
96	0.418	0.621	0.160

TABLE III
MSE With Different Batch Size

Batch size	Datasets		
	BTH	Beijing	Alabama
16	0.417	0.626	0.155
32	**0.412**	**0.610**	**0.150**
64	0.413	0.612	0.156
128	0.414	0.625	0.153

TABLE IV
Prediction Results of STMF With Different H, R, and G

(H, R, G)	MSE	MAE
(2, 4, 1)	0.435	0.391
(2, 8, 2)	0.426	0.375
(2, 16, 3)	0.420	0.367
(4, 4, 1)	0.428	0.382
(4, 4, 2)	**0.412**	**0.363**
(4, 4, 3)	0.419	0.365
(4, 8, 1)	0.424	0.373
(4, 8, 2)	0.418	0.365
(4, 16, 3)	0.422	0.372
(8, 4, 1)	0.430	0.384
(8, 8, 2)	0.421	0.370
(8, 16, 3)	0.423	0.367

D. Benchmark Models

To verify the effectiveness of STMF, six baseline models are selected for comparison, which include PatchTST, Autoformer, and FEDformer representing Transformer-based models, DLinear representing MLP-based approaches, and STSGCN and ASTGCN representing spatiotemporal forecasting models. The benchmark methods are listed as follows.

1) *PatchTST [37]:* It utilizes patches to capture local patterns in the time-series data.
2) *Autoformer [38]:* It introduces a decomposition-based transformer architecture to capture long-term dependencies and seasonal trends.
3) *FEDformer [39]:* It adopts frequency-enhanced block wise decomposition to jointly model global and local temporal dynamics.
4) *DLinear [40]:* It uses a simple linear model to forecast long-term trends with high computational efficiency.
5) *STSGCN [41]:* It models both spatial and temporal dependencies synchronously instead of treating them separately.
6) *ASTGCN [42]:* It combines spatiotemporal graph convolution with attention mechanisms to dynamically capture spatial and temporal dependencies.

E. Comparative Experiments

The comparison experiment is conducted on a server equipped with an Intel Xeon 6248R processor and a GTX3090 GPU, ensuring the necessary computational resources for efficient model training and evaluation. STMF and other benchmark models are implemented using PyTorch.

TABLE V
COMPARISON OF AVERAGE PREDICTION PERFORMANCE OF DIFFERENT MODELS ON VARIOUS DATASETS

Models	STMF (Ours)		PatchTST (2023)		Autoformer (2021)		FEDformer (2022)		DLinear (2023)		STSGCN (2020)		ASTGCN (2019)	
Metric	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
BTH	**0.386**	**0.331**	0.447	0.362	0.519	0.416	0.560	0.458	0.463	0.388	0.511	0.407	0.517	0.411
Beijing	**0.575**	**0.484**	0.640	0.521	0.734	0.587	0.736	0.588	0.781	0.678	0.727	0.583	0.731	0.586
Alabama	**0.138**	**0.180**	0.192	0.216	0.245	0.325	0.243	0.311	0.162	0.214	0.234	0.316	0.236	0.318

TABLE VI
COMPARISON OF PREDICTION PERFORMANCE OF STMF AND OTHER BASELINE MODELS ON VARIOUS DATASETS

	Models	STMF		PatchTST		Autoformer		FEDformer		DLinear		STSGCN		ASTGCN	
	Horizon	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
BTH	24	**0.196**	**0.219**	0.217	0.235	0.312	0.309	0.337	0.358	0.242	0.270	0.295	0.292	0.304	0.306
	48	**0.279**	**0.264**	0.307	0.296	0.385	0.361	0.421	0.402	0.332	0.328	0.368	0.344	0.377	0.358
	72	**0.334**	**0.313**	0.388	0.341	0.458	0.396	0.500	0.440	0.409	0.371	0.440	0.378	0.452	0.391
	96	**0.412**	**0.363**	0.466	0.383	0.531	0.428	0.573	0.468	0.485	0.407	0.512	0.408	0.524	0.426
	120	**0.468**	**0.380**	0.541	0.414	0.605	0.458	0.654	0.502	0.559	0.440	0.588	0.440	0.597	0.453
	192	**0.659**	**0.448**	0.763	0.502	0.824	0.535	0.874	0.576	0.753	0.514	0.815	0.518	0.818	0.530
BeiJing	24	**0.468**	**0.403**	0.516	0.441	0.684	0.553	0.677	0.551	0.611	0.583	0.661	0.535	0.670	0.548
	48	**0.557**	**0.462**	0.611	0.496	0.720	0.572	0.723	0.578	0.737	0.656	0.705	0.560	0.714	0.569
	72	**0.603**	**0.497**	0.661	0.535	0.742	0.590	0.741	0.590	0.796	0.688	0.723	0.573	0.732	0.581
	96	**0.610**	**0.505**	0.671	0.540	0.757	0.600	0.759	0.602	0.843	0.712	0.738	0.587	0.748	0.599
	120	**0.616**	**0.526**	0.696	0.557	0.756	0.601	0.772	0.614	0.868	0.724	0.741	0.590	0.752	0.597
	192	**0.597**	**0.513**	0.686	0.558	0.745	0.603	0.757	0.606	0.833	0.707	0.731	0.592	0.739	0.594
Alabama	24	**0.069**	**0.117**	0.074	0.126	0.192	0.278	0.148	0.247	0.088	0.151	0.135	0.234	0.142	0.241
	48	**0.093**	**0.149**	0.106	0.161	0.207	0.296	0.206	0.291	0.120	0.184	0.193	0.277	0.200	0.287
	72	**0.124**	**0.174**	0.134	0.187	0.225	0.309	0.236	0.312	0.147	0.206	0.223	0.298	0.224	0.304
	96	**0.150**	**0.192**	0.161	0.207	0.251	0.332	0.255	0.323	0.170	0.224	0.235	0.316	0.243	0.324
	120	**0.166**	**0.205**	0.179	0.220	0.268	0.351	0.275	0.332	0.191	0.239	0.255	0.338	0.261	0.344
	192	**0.228**	**0.243**	0.245	0.261	0.326	0.384	0.336	0.362	0.256	0.280	0.312	0.368	0.319	0.377

Table V compares the average prediction performance of STMF and other baseline models. In the BTH dataset, compared to Transformer-based models that utilize only time series as the input, including PatchTST, Autoformer, and FEDformer, STMF reduces the average MSE and MAE by 23% and 19%, respectively. Compared to the MLP-based time-series model DLinear, the reductions in MSE and MAE are 20% and 17%, respectively. This performance advantage is mainly attributed to STMF's ability to incorporate spatial information and key environmental factors, enabling it better to capture nonlinear variations and complex water quality patterns. Furthermore, compared to spatiotemporal forecasting models, such as STSGCN and ASTGCN, STMF achieves reductions of 21% and 19% in MSE and MAE, respectively. Unlike these models, which primarily rely on graph structures or static spatial relationships, STMF integrates spatiotemporal dependencies with remote-sensing precipitation images, enabling a broader understanding of dynamic environmental influences and enhancing accuracy and robustness.

Table VI presents the prediction performance of STMF and six baseline models across different forecasting horizons in the set of {24, 48, 72, 96, 120, 192}. The bold texts show the best prediction results. Across all datasets, STMF consistently demonstrates superior performance, achieving lower MSE and MAE values compared to other models. In the BTH dataset, compared with the best performing benchmark model, STMF improves MSE and MAE by 14% and 12%, respectively, in the 192-step prediction. In the 24-step prediction, STMF improves MSE and MAE by 10% and 7%, respectively. These results highlight that STMF's advantages are more evident in long-term forecasting. Its ability to integrate spatiotemporal dependencies and precipitation information enables it to capture extended trends, account for accumulated environmental effects, and enhance long-term forecasting performance. Figs. 8–13 show the MSE and MAE values of different models under forecasting horizons in the set of {24, 48, 72, 96, 120, 192}.

Moreover, to assess the statistical significance of the experimental results, a nonparametric Wilcoxon signed-rank test [43]

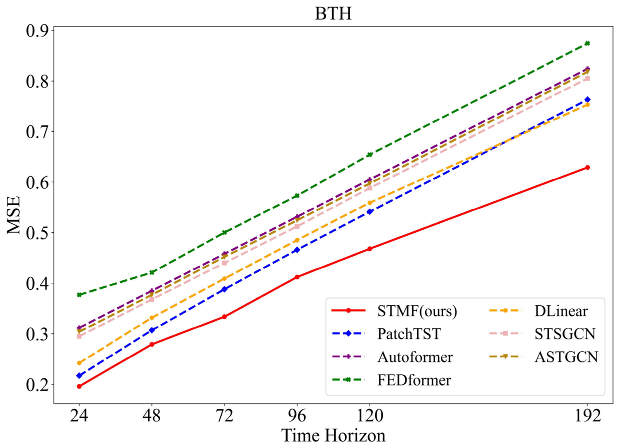

Fig. 8. MSE of multistep prediction on the BTH dataset.

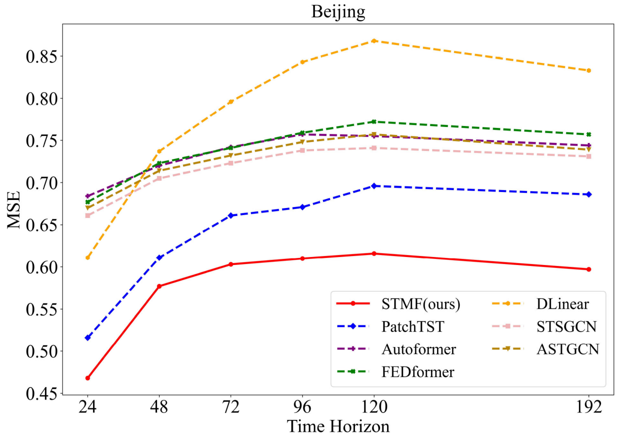

Fig. 9. MSE of multistep prediction on the Beijing dataset.

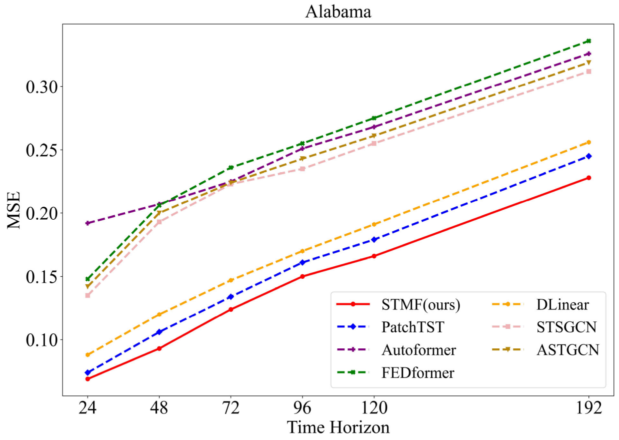

Fig. 10. MSE of multistep prediction on the Alabama dataset.

is performed to compare the performance of STMF with other baseline models across all prediction horizons and datasets. The test is conducted with a significance level of 0.05 ($\alpha = 0.05$), assuming the one-sided hypothesis that STMF yields lower MSE and MAE. The results indicate that STMF significantly outperforms the other models in both MSE and MAE, with p-values less than 0.05 for all comparisons.

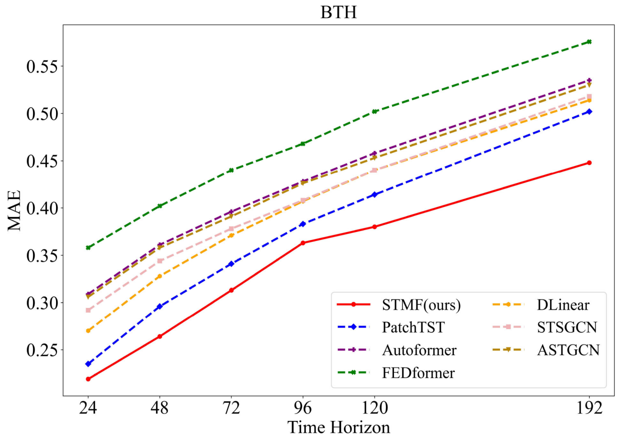

Fig. 11. MAE of multistep prediction on the BTH dataset.

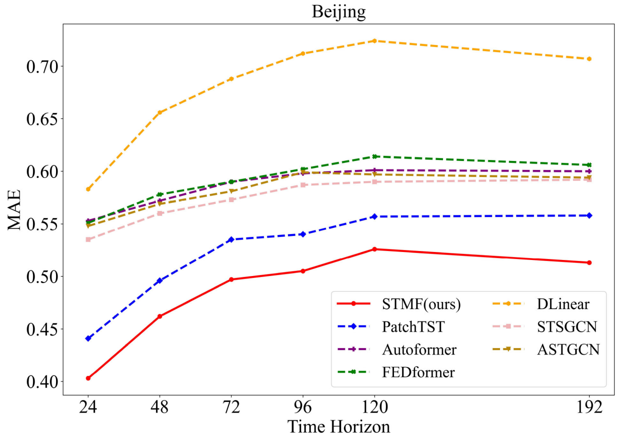

Fig. 12. MAE of multistep prediction on the Beijing dataset.

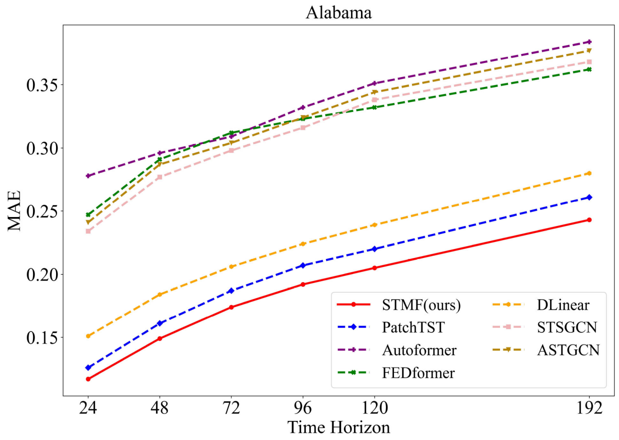

Fig. 13. MAE of multistep prediction on the Alabama dataset.

Fig. 15 compares the predicted and actual values of the TN indicator at the Huairou Reservoir Station. The line represents the ground truth, while the red line denotes the predictions by STMF. It is evident that STMF's predictions closely follow the true values, maintaining trend consistency and exhibiting lower error magnitudes than other models, thereby demonstrating its superior predictive performance.

To evaluate the fusion capability of LRCIF within STMF, a comparison is conducted against three widely used multimodal

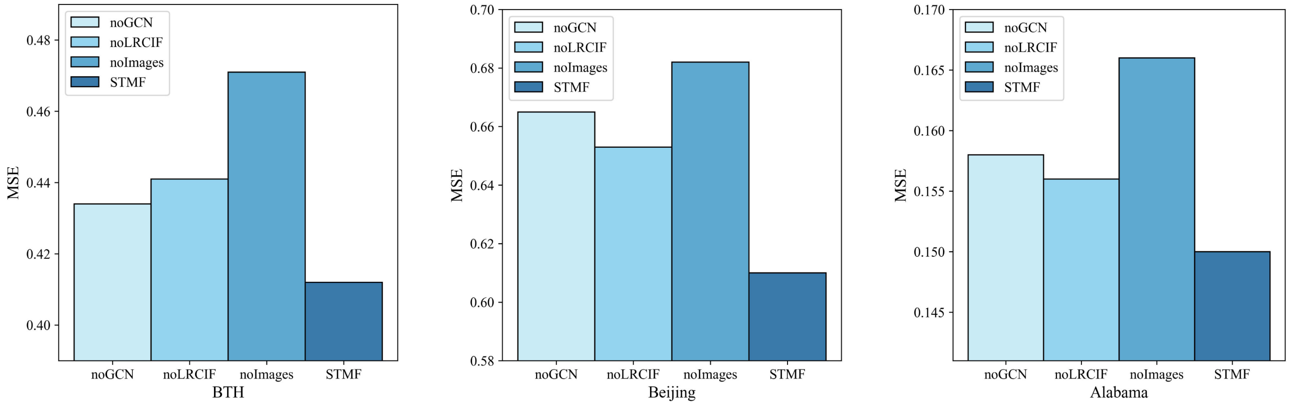

Fig. 14. Ablation studies on three real-world datasets.

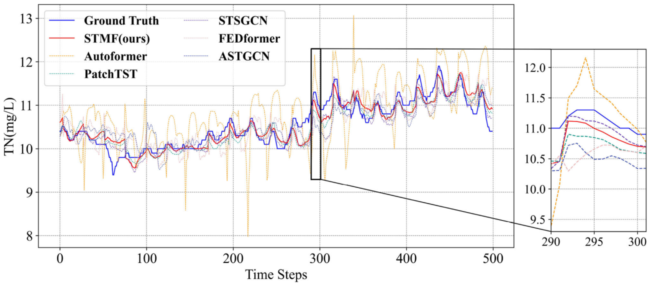

Fig. 15. Comparison between predicted and true values across all models.

TABLE VIII
COMPUTATIONAL COMPLEXITY COMPARISON BETWEEN STGFT AND OTHER BASELINE MODELS

Models	Number of parameters	FLOPs
PatchTST	3.31×10^6	3.67×10^9
Autoformer	1.05×10^7	1.90×10^{10}
FEDformer	1.47×10^7	2.83×10^{10}
DLinear	4.70×10^3	8.94×10^5
STSGCN	7.15×10^7	5.83×10^{10}
ASTGCN	6.19×10^5	1.53×10^9
STMF	5.91×10^5	1.11×10^{10}

TABLE VII
COMPARISON OF PREDICTION PERFORMANCE OF DIFFERENT FUSION MODULES ON VARIOUS DATASETS

Models	LRCIF		MBT		TFN		Concatenation	
Metric	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
BTH	**0.412**	**0.363**	0.423	0.358	0.437	0.362	0.441	0.367
Beijing	**0.610**	**0.505**	0.628	0.513	0.636	0.520	0.653	0.527
Alabama	**0.150**	**0.192**	0.152	0.176	0.153	0.182	0.156	0.189

feature-level fusion methods: 1) concatenation; 2) tensor fusion network (TFN) [44]; and 3) multimodal bottleneck transformer (MBT) [45]. Concatenation directly combines multimodal features along a specific dimension to create a unified, simple and efficient representation while retaining the information from each modality. TFN employs a tensor fusion strategy to process multimodal data, enabling end-to-end learning of both intra- and inter-modal information. MBT utilizes a self-attention mechanism and bilinear tensor structure to capture dependencies between different modalities flexibly. By comparing these models, the advantages and effectiveness of LRCIF in multimodal data fusion are better demonstrated. The comparative prediction performance of varying fusion models is shown in Table VII. Table VIII shows the computational complexity comparison between STMF and other baseline models. STMF has a moderate number of parameters and floating-point operations (FLOPs) among the models evaluated.

F. Ablation Studies

The ablation study aims to evaluate the contribution of each core component in the STMF model. Specifically, we compare STMF with three variant models: 1) STMF-noGCN, removing the spatiotemporal modeling module while keeping all other components unchanged; 2) STMF-noLRCIF, replacing the proposed LRCIF fusion mechanism with a simple concatenation strategy; and 3) STMF-noImages, removing the remote-sensing image input and relying solely on hydrological time-series data. This last variant can be regarded as a traditional spatiotemporal forecasting model without multimodal input. Fig. 14 presents the ablation results on the BTH, Beijing, and Alabama datasets, with the prediction horizon of 96. The results show that STMF consistently outperforms all ablated variants in terms of MSE, indicating that multimodal input, fusion strategy, and spatiotemporal modeling each play a critical role in enhancing prediction accuracy.

Among the three variants, STMF-noImages exhibits a relatively severe performance decline, highlighting the critical role of remote-sensing images in capturing external environmental factors like precipitation. In real-world scenarios, the uneven spatial distribution of precipitation often causes abrupt changes in water quality indicators across regions. For instance, a sudden rainfall event in an upstream area may rapidly alter local water conditions. In contrast, a downstream site without rainfall may still exhibit significant changes due to river flow and other hydrological processes. Such nonlinear and nonstationary patterns are intricate to capture using time-series data alone. Remote-sensing images provide large-scale spatial coverage, allowing the model to capture abrupt variations and compensate for the sparse spatiotemporal distribution of water quality monitoring stations. As a result, the model gains an improved understanding of external environmental drivers.

STMF-noGCN also shows a notable decline in performance. Without the GCN module, the model fails to capture spatial dependencies among monitoring stations, weakening its ability to learn spatial correlations in water quality across regions. In real-world river systems, stations are typically arranged along upstream–downstream paths, where water quality changes in one region may directly or indirectly affect downstream areas. In addition, STMF-noLRCIF leads to a moderate performance drop. Although feature fusion is still performed via a concatenation strategy, this simple approach fails to capture the deeper interactions between modalities. In contrast, the proposed LRCIF method enables the extraction of more relevant and complementary features from remote-sensing images and time-series data, thereby improving cross-modal synergy, enhancing the model's representational capacity, and increasing prediction robustness.

V. Conclusion and Future Work

Water quality forecasting is critical in water environment management and is essential in preventing and controlling water pollution. With the increasing deployment of monitoring devices, water environment data has become more diverse and multimodal. Several factors, including the spatial relationships between monitoring stations, pollutant emissions, and precipitation, influence water quality prediction outcomes. However, most existing water quality forecasting models rely solely on a single time series as input, failing to fully exploit and leverage the interrelationships among multimodal data. This work proposes an STMF model for long-term water quality forecasting. It captures the dynamic correlations of water quality variations between spatially adjacent monitoring stations through spatiotemporal modeling. Additionally, STMF introduces the LRCIF method, which facilitates deep interaction and fusion of spatiotemporal features with precipitation features from remote-sensing images. This multimodal integration of time-series data with spatial location information and precipitation as a meteorological factor significantly improves the accuracy of long-term water quality forecasting. Experimental results demonstrate that STMF substantially outperforms existing state-of-the-art models on three real-world water quality datasets, validating its effectiveness and superiority. Specifically, for the long-term forecasting task with a 192-step horizon, STMF improves MSE and MAE by 14% and 12%, respectively, compared to unimodal models.

In future work, we intend to employ intelligent optimization algorithms [46], [47], [48] to fine-tune model parameters to enhance prediction accuracy further. Moreover, we also plan to introduce a dynamic mechanism into the fusion module to adjust feature weights for each modality adaptively. This enhancement will improve the adaptability and robustness of the model and enable a more flexible integration of the modality under varying environmental conditions.

References

[1] B. Wang, R. Luo, and H. Chen, "Multichannel closed-loop seismic acoustic impedance estimation with nonlinear correction," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, Oct. 2024, Art. no. 5933609.

[2] Z. He, "Refining time-space traffic diagrams: A simple multiple linear regression model," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 2, pp. 1465–1475, Feb. 2024.

[3] M. H. R. Sales, S. de Bruin, C. Souza, and M. Herold, "Land use and land cover area estimates from class membership probability of a random forest classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Jun. 2022, Art. no. 4402711.

[4] Y. Wang, L. Gao, Y. Gao, and X. Li, "A graph guided convolutional neural network for surface defect recognition," *IEEE Trans. Autom. Sci. Eng.*, vol. 19, no. 3, pp. 1392–1404, Jul. 2022.

[5] J. Bi, X. Zhang, H. Yuan, J. Zhang, and M. Zhou, "A hybrid prediction method for realistic network traffic with temporal convolutional network and LSTM," *IEEE Trans. Autom. Sci. Eng.*, vol. 19, no. 3, pp. 1869–1879, Jul. 2022.

[6] X. Tao, C. Adak, P.-J. Chun, S. Yan, and H. Liu, "ViTALnet: Anomaly on industrial textured surfaces with hybrid transformer," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–13, Feb. 2023.

[7] G. Wang, H. Chen, S. Jiang, H. Han, and J. Qiao, "Neurodynamics-driven prediction model for state evolution of coastal water quality," *IEEE Trans. Instrum. Meas.*, vol. 73, pp. 1–10, 2025.

[8] 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.

[9] Z. Wang, X. Wu, J. Bi, H. Yuan, J. Zhang, and M. Zhou, "Long-term water quality prediction based on intelligent optimization and seasonal-trend decomposition," in *Proc. IEEE 20th Int. Conf. Autom. Sci. Eng. (CASE)*, Bari, Italy, 2024, pp. 264–269.

[10] L. Li, L. Qin, X. Qu, J. Zhang, H. Li, and B. Ran. "Travel time prediction for highway network based on the ensemble empirical mode decomposition and random vector functional link network," *Appl. Soft Comput.*, vol. 73, pp. 921–932, Oct. 2018.

[11] L. Li, H. Zhou, H. Liu, C. Zhang, and J. Liu, "A hybrid method coupling empirical mode decomposition and a long short-term memory network to predict missing measured signal data of SHM systems," *Struct. Health Monit.*, vol. 20, no. 4, pp. 1778–1793, Jul. 2021.

[12] J. Bi et al., "Long-term water quality prediction with transformer-based spatial-temporal graph fusion," *IEEE Trans. Autom. Sci. Eng.*, vol. 22, pp. 11392–11404, 2025.

[13] F. Wu, R. Jing, X.-P. Zhang, F. Wang, and Y. Bao, "A combined method of improved grey bp neural network and MEEMD-ARIMA for day-ahead wave energy forecast," *IEEE Trans. Sustain. Energy*, vol. 12, no. 4, pp. 2404–2412, Oct. 2021.

[14] M. Khakifirooz, C.-F. Chien, and Y.-J. Chen, "Dynamic support vector regression control system for overlay error compensation with stochastic metrology delay," *IEEE Trans. Autom. Sci. Eng.*, vol. 17, no. 1, pp. 502–512, Jan. 2020.

[15] S. Gao, W. Hao, Q. Wang, and Y. Zhang, "Missing-data filling method based on improved informer model for mechanical-bearing fault diagnosis," *IEEE Trans. Instrum. Meas.*, vol. 73, pp. 1–10, Aug. 2024.

[16] I. Assidiqie, D. Adytia, and I. A. Hakiki, "iTransformer application in sea level forecasting: Case study of Bali," in *Proc. Int. Conf. Artif. Intell., Blockchain, Cloud Comput., Data Anal. (ICoABCD)*, Indonesia, Bali, 2024, pp. 13–18.

[17] L. Li, L. Qin, X. Qu, J. Zhang, Y. Wang, and B. Ran, "Day-ahead traffic flow forecasting based on a deep belief network optimized by the multi-objective particle swarm algorithm," *Knowl. Based Syst.*, vol. 172, pp. 1–14, Jan. 2019.

[18] C. Zheng et al., "Spatio-temporal joint graph convolutional networks for traffic forecasting," *IEEE Trans. Knowl. Data Eng.*, vol. 36, no. 1, pp. 372–385, Jan. 2024.

[19] Y. Wu et al., "HSTA: A hierarchical spatio-temporal attention model for trajectory prediction," *IEEE Trans. Veh. Technol.*, vol. 70, no. 11, pp. 11295–11307, Nov. 2021.

[20] J. Li, X. Xu, X. Ding, J. Liu, and B. Ran, "Bayesian spatio-temporal graph convolutional network for railway train delay prediction," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 7, pp. 8193–8208, Jul. 2024.

[21] B. Du, C. Lin, L. Sun, Y. Zhao, and L. Li, "Response prediction based on temporal and spatial deep learning model for intelligent structural health monitoring," *IEEE Internet Things J.*, vol. 9, no. 15, pp. 13364–13375, Aug. 2022.

[22] G. Yang, H. Tao, S. He, W. Feng, R. Du, and Y. Zhong, "Multimodal time series forecasting for online oil monitoring of petrochemical pelletizer gearbox using multiscale inverted transform network," *IEEE Internet Things J.*, vol. 12, no. 8, pp. 10876–10884, Apr. 2025.

[23] W. Jiang, Y. Zhang, H. Han, Z. Huang, Q. Li, and J. Mu, "Mobile traffic prediction in consumer applications: A multimodal deep learning approach," *IEEE Trans. Consum. Electron.*, vol. 70, no. 1, pp. 3425–3435, Feb. 2024.

[24] K. Lv, L. Yuan, and X. Ni, "Learning autoencoder diffusion models of pedestrian group relationships for multimodal trajectory prediction," *IEEE Trans. Instrum. Meas.*, vol. 73, pp. 1–12, 2024.

[25] W. Gua et al., "Egocentric early action prediction via multimodal transformer-based dual action prediction," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 9, pp. 4472–4483, Sep. 2023.

[26] H. Peng, Y. Yu, and S. Yu, "Re-thinking the effectiveness of batch normalization and beyond," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 1, pp. 465–478, Jan. 2024.

[27] L. Fang, Y. Luo, K. Feng, K. Zhao, and A. Hu, "A knowledge-enriched ensemble method for word embedding and multi-sense embedding," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 6, pp. 5534–5549, Jun. 2023.

[28] Y. Liao, H. Li, Y. Cao, Z. Liu, W. Wang, and X. Liu, "Fast fourier transform with multihead attention for specific emitter identification," *IEEE Trans. Instrum. Meas.*, vol. 73, pp. 1–12, Dec. 2024.

[29] 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, Feb. 2023.

[30] S. Li, W. Li, L. Chen, H. Jiang, J. Zhang, and D. W. Gao, "Real-time robust state estimation for large-scale low-observability power-transportation system based on meta physics-informed graph timesnet," *IEEE Trans. Smart Grid*, vol. 15, no. 6, pp. 5500–5513, Nov. 2024.

[31] F. Gurkan, L. Cerkezi, O. Cirakman, and B. Gunsel, "TDIOT: Target-driven inference for deep video object tracking," *IEEE Trans. Image Process.*, vol. 30, pp. 7938–7951, 2021.

[32] Y.-Q. Liu, X. Du, H.-L. Shen, and S.-J. Chen, "Estimating generalized gaussian blur kernels for out-of-focus image deblurring," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 3, pp. 829–843, Mar. 2021.

[33] D.-H. Zhai, Z. Yan, and Y. Xia, "Lightweight multiscale spatiotemporal locally connected graph convolutional networks for single human motion forecasting," *IEEE Trans. Autom. Sci. Eng.*, vol. 21, no. 3, pp. 4768–4777, Jul. 2024.

[34] 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.

[35] Y.-L. He, X.-Y. Li, Y. Xu, Q.-X. Zhu, and S. Lu, "Novel distributed GRUs based on hybrid self-attention mechanism for dynamic soft sensing," *IEEE Trans. Autom. Sci. Eng.*, vol. 21, no. 4, pp. 5161–5172, Oct. 2024.

[36] L.-F. Shi, Y.-F. Dai, H. Yin, and Y. Shi, "Pedestrian trajectory projection based on adaptive interpolation factor linear interpolation quaternion attitude estimation method," *IEEE Trans. Instrum. Meas.*, vol. 74, pp. 1–9, Mar. 2025.

[37] Y. Liu, W. Wang, L. Chang, and J. Tang, "MSWI multi-temperature prediction based on patch time series transformer," in *Proc. 36th Chin. Control Decis. Conf. (CCDC)*, Xi'an, China, 2024, pp. 2369–2373.

[38] C. Yang, C. Yang, X. Zhang, and J. Zhang, "Multisource information fusion for autoformer: Soft sensor modeling of FeO content in iron ore sintering process," *IEEE Trans. Ind. Informat.*, vol. 19, no. 12, pp. 11584–11595, Dec. 2023.

[39] B. Deng, Y. Wu, S. Liu, and Z. Xu, "Wind speed forecasting for wind power production based on frequency-enhanced transformer," in *Proc. 4th Int. Conf. Mach. Learn., Big Data Bus. Intell. (MLBDBI)*, Shanghai, China, 2022, pp. 151–155.

[40] G. Wang, Y. Liao, L. Guo, J. Geng, and X. Ma, "DLinear photovoltaic power generation forecasting based on reversible instance normalization," in *Proc. IEEE 12th Data Driven Control Learn. Syst. Conf. (DDCLS)*, Xiangtan, China, 2023, pp. 990–995.

[41] Y. Hu et al., "A multiview spatial-temporal adaptive transformer-GRU framework for traffic flow prediction," *IEEE Internet Things J.*, vol. 12, no. 6, pp. 7114–7132, Mar. 2025.

[42] Y. Chen et al., "Graph attention network with spatial-temporal clustering for traffic flow forecasting in intelligent transportation system," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 8, pp. 8727–8737, Aug. 2023.

[43] Y. K. Saheed, B. F. Balogun, B. J. Odunayo, and M. Abdulsalam, "Microarray gene expression data classification via Wilcoxon sign rank sum and novel grey wolf optimized ensemble learning models," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 20, no. 6, pp. 3575–3587, Aug. 2023.

[44] Z.-Z. Liu, T. Sun, X.-M. Sun, and W.-Y. Cui, "Estimating remaining useful life of aircraft engine system via a novel graph tensor fusion network based on knowledge of physical structure and thermodynamics," *IEEE Trans. Instrum. Meas.*, vol. 74, pp. 1–14, Mar. 2025.

[45] Z. Wang, L. Yu, X. Ding, X. Liao, and L. Wang, "Shared-specific feature learning with bottleneck fusion transformer for multi-modal whole slide image analysis," *IEEE Trans. Med. Imag.*, vol. 42, no. 11, pp. 3374–3383, Nov. 2023.

[46] 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.

[47] H. Yuan, J. Bi, Z. Wang, J. Yang, and J. Zhang, "Partial and cost-minimized computation offloading in hybrid edge and cloud systems," *Expert Syst. Appl.*, vol. 250, pp. 1–13, Sep. 2024.

[48] 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.

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.

From 2013 to 2015, she was a Postdoctoral Researcher with the Department of Automation, Tsinghua University, Beijing, China. From 2018 to 2019, she was a Visiting Research Scholar with the Department of Electrical and Computer Engineering, New Jersey Institute of Technology, Newark, NJ, USA. She is currently a Professor with the College of Computer Science, Beijing University of Technology, Beijing. 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.

Prof. Bi is currently an Associate Editor of IEEE TRANSACTIONS ON SYSTEMS MAN AND CYBERNETICS: SYSTEMS.

Xiangxi Wu (Student Member, IEEE) received the B.E. degree in software engineering from Beijing University of Technology, Beijing, China, in 2024, 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.

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 a Deputy Director of the Department of Science and Technology Innovation, Wenchang International Aerospace City, Hainan, China. He is currently an Associate Professor with the School of Automation Science and Electrical Engineering, Beihang University, Beijing, China. His research interests include Internet of Things, edge computing, deep learning, data-driven optimization, and computational intelligence algorithms.

Dr. Yuan 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 named in the world's top 2% of Scientists List. He is an Associate Editor of IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS: SYSTEMS, IEEE INTERNET OF THINGS JOURNAL, and *Expert Systems With Applications*.

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 mobile-edge computing, task scheduling, intelligent optimization algorithms, and big data.

Mr. Wang received the Best Paper Award in 2024 ICAIS and ISAS and the Best Application Paper Award in the 21st IEEE ICNSC.

Daming Wei received the master's degree in water chemistry and microbiology from the University of Poitiers, Poitiers, France.

He is currently a Senior Engineer with the Chinese Academy of Environmental Planning, Beijing, China. He has published more than 30 papers in national and international journals. His research areas include watershed pollution control and water ecological environmental protection.

Renren Wu received the Ph.D. degree in environmental engineering from South China University of Technology, Guangzhou, China, in 2011.

He is currently a Professor with South China Institute of Environmental Sciences, Ministry of Ecology and Environment of the People's Republic of China, Guangzhou, where he is currently 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, Chicago, IL, USA, in 2000.

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, Dallas, TX, USA. 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.

Junfei Qiao (Senior Member, IEEE) received the B.E. and M.E. degrees in control engineering from Liaoning Technical University, Huludao, China, in 1992 and 1995, respectively, and the Ph.D. degree from Northeastern University, Shenyang, China, in 1998.

From 1998 to 2000, he was a Postdoctoral Fellow with the School of Automatics, Tianjin University, Tianjin, China. In 2000, he joined Beijing University of Technology, Beijing, China, where he is currently a Professor and the Vice-President. He is also the Director of the Intelligence Systems Laboratory, Beijing. His current research interests include neural networks, intelligent systems, self-adaptive/learning systems, and process control systems.

Rajkumar Buyya (Fellow, IEEE) received the B.E. degree in computer science and engineering from the University of Mysore, Mysuru, India, in 1992, the M.E. degree in computer science and engineering from Bengaluru University, Bengaluru, India, in 1995, and the Ph.D. degree in computer science and software engineering from Monash University, Melbourne, VIC, Australia, in 2002.

He is a Redmond Barry Distinguished Professor and the Director of the Cloud Computing and Distributed Systems (CLOUDS) Laboratory, University of Melbourne, Melbourne. He has authored over 850 publications and seven textbooks.

Dr. Buyya is a highly cited computer science and software engineering author worldwide, with over 155 600 citations and an H-index of 170. Thomson Reuters recognized him as a "Web of Science Highly Cited Researcher" from 2016 to 2021. He was a Future Fellow of the Australian Research Council from 2012 to 2016.