2024 IEEE 20th International Conference on Automation
Science and Engineering (CASE)
August 28 - September 1, 2024. Bari, Italy

2024 IEEE 20th International Conference on Automation Science and Engineering (CASE) | 979-8-3503-5851-3/24/$31.00 ©2024 IEEE | DOI: 10.1109/CASE59546.2024.10711632

WaterTS: Integrating Enhanced Transformer, Sliding Block, and Channel Independence for Long-term Water Quality Prediction

Jing Bi[1], Lifeng Xu[1], Ziqi Wang[1], Haitao Yuan[2], Shichao Chen[3], Mu Gu[4] and MengChu Zhou[5]

***Abstract*—Nowadays, the deterioration of water resources leads to negative ecological impacts. To effectively inhibit the deterioration of water resources, a water quality prediction model based on enhanced transformer, sliding block, and channel independence (WaterTS) is proposed by comprehensively analyzing the indicators of water resources and making long-term predictions of the dissolved oxygen index. WaterTS adopts a sliding block method to extract the short-term temporal features of the water quality series and combine them with channel independence to make independent predictions of multi-featured data. Moreover, it upgrades the internal encoder structure of the transformer and improves the attention mechanism to Probsparse-attention and Auto-Correlation to speed up the prediction speed. Furthermore, Post LayerNormal is adjusted to Pre LayerNormal, which makes the training gradient more stable and enhances the accuracy of predictions. Experiments are conducted using real-world water environment data, and comparison results with state-of-the-art prediction models show that the WaterTS achieves accurate predictions on both short-term and long-term water quality data.**

***Index Terms*—Water quality prediction, channel independence, sliding block, Pre LayerNormal.**

I. INTRODUCTION

Water stands as one of the most precious resources on Earth, playing a pivotal role in human survival and development. However, the water environment confronts substantial challenges due to climate change, burgeoning population, and escalating industrialization. The pollution of water bodies originates from diverse sources, including discharges from chemical plants, the application of fertilizers, wastewater from urban drainage systems, and illicit dumping near rivers and lakes. These contaminants comprise hazardous substances, *e.g.*, heavy metals, organic pollutants, and bacteria, posing severe threats to aquatic life, food supply chains, and human health. To effectively tackle these challenges, it is imperative to accurately predict the water environment in the future for timely responses.

The relentless advancement of Internet of Things (IoT) technology has given us an extensive reservoir of data for investigating water pollution. To collect relevant water quality data, automatic monitoring stations are established within pertinent watersheds, systematically collecting continuous water quality data at regular intervals. This dataset serves as the foundation for subsequent research on water quality prediction. Moreover, there are many methods for time series prediction, but fewer methods are applied in the field of water environment. Since water quality data are also time series data, we aim to design a time series prediction method to conduct research in the field of water quality prediction.

Traditional time series prediction methods, *e.g.*, autoregressive integrated moving average [1], autoregressive moving average [2], and seasonal autoregressive integrated moving average [3] are pertinent and robust for time-series prediction. However, they are less generalized and cannot be well applied to complex problems. Machine learning-based methods, *e.g.*, light gradient boosting machine [4] and extreme gradient boosting [5] have better predictive performance, but they require some feature engineering processing, which is time-consuming and complex. With the development of deep learning, some prediction methods such as long short-term memory [6], gated recurrent unit [7], sequence to sequence [8], wavenet [9], 1D-convolutional neural Network [10] are highly generalizable and can be well applied to water quality prediction. Furthermore, Transformer [11] and the derived long-term prediction models such as Informer [12], Autoformer [13], FEDformer [14] have achieved great success in time-series prediction. Therefore, this paper is based on Transformer and it is further improved and optimized to enhance its effectiveness in long-term water quality prediction.

Based on the aforementioned analysis, this paper proposes a water quality prediction model based on enhanced transformer, sliding block, and channel independence (WaterTS). Water quality data is a multivariate time series data containing multiple feature information. Based on the input characteristics of the Transformer [15], it can accept both single-channel and multi-channel data. For Transformer, the multivariate data are mixed and processed to map the multivariate information to a uniform embedding dimension. However, it leads to lower prediction performance of the model. Therefore, we aim to disassemble the multi-channel and use each channel as the input information for the Transformer, and the

This work was supported by Beijing Natural Science Foundation (L233005), National Natural Science Foundation of China (NSFC) under Grants 62173013 and 62073005, and the Key Research and Development Program of Rizhao (2023ZDYF010153).

[1]J. Bi, L. Xu and Z. Wang are with the Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China. bijing@bjut.edu.cn

[2]H. Yuan is with the School of Automation Science and Electrical Engineering, Beihang University, Beijing 100191, China. yuan@buaa.edu.cn

[3]S. Chen is with the State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences and Beijing Huairou Academy of Parallel Sensing, Beijing 101408, China. shichao.chen@ia.ac.cn

[4]M. Gu is with the Beijing aerospace smart manufacturing technology development Co., Ltd, Beijing 100045, China. gumu2005@163.com

[5]M. Zhou is with the Department of Electrical and Computer Engineering, New Jersey Institute of Technology, Newark, NJ 07102, USA. zhou@njit.edu

multi-channel data shares the same Transformer parameters. Moreover, time series prediction uses each point individually as a prediction step. However, this does not work well for long-term series prediction, as individual points represent limited information. Thus, accurate predictions require a significant amount of prior information. In that case, we splice the data within a sliding window into localized data blocks and arrange them in a time sequence. These data blocks are then input to the Transformer to replace the single-point sequence. Furthermore, the Attention mechanism has a significant impact on the model's predictive effectiveness. The complexity of the traditional Attention mechanism is $O(N^2)$, and the complexity of improved Probsparse [16] and Auto-Correlation [17] are both $O(NLogN)$. Therefore, replacing the original attention mechanism can reduce the overall complexity of the model. Finally, referring to the network structure of the large language model, it is proved that with the increase of encoder layers, the gradient of the model is relatively smooth when using Pre LayerNormal pattern [18] to train it, and the effect is better than that of Post LayerNormal [19]. Therefore, WaterTS is trained under the Pre LayerNormal pattern.

II. PROPOSED FRAMEWORK

This section first introduces each component in the WaterTS and then gives the overall architecture of it.

A. Channel Independence

WaterTS adopts channel independence to enhance its prediction accuracy. Multivariate data for the water environment is considered in this paper. Given a set of multivariate water quality samples $X{=}(x_1, \cdots, x_L)$ with a sliding window of L where each x_i contains K dimensional data. The prediction model needs to predict the future M data points, *i.e.*, $(x_{L+1}, \cdots, x_{L+M})$. The univariate water quality data from 1 to L are represented by $x_{1:L}^{(i)}{=}(x_1^i, \cdots, x_L^i)$ where $i{\in}[1, 2, \cdots, K]$. All feature sequences are split into individual feature sequences and then they are delivered to the Transformer by channel independence. The model gives the prediction result in $\hat{x}_{1:L}^{(i)}{=}(\hat{x}_1^i, \cdots, \hat{x}_L^i)$ for each feature sequence, and then the individual feature prediction sequences are merged and restored to the original form of prediction sequence.

B. Sliding Block

In traditional time series forecasting, the information is obtained from a single time point, which is then combined with the past and future time points to make a long-term prediction [20]. However, this method ignores the information around that time point. In this case, a sliding-block [21] is adopted to divide the input series into multiple blocks. They are shorter sequences containing the original sequence with some local sequence information, and the number of blocks depends on the length of the window (W) and the stride length (S). The number of blocks N is calculated as: $N{=}\left\lfloor \frac{(L-W)}{S} \right\rfloor +2$. The segmentation process is shown in Fig. 1. Specifically, it is assumed that there is a sequence of 18-time steps and the length of the sliding window is set to 6. Moreover, the stride length is also set to 6. Therefore, the sequence is divided into three blocks.

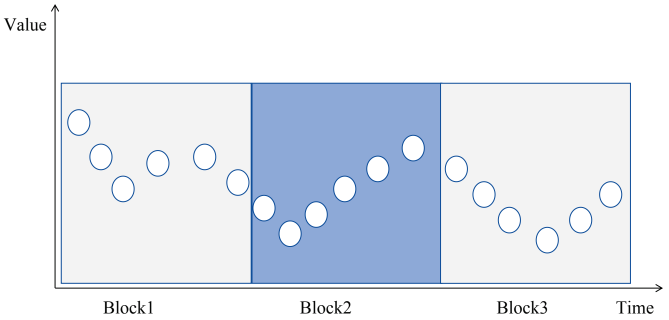

Fig. 1: Working principle of a sliding block.

C. ProbSparse-attention

The ProbSparse-attention mechanism is adopted in WaterTS. It can not only maintain the predicted performance of the model but also reduce its computational complexity. The principle of ProbSparse-attention ($\mathcal{A}(\cdot)$) is shown in (1).

$$\mathcal{A}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \operatorname{Softmax}\left(\frac{\overline{\mathbf{Q}}\mathbf{K}^{\top}}{\sqrt{d}}\right)\mathbf{V} \tag{1}$$

where $\mathbf{Q}$, $\mathbf{K}$ and $\mathbf{V}$ denote query, key and value vectors, respectively. $\overline{\mathbf{Q}}$ is a coefficient matrix and contains only the top k query vectors, all query vectors are obtained by calculating the sparsity, and the computation with the key takes the first k query vectors. d denotes the scaling factor and it is the first dimension of $\mathbf{K}$. $Softmax(\cdot)$ denotes the normalization process. Moreover, the algorithmic complexity is reduced to $O(NLogN)$.

D. Auto-Correlation

WaterTS adopts Auto-Correlation [22] to realize efficient sequence-level connections. Specifically, it first calculates the correlation between the original sequence and the lagging sequence. Then, it finds the subsequence with similar periodicity and selects the most likely top k cycle lengths. Finally, it performs the attention score calculation. In addition, compared to the traditional attention mechanism, it reduces the time complexity to $O(NLogN)$. The principle of Auto-Correlation (Auto-Correlation($\cdot$)) is shown as follow:

$$\tau_1, \cdots, \tau_k = \underset{\tau \in \{1, \cdots, L\}}{\arg \operatorname{Topk}} \left(\mathcal{R}_{\mathcal{Q},\mathcal{K}}(\tau)\right) \tag{2}$$

$$\begin{aligned} &\widehat{\mathcal{R}}_{\mathcal{Q},\mathcal{K}}(\tau_1), \cdots, \widehat{\mathcal{R}}_{\mathcal{Q},\mathcal{K}}(\tau_k) = \\ &\operatorname{Softmax}\left(\mathcal{R}_{\mathcal{Q},\mathcal{K}}(\tau_1), \cdots, \mathcal{R}_{\mathcal{Q},\mathcal{K}}(\tau_k)\right) \end{aligned} \tag{3}$$

$$\text{Auto-Correlation}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \sum_{i=1}^{k} \operatorname{Roll}(\mathcal{V}, \tau_i)\, \widehat{\mathcal{R}}_{\mathcal{Q},\mathcal{K}}(\tau_i) \tag{4}$$

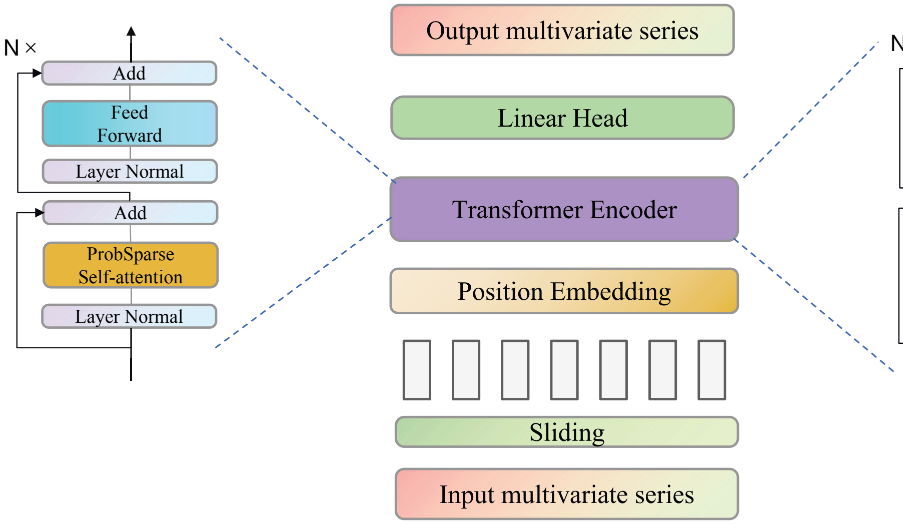

Fig. 2: Network structure of the WaterTS.

where argTopk$(\cdot)$ shows the top k independent variables of the autocorrelation, τ_k denotes the delay k. $\mathcal{R}_{\mathcal{Q},\mathcal{K}}$ represents the autocorrelation of $\mathbf{Q}$ and $\mathbf{K}$. $\widehat{\mathcal{R}}_{\mathcal{Q},\mathcal{K}}\left(\tau_i\right)$ denotes the result of τ_i after the softmax process. $\operatorname{Roll}\left(\mathcal{V},\tau_i\right)$ represents the effect of the time delay τ_i on $\mathcal{V}$.

E. Pre LayerNormal

The general structure of the Transformer is Post LayerNormal, which performs layer normalization (LN) after the residuals, but we aim to put the LN before the residuals, which is Pre LayerNormal. The reason for this is that when the number of layers of the Transformer is increased, the backpropagation can avoid the gradient explosion and disappearance of the live gradient. Therefore, the Pre LayerNormal is better for a larger number of layers.

F. Architecture of the WaterTS

The WaterTS consists of three main components, including channel independence, sliding block, and Transformer. The multivariate data is disassembled using channel independence to make separate predictions for individual dimensions and finally integrated. Moreover, the sliding block makes the model focus on local information and reduces its computational complexity. Furthermore, the internal structure of the Transformer is changed from Post LayerNormal to Pre LayerNormal to accommodate a large number of layers, which makes model training more stable. Finally, the traditional multi-head attention is adjusted to Probsparse-attention or Auto-Correlation, which is more adapted to time series prediction tasks.

The structure of the WaterTS is shown in Fig. 2. Specifically, the multivariate data is first divided into features and used as the input. Then, each feature is split independently to form multiple single-feature time series data. Next, the multiple single-feature time series data are subjected to sliding block processing, which divides them into multiple consecutive time blocks containing local features. The segmented time blocks are connected to serve as the input data for the Transformer. It is worth noting that before inputting into the Transformer, positional encoding is performed to maintain the order between sliding blocks. Then, the data enters into the encoder part of the Transformer and undergoes Pre LayerNormal, which normalizes the data distribution. After that, it undergoes ProbSparse-attention or Auto-Correlation to regulate the weight distribution. In addition, the specific kind of attention to use needs to be chosen rationally based on different datasets. Therefore, for WaterTS, the one that adopts the Auto-Correlation is named Auto_WaterTS, while the one with the ProbSparse-attention is named PS_WaterTS. After that, the residuals are connected and after Pre LayerNormal and feed-forward neural network, the residuals are connected again, *i.e.*, a complete encoder block. It is important to note that the model needs to be stacked with N such blocks. Finally, after the multi-layer Transformer Encoder, the data is mapped through a linear head to get the final multivariate prediction result.

III. Experimental Results and Discussion

This section presents the experimental results of the WaterTS and the comparative models. All models are trained and validated on the RTX 4090 server.

A. Dataset and Evaluation Metrics

WaterTS is evaluated on two high-quality water quality datasets, *i.e.*, the water body data released by the U.S. Geological Survey (USGS) for the state of California from May 2012 to August 2020, and the water body data collected in the Beijing-Tianjin-Hebei region within China (China_Water) from August 2018 to December 2021. It is worth noting that the dissolved oxygen (DO) indicator in water is one of the most important indicators reflecting the quality of the water

TABLE I: Predicted results of different models with different datasets

Models		PS_WaterTS		Auto_WaterTS		Autoformer		Informer		Transformer	
Metric		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
USGS	1	0.0020	0.0302	**0.0015**	**0.0268**	0.0443	0.1550	0.0480	0.1011	0.0212	0.0894
	4	0.0023	0.0327	**0.0020**	**0.0300**	0.0331	0.1263	0.0428	0.0841	0.0247	0.1128
	8	0.0045	0.0449	**0.0037**	**0.0399**	0.0378	0.1350	0.2420	0.1400	0.0269	0.1214
	16	**0.0069**	**0.0520**	0.0081	0.0566	0.0553	0.1716	0.0362	0.1314	0.0300	0.1198
	96	**0.0369**	0.1231	0.0371	**0.1226**	0.0732	0.1926	0.0895	0.2024	0.0718	0.1953
	192	**0.0624**	**0.1626**	0.0639	0.1641	0.1428	0.2689	0.1739	0.2948	0.1176	0.2366
	384	**0.1139**	**0.2219**	0.1423	0.2561	0.2617	0.3582	0.4448	0.4961	0.3190	0.3995
	672	**0.1921**	**0.2881**	0.2043	0.3010	0.4448	0.4904	0.4992	0.5298	0.4301	0.4675
China_Water	1	0.1986	0.2599	**0.1981**	**0.2577**	0.4431	0.4910	0.4205	0.4939	0.2590	0.3259
	2	**0.2472**	0.2898	0.02479	**0.2892**	0.5787	0.5503	0.4587	0.4939	0.4292	0.4790
	3	0.2794	**0.3142**	**0.2789**	0.3144	0.6005	0.5632	0.4620	0.4655	0.4385	0.4733
	4	0.3041	0.3350	**0.3033**	**0.3318**	0.6210	0.5724	0.4874	0.4892	0.5190	0.5418
	6	**0.3411**	**0.3564**	0.3414	0.3584	0.6469	0.5860	0.4855	0.4828	0.4975	0.5064
	18	**0.4888**	**0.4612**	0.4841	0.4566	0.6670	0.5965	0.5837	0.5388	0.8632	0.7173
	30	0.6258	0.5492	**0.5887**	**0.5271**	0.7272	0.6308	0.6379	0.5998	1.4194	0.9372
	42	**0.6340**	**0.5544**	0.6478	0.5621	0.7485	0.6404	0.8182	0.6973	1.0221	0.8139

body, which is predicted by our experiments. Moreover, to verify the prediction ability of the WaterTS, the error between the predicted value and the real value is calculated by using two error evaluation indexes, including mean squared error (MSE) and mean absolute error (MAE). They complement each other, MSE is easy to compute, and MAE has better robustness to anomalies.

B. Parameter Tuning

The parameters affecting the effectiveness of the WaterTS mainly contain three elements including the length of the sliding block, the number of encoder layers, and the number of attention heads. It is worth noting that Auto_WaterTS and PS_WaterTS only have different attention layers, all other structures are the same. As a result, the trends and optimal points in the parameter tuning of the two models are consistent. Therefore, the average value [23] is taken to show the tuning results.

Firstly, we research the influence of the length of the sliding block on the prediction effect, we choose $W\in\{4,8,16,32,64\}$, and choose the length of fixed stride to ensure the fairness of the experiment. The experimental result is shown in Fig. 4. It is shown that the best effect is achieved when W is set as 32. Subsequently, we examine the impact of varying the number of layers in the encoder (E) on prediction performance. As the number of layers in the encoder increases, the data tends to converge towards the true value. We select $E\in\{2,4,8,16,32\}$ for our analysis. The results are depicted in Fig. 5, revealing that the optimal number of layers is set to 4. Finally, we investigate the impact of varying the number of attention heads. A greater amount of attention heads widens the scope of observation, leading to increased synthesis of information and improved prediction accuracy. A denotes the number of attention heads and we explore $A\in\{2,4,8,16,32\}$. The experimental results are illustrated in Fig. 6. It is shown that the optimal effect is achieved when A is taken as 4. It is worth noting that "more is better" does not necessarily hold, and the number of attention heads needs to be balanced.

C. Comparison Experiments

The long-term time series prediction models that have been studied in recent years are selected for comparison, *i.e.*, Transformer, Informer, and Autoformer. They are compared with our PS_WaterTS and Auto_WaterTS in the two types of water quality datasets, including short-term and long-term prediction. Moreover, the evaluated metric functions are MSE and MAE.

The comparison results are shown in Table I. The best results are shown in **bold**, while the second is *underlined*. For the USGS dataset, we select a range of $T\in\{1,4,8,16,96,192,384,672\}$, where each step represents 15 minutes, so the corresponding prediction ranges are 15min, 1 hour, 2 hours, 4 hours, 1 day, 2 days, 4 days, and 8 days. For the China_Water dataset, we select a range of $T'\in\{1,2,3,4,6,18,30,42\}$, where each step represents 4 hours, so the corresponding prediction ranges are 4 hours, 8 hours, 12 hours, 16 hours, 1 day, 3 days, 5 days, and 7 days. The prediction effect of USGS dataset is shown in Fig. 3, and the prediction effect of China_Water dataset is shown in Fig. 7.

TABLE II: Training time of different models

Model	Epoch training time (s)	Total training time (min)
PS_WaterTS	**11**	**9.17**
Auto_WaterTS	**20**	**16.67**
Transformer	30	25.00
Autoformer	110	91.67
Informer	251	209.17

It is shown in Figs. 3 and 7 that the prediction results of Auto_WaterTS (blue curve) and PS_WaterTS (purple curve) are closer to the real data (green curve) compared to benchmark models on both datasets. In addition, it is illustrated in Table I that Auto_WaterTS and PS_WaterTS achieve the top two results on different datasets with different step sizes, proving that WaterTS has excellent predictive ability and high robustness. Furthermore, the training time of each model is shown in Table II. It shows the time consumed by different models in one epoch and total training for a prediction step

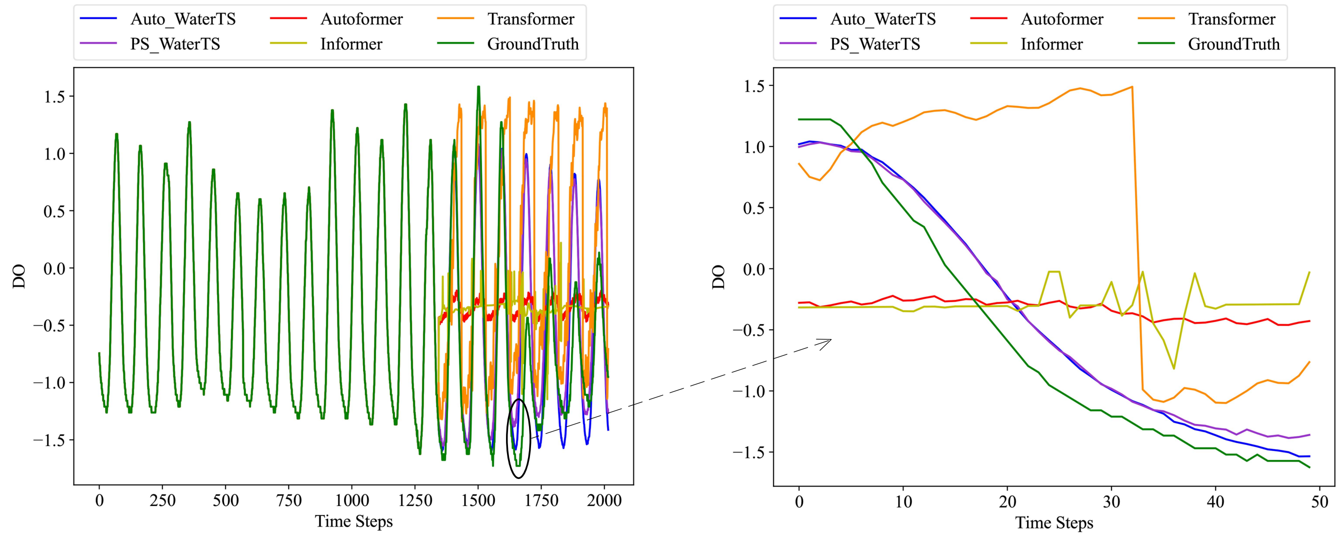

Fig. 3: Predicted results for a step size of 672 on the USGS dataset.

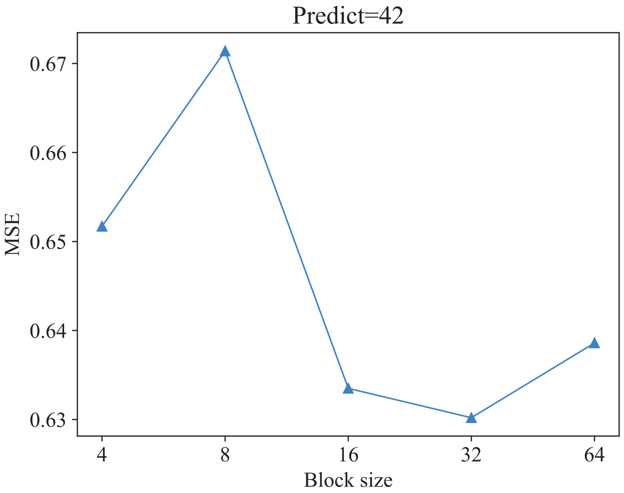

Fig. 4: MSE of WaterTS under different W.

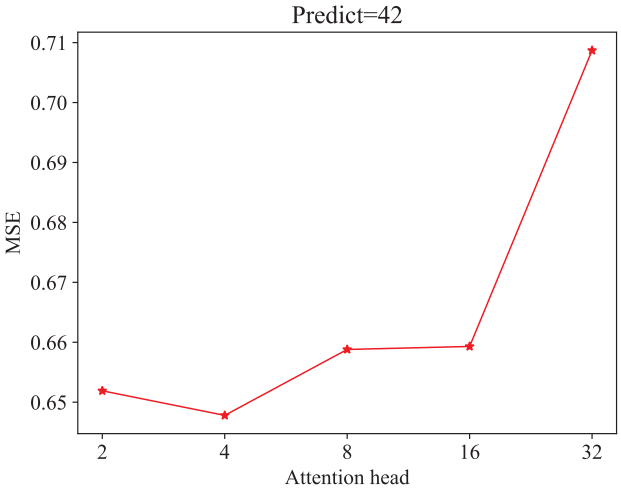

Fig. 6: MSE of WaterTS under different A.

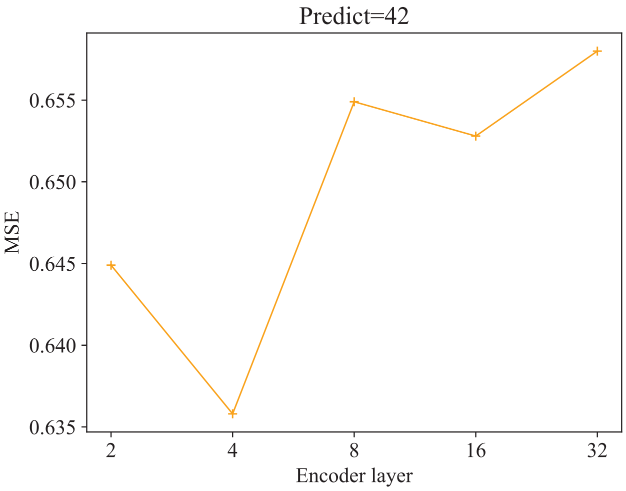

Fig. 5: MSE of WaterTS under the different E.

of 42 as an example. It can be seen that WaterTS consumes the shortest training time.

IV. CONCLUSIONS

At present, the global water resources environment is gradually deteriorating. It is affecting people's health and social development. Thus, real-time prediction of water quality data is needed for timely response. Therefore, we propose a water quality prediction model based on enhanced transformer, sliding block, and channel independence (WaterTS) in this work. It combines sliding block and channel independence for independent prediction of the multi-featured data. Moreover, it adopts Probsparse-attention and Auto-Correlation to improve the predictive speed of the model, allowing it to respond more quickly to changes in water quality. In addition, the original Transformer structure is improved to replace the Post LayerNormal with the Pre LayerNormal

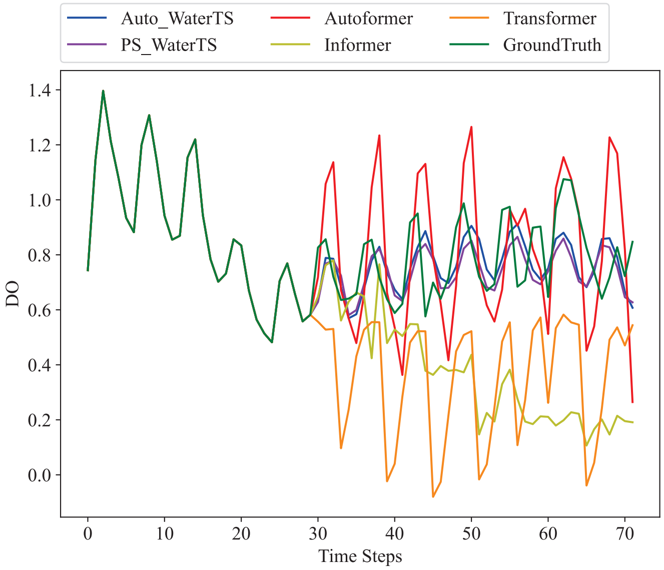

Fig. 7: Predicted results for a step size of 42 on the China_Water dataset.

to enhance the predictive accuracy. The proposed WaterTS is validated against two real-world water quality datasets, and the results verify that each component in WaterTS is effective. In addition, its predictive performance on these datasets is superior to the compared models. In future work, we will further investigate the integration of block thinking and channel independence prediction with advanced deep learning methods on time series prediction. We also intend to incorporate optimization algorithms [24], [25] to choose the parameters of the model, avoiding manual selection errors.

REFERENCES

[1] M. Güvercin, N. Ferhatosmanoglu and B. Gedik, "Forecasting Flight Delays Using Clustered Models Based on Airport Networks," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 5, pp. 3179–3189, May 2021.

[2] L. Falconi, A. Ferrante and M. Zorzi, "A Robust Approach to ARMA Factor Modeling," *IEEE Transactions on Automatic Control*, vol. 69, no. 2, pp. 828–841, Feb. 2024.

[3] E. Rokhsatyazdi, S. Rahnamayan, H. Amirinia and S. Ahmed, "Optimizing LSTM Based Network For Forecasting Stock Market," *2020 IEEE Congress on Evolutionary Computation (CEC)*, 2020, Glasgow, UK, pp. 1–7.

[4] Z. Zhou, M. Wang, J. Huang, S. Lin and Z. Lv, "Blockchain in Big Data Security for Intelligent Transportation With 6G," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 7, pp. 9736–9746, Jul. 2022.

[5] J. Xie, Z. Li, Z. Zhou and S. Liu, "A Novel Bearing Fault Classification Method Based on XGBoost: The Fusion of Deep Learning-Based Features and Empirical Features," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–9, Dec. 2021.

[6] J. Bi, L. Zhang, H. Yuan and J. Zhang, "Multi-indicator Water Quality Prediction with Attention-assisted Bidirectional LSTM and Encoder-Decoder," *Information Sciences*, vol. 625, pp. 54–80, May 2023.

[7] Y. Li, S. Wang, Y. Wei and Q. Zhu, "A New Hybrid VMD-ICSS-BiGRU Approach for Gold Futures Price Forecasting and Algorithmic Trading," *IEEE Transactions on Computational Social Systems*, vol. 8, no. 6, pp. 1357–1368, Dec. 2021.

[8] Z. Wang, X. Su and Z. Ding, "Long-Term Traffic Prediction Based on LSTM Encoder-Decoder Architecture," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 10, pp. 6561–6571, Oct. 2021.

[9] C. Lyu, B. Yang, J. Tian, J. Jin, C. Ge and J. Yang, "Three-Fingers FBG Tactile Sensing System Based on Squeeze-and-Excitation LSTM for Object Classification," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–11, Jun. 2022.

[10] M. Sharma and T. Maity, "Smart and Fault-Tolerant Multisensor Fusion Model for UCM Methane Hazard Monitoring Based on Belief Divergence Backed DS Filter and Hybrid CNN-LSTM Classifier," *IEEE Internet of Things Journal*, vol. 11, no. 2, pp. 3264–3273, Jan. 2024.

[11] J. Lin, M. Rickert and A. Knoll, "LieGrasPFormer: Point Transformer-Based 6-DOF Grasp Detection with Lie Algebra Grasp Representation," *2023 IEEE 19th International Conference on Automation Science and Engineering (CASE)*, 2023, Auckland, New Zealand, pp. 1–7.

[12] Y. Zhang, C. Li, Y. Tang, F. Zhou and X. Zhang, "Fault Trend Prediction of Centrifugal Blowers Considering Incomplete Data," *2023 IEEE 19th International Conference on Automation Science and Engineering (CASE)*, 2023, Auckland, New Zealand, pp. 1–6.

[13] 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 Transactions on Industrial Informatics*, vol. 19, no. 12, pp. 11584–11595, Dec. 2023.

[14] A. S. Editya, T. Ahmad and H. Studiawan, "Forensic Investigation of Drone Malfunctions with Transformer," *2023 International Conference on Smart Systems for applications in Electrical Sciences (ICSSES)*, 2023, Tumakuru, India, pp. 1–5.

[15] Y. Cao, K. Ngo and D. Dong, "A Scalable Electronic-Embedded Transformer, a New Concept Toward Ultra-High-Frequency High-Power Transformer in DC–DC Converters," *IEEE Transactions on Power Electronics*, vol. 38, no. 8, pp. 9278–9293, Aug. 2023.

[16] Y. Jiang, Y. Dai, R. Si, J. Chen, T. Gao, and J. Zhang, "Short-Term State Electricity Load Forecasting Based on Transfer-Informer," *2022 IEEE 2nd International Conference on Digital Twins and Parallel Intelligence (DTPI)*, 2022, Boston, USA, pp. 1–6.

[17] D. A. Hague, "Adaptive Transmit Waveform Design Using Multitone Sinusoidal Frequency Modulation," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 57, no. 2, pp. 1274–1287, Apr. 2021.

[18] S. Jeong, M. Seo, X. T. Nguyen and H. -J. Lee, "A Low-Latency and Lightweight FPGA-Based Engine for Softmax and Layer Normalization Acceleration," *2023 IEEE International Conference on Consumer Electronics-Asia (ICCE-Asia)*, 2023, Busan, Korea, 2023, pp. 1–3.

[19] 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.

[20] H. Novak, F. Bronić, A. Kolak and V. Lesic, "Data-Driven Modeling of Urban Traffic Travel Times for Short- and Long-Term Forecasting," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 10, pp. 11198–11209, Oct. 2023.

[21] K. Kim, S. Moon, J. Han, E. Alon and A. M. Niknejad, "Precursor ISI Cancellation Sliding-Block DFE for High-Speed Wireline Receivers," *IEEE Transactions on Circuits and Systems*, vol. 70, no. 10, pp. 4169–4182, Oct. 2023.

[22] M. Chen, H. Peng, J. Fu and H. Ling, "AutoFormer: Searching Transformers for Visual Recognition," *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, Montreal, Canada, pp. 12250–12260.

[23] J. Lin, F. Gao, X. Shi, J. Dong and Q. Du, "SS-MAE: Spatial–Spectral Masked Autoencoder for Multisource Remote Sensing Image Classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–14, Nov. 2023.

[24] J. Bi, Z. Wang, H. Yuan, J. Zhang, M. Zhou, "Self-adaptive Teaching-learning-based Optimizer with Improved RBF and Sparse Autoencoder for High-dimensional Problems," *Information Sciences*, vol. 630, pp. 463–481, Jun. 2023.

[25] 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 of Things Journal*, vol. 11, no. 9, pp. 16672–16683, May 2024.