A Water Quality Prediction Model Based on Low-Rank Multimodal Fusion Patch Transformer

Xiankun Shi
College of Computer Science
Beijing University of Technology
Beijing, China
shixk@emails.bjut.edu.cn

Junqi Zhang
College of Computer Science
Beijing University of Technology
Beijing, China
jq-zhang15@tsinghua.org.cn

Yibo Li
College of Computer Science
Beijing University of Technology
Beijing, China
liyibo0206@163.com

Ziqi Wang
School of Software Technology
Zhejiang University
Ningbo, China
wangziqi0312@163.com

Renen Wu
South China Institute of Environmental Sciences
Ministry of Ecology and Environment
of the People's Republic of China
Guangzhou, China
wurenren@scies.org

Jing Bi
College of Computer Science
Beijing University of Technology
Beijing, China
bijing@bjut.edu.cn

2025 International Conference on Networking, Sensing and Control (ICNSC) | 979-8-3315-9749-8/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICNSC66229.2025.00034

***Abstract*—Accurate prediction of water quality parameters is essential for proactive environmental management and informed resource allocation. It also provides a scientific foundation for protecting aquatic ecosystems. Traditional prediction methodologies often struggle to capture seasonal fluctuations and long-term trends in water quality data. They also fail to effectively incorporate external factors such as precipitation. To address these challenges, this work proposes a novel model called Fusion Multimodal Patch Transformer (FMPT). This model improves water quality prediction accuracy through three innovative mechanisms: First, the model employs convolutional neural networks to process spatial precipitation patterns and Transformers to analyze temporal water quality data, enabling the extraction of spatio-temporal features. Second, FMPT applies patch-based decomposition to both modalities. This enhances the capture of local semantic information and expands the receptive field for more effective representation. Finally, it utilizes a low-rank multimodal fusion approach to integrate features from image and time series data, thereby obtaining a comprehensive representation of information from different modalities. Empirical validation was conducted using real-world hydrological datasets. The results show that the FMPT model performs better than existing state-of-the-art unimodal methods in water quality prediction tasks. It achieves an average improvement of 5.38% in prediction accuracy compared to other benchmark models.**

***Index Terms*—Water quality prediction, Multimodal fusion, Transformer, Spatiotemporal feature extraction, Low-rank tensor fusion**

I. INTRODUCTION

Water is one of the most precious natural resources on earth, and a healthy aquatic environment plays a crucial role in maintaining ecological balance, ensuring human health, and promoting sustainable socioeconomic development. However, the acceleration of industrialization and urbanization led to increasingly significant issues related to water pollution and quality. Therefore, establishing accurate and reliable water quality prediction models [1] to promptly monitor dynamic changes in the aquatic environment has significant practical implications.

Among the numerous factors influencing changes in water quality, precipitation is a critical element that cannot be overlooked [2]. Precipitation processes directly affect watershed runoff, hydraulic scouring, and the loading and transport-diffusion of pollutants. Furthermore, the increasing frequency and intensity of extreme precipitation events, driven by global climate change, have heightened the difficulty and uncertainty of predicting water quality. Consequently, effectively integrating precipitation information with water quality monitoring data and exploring their dynamic coupling mechanisms has become a focal point and challenge in contemporary water quality prediction research.

In an era characterized by the rapid proliferation of information and data, multimodal fusion [3] prediction emerges as a crucial approach for analyzing complex water environmental systems. This approach enhances prediction accuracy and stability through feature extraction, data fusion, and model integration. For instance, Zhang *et al.* [4] propose NowcastNet, an end-to-end neural network prediction model that leverages radar observation data to forecast rainfall over the next three hours. In water quality prediction, multimodal approaches can integrate precipitation remote sensing imagery with traditional water quality monitoring data to create a more comprehensive and multidimensional data support system. By integrating these heterogeneous information sources, the complex dynamic processes of the water environment can be characterized more comprehensively.

When collecting water quality data, multimodal datasets often exhibit various representations and scales of characteristics. Traditional fusion strategies typically use simple feature concatenation and linear transformations [5]. These methods often ignore the potential correlations among mul-

This work was supported by the National Natural Science Foundation of China under Grants 62473014 and 62173013, the Beijing Natural Science Foundation under Grants L233005 and 4232049. (Corresponding author: Junqi Zhang.)

2766-8665/25/$31.00 ©2025 IEEE
DOI 10.1109/ICNSC66229.2025.00034

timodal data. This oversight results in the loss of valuable information, reduces predictive accuracy, and limits the broader application of deep learning in water environment decision support. Currently, few methods can effectively align and adaptively fuse multimodal data [6]. Developing robust alignment and fusion techniques has become a critical research priority in modern water environment management. Integrating water quality data from diverse sources helps improve predictive precision. It also reduces heterogeneity caused by inconsistent data collection times and methods, achieving optimal matching across different feature spaces.

To solve these problems, this work proposes a Fusion Multimodal Patch Transformer, called FMPT for short, to forecast time series of water quality. The primary contributions of this work can be summarized as follows:

1) FMPT employs convolutional neural networks to extract spatial features from remote sensing images while utilizing Transformers to capture temporal features.
2) FMPT constructs patches for both spatial and temporal features, thus improving the capture of local semantic information and expanding the receptive field to improve feature extraction.
3) To effectively integrate data from various dimensions, FMPT employs the Low-Rank Multimodal Fusion (LMF) method.

The remainder of this work is organized as follows. Section II presents the details of the proposed FMPT, Section III discusses the experimental results, and Section IV summarizes and concludes the findings of this study.

II. Model Framework

This section introduces the FMPT model for predicting future water quality based on historical data. Fig. 1 illustrates the structure of the FMPT model. This model employs a foundational framework composed of spatial and temporal patches to concurrently process multimodal inputs, including time series data X and remote sensing image data X^R. For remote sensing imagery, spatial patches are inspired by the Vision Transformer [7]. This approach segments the 2D image into non-overlapping grid patches of fixed size. Each patch retains local spatial information and texture features, functioning as an individual visual token. Meanwhile, temporal patches focus on time series data by grouping multiple consecutive time steps into subsequence patches, thereby capturing local temporal dependencies [8]. This approach improves the recognition of time-dependent features such as seasonal variations, trends, and periodic events. Following patch processing, both image and time series patches undergo temporal embedding. After alignment and fusion, they are linearly projected to generate queries (Q), keys (K), and values (V), which are subsequently fed into a multi-head self-attention mechanism [9]. Finally, the output $\hat{Y}$ is produced through a feed-forward network [10].

A. Notations and Problem Statement

Let $X=\{x_t\}_{t=1}^{T}$ denote water quality time series, where $x_t \in \mathbb{R}^F$ is the F-dimensional observation at time t, and T is the sequence length. Rainfall remote sensing images are $X^R=\{x_t^R\}_{t=1}^{T}$, with $x_t^R \in \mathbb{R}^{H\times W\times C}$ representing the image at time t, where $H\times W$ is the resolution and C is the number of channels. Other notations are summarized in Table I.

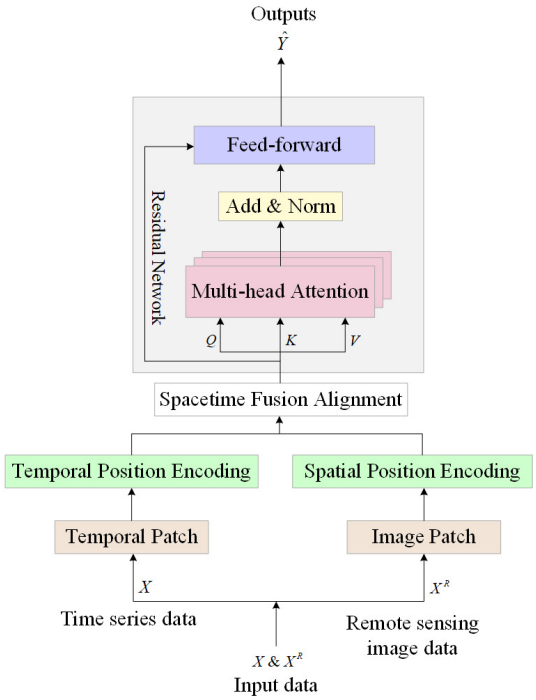


Fig. 1. The structure of FMPT.

TABLE I
Main Notations

Notation	Definition
$X\in\mathbb{R}^{T\times F}$	Indicator value of the water quality
X^R	Remote sensing image data
F	Number of features
H	Height of remote sensing image
W	Width of remote sensing image
T	Number of input time steps
P	Patch length
S	Stride length
N	Number of patches
τ	Size of the prediction step
D	Hidden size
L	Number of layers
ξ	Number of heads
$x_t \in \mathbb{R}^F$	Input value at time step t
$\hat{Y} \in \mathbb{R}^\tau$	Predicted values in next τ time steps

B. Spatial & Temporal Patching

1) Spatial Patching: To effectively align with temporal patches, the model divides remote sensing images into fixed-

size segments, linearly embeds each segment, and incorporates positional embeddings.

Since the Transformer accepts an input 1D sequence of token embeddings, FMPT reshapes the remote sensing image $x_t^R \in \mathbb{R}^{H\times W\times C}$ into a sequence of flattened 2D patches $\bar{x}_t^R \in \mathbb{R}^{P_s^2\times C}$, where P_s^2 indicates the resolution of each patch. The number of image patches N_s is expressed as:

$$N_s = \frac{HW}{P_s^2} \tag{1}$$

After flattening the patches, we apply a learnable linear projection to map them into D dimensions. The output of this projection is referred to as the patch embeddings. The trainable linear projection is defined as follows:

$$z_0 = \mathrm{MLP}[\bar{x}_{1,t}^R W_R; \bar{x}_{2,t}^R W_R; \ldots; \bar{x}_{N_s,t}^R W_R] + W_{pos} \tag{2}$$

where $W_R \in \mathbb{R}^{(P_s^2\cdot C)\times D}$ is a trainable projection matrix, and $W_{pos} \in \mathbb{R}^{N_s\times D}$ represents the relative position of the image patch in space.

2) Temporal patching: Following spatial patching, we segment the input time series X into overlapping patches to capture local temporal semantics. The number of patches N is calculated as follows:

$$N = \lfloor \frac{(T-P)}{S} \rfloor + 2 \tag{3}$$

where P represents the patch length and S denotes the stride between adjacent patches. Each patch is represented as $x_p \in \mathbb{R}^{P\times N\times F}$. We pad the sequence by replicating the terminal value x_T to ensure complete coverage. This process transforms the original sequence length T to $\frac{T}{S}$, thereby reducing the computational complexity of the Transformer and optimizing memory utilization during training. Furthermore, this approach expands the effective receptive field, enhances the capture of temporal dependencies, and ultimately improves prediction performance.

3) Embedding Patch Time Series: Conventional Transformer architectures typically encode time steps independently. They often fail to capture comprehensive temporal representations across the entire sequence. This limitation weakens the model's ability to extract essential patterns. It also restricts generalization across diverse temporal distributions. To address this deficiency, we implement learnable positional embeddings. These embeddings encode structural information from the patched and filtered sequences. The mathematical formulation of this embedding process is expressed as follows:

$$\tilde{x}_p = W_p \bar{x}_p + W_{pos} \tag{4}$$

where $\tilde{x}_p \in \mathbb{R}^{D\times N}$ represents the embedding result, and this embedding is then input into the Transformer. $W_{pos} \in \mathbb{R}^{D\times N}$ is a learnable position encoding. It captures the sequential ordering across patches. W_p is a trainable linear projection matrix. It transforms the input patches from their original feature dimension to the embedding dimension D.

4) Aligning of images and temporal patches: Aligning time series data with remote sensing images is essential. This integration enhances the richness of contextual information. The LMF method retains significant information after fusion. It also improves computational efficiency. This approach allows us to minimize the excessive parameterization of the models. It applies cross-modal attention to all modality-to-modality combinations for each modality. The fused tensor B is computed as follows:

$$B = v_1 \otimes v_2 \otimes \cdots \otimes v_Z = \bigotimes_{z=1}^{Z} v_z \tag{5}$$

where $\bigotimes_{z=1}^{Z}$ denotes the tensor outer product of a set of vectors indexed by z, and v_z represents the expanded one-dimensional tensor input. Multimodal fusion is expressed as:

$$f = g(B; U, c) = U \cdot B + c \quad f, c \in R^D \tag{6}$$

where U denotes the multimodal fusion weight and c represents the bias term. However, as the number of input modalities increases, the dimensionality and weight scale of the tensor can lead to model overfitting. To mitigate these issues, low-rank multimodal fusion is employed to decompose high-order weights into low-order weights, thereby reducing the computational complexity of the model. Utilizing LMF, the weight vector U is decomposed into Z groups of modality-specific factors, simplifying the calculations. The vector U consists of d_f Z-order tensors, $U_k \in \mathbb{R}^{d_1\times\ldots\times d_Z}$, $k=1,\ldots,d_f$, where d_1 and d_Z represent the feature dimensions of the first and Z-th modalities respectively, d_f denotes the dimension of the fusion output, and $Z=2$ represents the number of modalities (spatial and temporal). Each U_k is decomposed as follows:

$$U_k = \sum_{i=1}^{E} \bigotimes_{z=1}^{Z} \phi_{z,k}^{(i)} \quad \phi_{z,k}^{(i)} \in R^{d_z} \tag{7}$$

where E represents the smallest variable that can be decomposed, it defines the rank of the tensor. The term $\phi_{z,k}^{(i)}$ refers to the decomposition factor of the original tensor rank. By initializing the rank to e and parameterizing the model with e decomposition factors, the base rank U is constructed, forming a low-rank weight tensor.

Further, simplify by parameterizing the entire weight tensor U directly. By initializing a smaller rank e (where $e \leq E$) and parameterizing the model with these e decomposition factors, we construct a more efficient low-rank weight tensor:

$$U = \sum_{i=1}^{e} \bigotimes_{z=1}^{Z} \phi_z^{(i)} \tag{8}$$

Finally, the fusion vector f is redefined as follows:

$$f = \left(\sum_{i=1}^{e} \bigotimes_{z=1}^{Z} \phi_z^{(i)} \right) \cdot B + c \tag{9}$$

C. Transformer Architecture

After the multimodal fusion, the combined features $\tilde{X}$ are processed through a multi-head attention to capture various

aspects of relationships. The output is subsequently processed through a feed-forward network:

$$\text{FFN}(M)=\text{ReLU}(MW_1+b_1)W_2+b_2 \tag{10}$$

where M represents the output of the Transformer, W_1, W_2, b_1, and b_2 are the parameters of the network.

Residual connections with layer normalization [11] are implemented throughout the network to enhance gradient flow during training:

$$\bar{M} = \text{LayerNorm}(M + \text{FFN}(M)) \tag{11}$$

To capture both long-term trends and short-term fluctuations, the residual network output $\bar{M}$ is decomposed into trend and detrended components using linear transformations:

$$\bar{M}_\Gamma = W_\Gamma \bar{M} + b_\Gamma \tag{12}$$

$$\bar{M}_\gamma = W_\gamma \bar{M} + b_\gamma \tag{13}$$

where W_Γ, b_Γ, W_γ, and b_γ are learnable parameters for trend and detrended component extraction, respectively.

Ultimately, the model integrates the trend $\bar{M}_\Gamma$ and the detrended component $\bar{M}_\gamma$ for prediction:

$$\hat{Y}=W_y(\bar{M}_\Gamma+\bar{M}_\gamma)+b_y \tag{14}$$

where W_y and b_y serve as trainable parameters in this linear projection, this fusion mechanism enables the model to account for both long-term trends and short-term fluctuations simultaneously.

D. Training Methodology

To accurately quantify the discrepancy between predicted values $\hat{Y}$ and actual values Y, the Mean Squared Error (MSE) is employed as the loss function:

$$\mathcal{L}=\left\|Y-\hat{Y}\right\|_2^2 \tag{15}$$

For efficient parameter optimization and faster convergence, the Adam optimizer [12], [13] is utilized to minimize the loss function of the Transformer.

III. Experimental Evaluation

A. Experimental Setup

The FMPT model and other benchmark models are implemented in PyTorch. All experiments are conducted on a server with an RTX 3090 GPU and an Intel Xeon 6248R CPU. The experiments involve two types of data sources. The first data source includes hydrological data from nine national-level surface water automatic monitoring stations. These stations are located in the Haihe River Basin within the Beijing-Tianjin-Hebei region. Specifically, the models are tasked with forecasting dissolved oxygen (DO) levels at the Huairou Reservoir station and total nitrogen (TN) concentrations at the Shawo station. Detailed information is presented in Table II. The second source comprised high-precision remote sensing precipitation data. This data is provided by the National Aeronautics and Space Administration.

To construct a multimodal prediction framework, remote sensing data originally sampled every 30 minutes are resampled into a six-point mode, synchronized with the water quality monitoring times at 00:00, 04:00, 08:00, 12:00, 16:00, and 20:00. This process established a one-to-one mapping between the hydrological and remote sensing datasets. Both datasets cover the same period and are divided into training, validation, and test sets in a 7:1:2 ratio to facilitate the analysis of multi-source information fusion.

TABLE II
Statistics of the Datasets

Parameter	Dataset	
	Huairou	Shawo
Time span	Jun. 2019–Aug. 2023	Jun. 2019–Aug. 2023
Time interval	4 hours	4 hours
Data length	2,760	2,760
Water indicator	DO	TN

B. Evaluation Metrics

This work evaluates the performance of FMPT and other baseline models using three metrics: Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Mean Absolute Percentage Error (MAPE).

C. Hyperparameter Setting

FMPT uses a batch size of 32. Key hyperparameters, including input time steps (T), hidden size (D), number of layers (L), number of heads (ξ), and patch length (P), are selected from predefined candidate sets. The Adam optimizer is adopted with an initial learning rate of 0.01, which decays by 1×10^{-6} every 20 epochs.

To optimize model performance, we conduct extensive hyperparameter tuning across various dimensions. When τ=12, we compare the RMSE performance of FMPT for different values of D and T, as shown in Fig. 2 and Fig. 3. The optimal performance is achieved with a dimensionality of D=128 and a historical context window of T=64. Table V quantifies the relationship between model depth and prediction error, with L=4 providing the best balance between expressiveness and computational efficiency. This study evaluates various numbers of attention heads, with ξ=2 heads yielding superior performance, as demonstrated in Table IV. Finally, we optimize the extraction of temporal patterns through patch length experiments, as shown in Table VI, which indicates minimal error metrics at P=8.

D. Comparative Experimental

A comparative analysis against benchmark models demonstrates the superior efficacy of the FMPT. Table III presents the long-term water quality prediction results across 12, 36, and 64 time steps,where **bold** values indicate the best performance and underlined values represent the second-best performance. The FMPT model consistently outperforms baseline models on the Huairou DO dataset across all evaluation metrics. For Shawo TN datasets, FMPT is slightly lower than the optimal model in short-term prediction, but

TABLE III
COMPARISON OF FMPT AND OTHER BASELINE MODELS

Dataset	Models	12 Steps			36 Steps			64 Steps		
		RMSE	MAE	MAPE	RMSE	MAE	MAPE	RMSE	MAE	MAPE
Huairou DO	LSTM	2.7084	2.1488	21.4689	3.0465	2.4756	23.5523	2.5370	2.0646	20.2229
	Seq2Seq	3.2228	2.6146	25.6064	3.2475	2.6990	25.8344	3.3297	2.7991	26.2231
	Transformer	2.7335	2.4944	23.2397	3.2006	2.8257	26.5618	2.5370	2.0646	20.2229
	Informer	2.3906	2.1177	20.1764	2.3961	1.9839	19.4573	2.3259	1.8210	18.3352
	Autoformer	2.1531	1.8850	20.4328	2.2459	1.8161	19.3638	2.4629	1.9592	21.2614
	PatchTST	1.9014	1.5955	17.4529	1.9938	1.5592	17.0455	2.1806	1.6759	18.2617
	FMPT	**1.7736**	**1.5200**	**16.6280**	**1.9069**	**1.5075**	**16.3166**	**2.0132**	**1.5326**	**16.7507**
Shawo TN	LSTM	1.6057	1.2437	18.4595	1.8011	1.4688	22.8862	1.6260	1.3103	19.2860
	Seq2Seq	2.2713	1.8248	28.1051	1.9430	1.5891	24.4670	1.6592	1.2601	17.9217
	Transformer	1.3250	1.1938	17.6286	1.5412	1.3338	20.1397	1.6337	1.4013	21.3893
	Informer	1.3460	1.2175	17.4413	1.3556	**1.1311**	**15.5677**	1.6691	1.4359	21.9534
	Autoformer	1.8521	1.6711	23.2816	2.0013	1.7476	25.0074	1.9580	1.6651	23.5358
	PatchTST	**1.1270**	**0.9884**	**13.5492**	1.5101	1.2649	17.2871	1.6965	1.4069	19.2792
	FMPT	1.1722	1.0444	15.0554	**1.3457**	1.1375	16.3807	**1.3510**	**1.1095**	**16.1658**

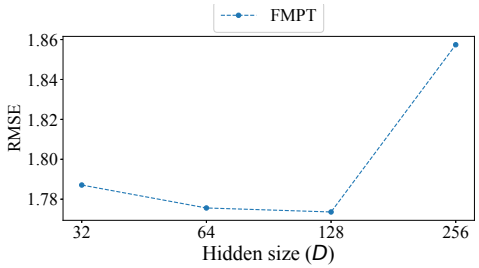


Fig. 2. Performance comparison concerning different D

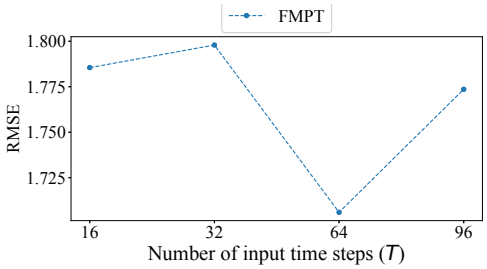


Fig. 3. Performance comparison concerning different T

TABLE IV
RMSE WITH DIFFERENT NUMBERS OF HEADS

ξ	RMSE
2	1.7151
4	1.7516
8	1.7736

TABLE V
RMSE WITH DIFFERENT NUMBERS OF LAYERS

L	RMSE
1	1.9241
2	1.7736
3	1.8894
4	1.7539

TABLE VI
RMSE WITH DIFFERENT PATCH LENGTH

P	RMSE
8	1.7350
16	1.7736
32	1.7663

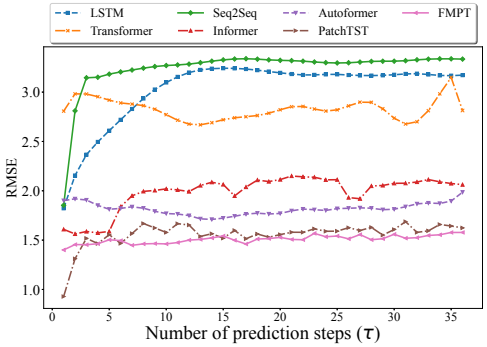
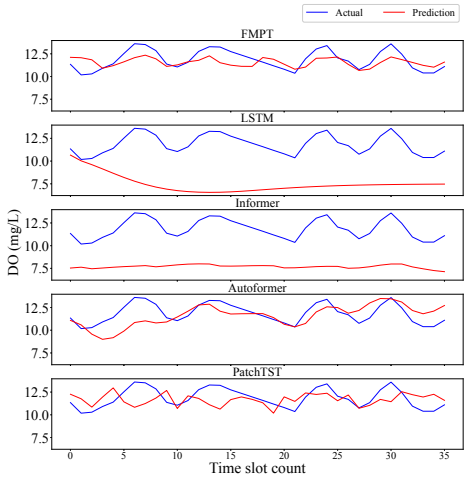


Fig. 4. Performance comparison with respect to different τ.

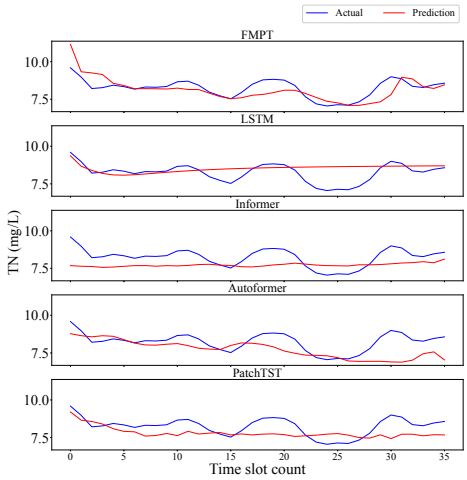
performs best in long-term prediction. As shown in Fig. 4, the prediction error increases with the forecast horizon for all models. Transformer-based architectures, including FMPT, Autoformer, and PatchTST, perform superior than other deep learning approaches. The enhanced accuracy of FMPT can be attributed to its multimodal fusion of remote sensing imagery and time series data.

Fig. 5 illustrates FMPT's multi-step predictions compared to baseline methodologies, demonstrating an enhanced ability to capture cyclical patterns and long-term trends in hydrological time series data. These results underscore FMPT's

accuracy in predicting water quality.



(a) Predicted curves from 1 to 36 steps for the Huairou DO.



(b) Predicted curves from 1 to 36 steps for the Shawo TN.

Fig. 5. Comparison of ground-truth ones and predicted values

IV. CONCLUSIONS

Accurate prediction of water quality parameters is crucial for sustainable environmental management and the protection of aquatic ecosystems. This work proposes a model called Fusion Multimodal Patch Transformer (FMPT). FMPT leverages convolutional neural networks to analyze spatial precipitation patterns and extract local spatial correlations. It also employs Transformers to capture long-term dependencies in temporal water quality data. This combination enables a comprehensive understanding of spatio-temporal dynamics. Additionally, a patch-based decomposition mechanism segments complex temporal data into local fragments. This enhances the model's sensitivity to local features and reduces information loss caused by global smoothing. The Low-Rank Multimodal Fusion method efficiently integrates multiple data modalities. It preserves key associative information and further enhances the robustness and accuracy of predictions. Experimental results show that the FMPT model outperforms traditional methods on real-world hydrological datasets. It performs especially well in long-term predictions, achieving an average improvement of 5.38% in prediction accuracy compared to other benchmark models.

In the future, FMPT will focus on enhancing its adaptability to a wider range of water quality parameters. It will also explore the integration of additional external environmental variables and large models [14] to further improve the comprehensiveness and reliability of its predictions.

REFERENCES

[1] J. Bi, X. Wu, H. Yuan, Z. Wang, D. Wei, R. Wu, J. Zhang, J. Qiao and R. Buyya, "STMF: A Spatio-Temporal Multimodal Fusion Model for Long-term Water Quality Forecasting," *IEEE Internet of Things Journal*, pp. 1–1, Jun. 2025.

[2] D. Kim, Y. O. Lee, C. Jun and S. Kang, "Understanding the Way Machines Simulate Hydrological Processes—A Case Study of Predicting Fine-Scale Watershed Response on a Distributed Framework," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–18, Jun. 2023.

[3] J. Bi, Z. Wang, H. Yuan, X. Wu, R. Wu, J. Zhang and M. Zhou, "Long-Term Water Quality Prediction With Transformer-Based Spatial-Temporal Graph Fusion," *IEEE Transactions on Automation Science and Engineering*, vol. 22, pp. 11392–11404, Jan. 2025.

[4] Y. Zhang, M. Long, K. Chen, L. Xing, R. Jin, M. Jordan, and J. Wang, "Skilful Nowcasting of Extreme Precipitation with Nowcastnet," *Nature*, vol. 619, no. 7970, pp. 526–532, Jul. 2023.

[5] H. Cheng, Z. Yang, X. Zhang and Y. Yang, "Multimodal Sentiment Analysis Based on Attentional Temporal Convolutional Network and Multi-Layer Feature Fusion," *IEEE Transactions on Affective Computing*, vol. 14, no. 4, pp. 3149-3163, Apr. 2023.

[6] L. Zhang, Z. Liu, X. Zhu, Z. Song, X. Yang, Z. Lei, and H. Qiao, "Weakly Aligned Feature Fusion for Multimodal Object Detection," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 36, no. 3, pp. 4145–4159, Aug. 2025.

[7] J. Bi, Y. Li, H. Yuan, M. Wang, Z. Wang, J. Zhang and M. Zhou, "Hybrid Water Quality Prediction With Multimodal Low-Rank Fusion and Localized Attention," *IEEE Internet of Things Journal*, vol. 12, no. 12, pp. 21158–21169, Mar. 2025.

[8] P. Lin, X. Zhang, L. Gong, J. Lin, J. Zhang, and S. Cheng, "Multi-timescale short-term urban water demand forecasting based on an improved PatchTST model," *Journal of Hydrology*, vol. 651, pp. 132599, Apr. 2025.

[9] Y. Zhou, F. Wang, J. Zhao, R. Yao, S. Chen and H. Ma, "Spatial-Temporal Based Multihead Self-Attention for Remote Sensing Image Change Detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 10, pp. 6615-6626, Oct. 2022.

[10] Z. Xu, Y. Liu, S. Qin and Z. Ming, "Output Range Analysis for Feed-Forward Deep Neural Networks via Linear Programming," *IEEE Transactions on Reliability*, vol. 72, no. 3, pp. 1191-1205, Sept. 2023.

[11] L. Huang, J. Qin, Y. Zhou, F. Zhu, L. Liu and L. Shao, "Normalization Techniques in Training DNNs: Methodology, Analysis and Application," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 8, pp. 10173-10196, Aug. 2023.

[12] M. Akhtar, M. Tanveer and M. Arshad, "RoBoSS: A Robust, Bounded, Sparse, and Smooth Loss Function for Supervised Learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 47, no. 1, pp. 149-160, Jan. 2025.

[13] Y. Zhang, C. Chen, T. Ding, Z. Li, R. Sun, Z.-Q. Luo, "Why Transformers Need Adam: A hessian perspective," *Advances in Neural Information Processing Systems*, vol. 37, pp. 131786–131823, Oct. 2024.

[14] J. Bi, Z. Wang, H. Yuan, X. Shi, Z. Wang, J. Zhang, "Large AI Models and Their Applications: Classification, Limitations, and Potential Solutions," *Software: Practice and Experience*, vol. 55, pp. 1003–1017. Jun. 2025.