Water Research X 29 (2025) 100402

Contents lists available at ScienceDirect

Water Research X

journal homepage: www.elsevier.com/locate/wroa

Long-term multivariate water quality forecasting for sustainable aquaculture management

Xiaodong Ji [a,b], Lu Liu [a,b], Bentao Duan [c], Ying Li [a,b], Haoran Xing [a,b], Bin Wang [a,*], Dashe Li [a,b]

[a] School of Computer Science and Technology, Shandong Technology and Business University, Yantai, 264005, Shandong, China
[b] Key Laboratory of Intelligent Information Processing, Shandong Technology and Business University, Yantai, 264005, Shandong, China
[c] Energy Management Service Center, Yantai University, Yantai, 264005, Shandong, China

ARTICLE INFO

Keywords:
Water quality
Long-term prediction
Multivariate
Aquaculture
Transformer

ABSTRACT

Accurate water quality prediction is essential for intelligent aquaculture management, enabling timely intervention, risk mitigation, and sustainable resource use. Key parameters such as dissolved oxygen, chlorophyll-a, and pH are influenced by complex spatiotemporal dynamics, making long-term forecasting particularly challenging in high-density aquaculture systems. Traditional methods struggle to balance local details and global trends, while circadian rhythms, feeding cycles, and seasonal shifts cause dynamic dependencies and distribution drift. To address these issues, we propose a novel deep learning framework with three core components: (1) a multi-scale decomposition module with time–frequency enhancement, which removes cross-scale redundancy, suppresses noise, and integrates local–global features via hierarchical decomposition and feature reorganization; (2) an adaptive sequence perception attention mechanism based on graph learning, which captures dynamic variable dependencies and models spatiotemporal interactions, including environmental coupling and aquaculture disturbances; and (3) a GRU-MoE network with a dynamic expert selection strategy that adjusts to data characteristics, mitigating distribution drift caused by human interventions like feeding and oxygenation. Extensive experiments on four real-world water quality datasets show the proposed method outperforms six deep learning baselines, achieving an average MAE reduction of 53.17%, RMSE reduction of 51.68%, R^2 improvement of 0.4945, and KGE improvement of 0.1979. Furthermore, Kolmogorov–Smirnov test results confirm the model's ability to recover real data distributions and their temporal evolution. This high-precision long-term prediction method enhances aquaculture system resilience, reduces risks from water quality fluctuations, and provides a robust foundation for informed decision-making and sustainable aquaculture management.

1. Introduction

Accurate prediction of aquaculture water quality parameters is the core foundation for realizing intelligent management of aquaculture (Uddin et al., 2022). By predicting key water quality parameters such as dissolved oxygen, chlorophyll a, temperature, turbidity, salinity and pH value, the aquaculture environment can be monitored in real time, water quality abnormalities can be detected in time, and aquatic animals can be prevented from growth restriction or disease due to harsh environment (Jayasiri et al., 2022), thereby improving aquaculture production and economic benefits; in addition, accurate water quality prediction can also optimize bait delivery and water body regulation strategies, reduce resource waste, and improve aquaculture sustainability. At the same time, it helps to assess the impact of climate change on marine aquaculture ecology and provide a basis for scientific management and policy formulation. Therefore, it is urgent to establish a high-precision long-term prediction model for aquaculture water quality parameters, which can not only provide dynamic early warning support for aquaculture models such as factory-scale recirculating water aquaculture and offshore cages, but also help environmental carrying capacity assessment and the formulation of regional aquaculture capacity standards, and provide solid data support for the transformation of the aquaculture industry to an eco-friendly and sustainable development model (Kim et al., 2023).

Water quality prediction models can generally be divided into two categories: physics-based models (Li et al., 2022) and data-driven models (Noori et al., 2020). Physics-based models simulate specific water chemical processes by constructing equations and parameterization schemes with clear physical meanings, and have been widely used in the field of water quality simulation and prediction (Quevedo-Castro

* Corresponding author.
E-mail address: Bin.seulement@gmail.com (B. Wang).

https://doi.org/10.1016/j.wroa.2025.100402
Received 7 April 2025; Received in revised form 19 August 2025; Accepted 20 August 2025
Available online 10 September 2025

et al., 2022). Such models are often combined with data assimilation technology to enhance their robustness and reliability (Rezaie-Balf et al., 2020). However, physics-based models have several inherent limitations: their dependence on idealized condition assumptions limits their predictive capabilities in complex or highly dynamic environments; model construction usually relies on detailed prior knowledge of the physical and chemical properties of water bodies, and obtaining such knowledge often requires a lot of experimental and observational resources, which significantly increases the technical threshold and cost of application (Wan et al., 2022a). More importantly, such models involve complex numerical calculation processes with high computational overhead, making it difficult to meet the timeliness requirements of real-time or near-real-time water quality prediction, thereby restricting their engineering applications and large-scale deployment in actual scenarios (Wan et al., 2022b).

With the advancement of data mining technology and the increasing abundance of environmental monitoring data, the research and application of data-driven water quality prediction models are becoming increasingly extensive (Li et al., 2020). The core goal of such models is not to reveal the physical and chemical mechanisms behind water quality changes, but to focus on mining the complex nonlinear mapping relationship between meteorological factors and water quality parameters (Sheng et al., 2023). According to the differences in modeling methodology, they can be divided into two categories: machine learning-based and deep learning-based models. Among them, traditional machine learning methods such as decision trees (Ahmed and Lin, 2021), support vector machines (SVMs) (Leong et al., 2021), and hidden Markov models (HMMs) (Li et al., 2023a) have made up for the shortcomings of traditional mechanism models to a certain extent with their powerful nonlinear modeling capabilities (Derot et al., 2020). However, these methods still face significant challenges in practical applications: high dependence on feature engineering, significant decrease in computational efficiency with increasing dimensions in high-dimensional and complex data scenarios, and limited model generalization ability are common. In addition, such methods are usually difficult to effectively capture the dynamic correlation between time steps in time series, which restricts their ability to model the time evolution characteristics of water quality (Wang et al., 2024a). Therefore, although they perform well in short-term prediction tasks, their prediction performance is obviously limited when dealing with water quality data with long time series dependence, multivariate coupling, and significant dynamic changes.

Deep neural networks have been widely used in time series prediction tasks due to their excellent feature learning ability, noise resistance and excellent generalization performance. In the field of water quality prediction, models based on convolutional neural networks (CNNs) have shown good performance due to their powerful local feature extraction capabilities. For example, methods such as MICN (Wang et al., 2023) and PDF (Dai et al., 2024) achieve high prediction accuracy by effectively capturing local periodic features and modeling long-term dependencies; while the improved model that introduces temporal convolutional networks (TCNs) (Li et al., 2023b) and double residual structures further enhances the interpretability of the model. However, such convolution-based methods overly rely on the design of convolution kernels when dealing with long-term dependencies, making it difficult to effectively model global dynamic associations. In contrast, recurrent neural networks (RNNs) can achieve dynamic transmission of time series information through their inherent directed cyclic structure. Its important variants, such as long short-term memory networks (LSTMs) (Zhou et al., 2023) and gated recurrent units (GRUs) (Seifi et al., 2024), are widely used in complex time series prediction tasks due to their stronger nonlinear modeling capabilities and memory mechanisms. In particular, in water quality prediction applications, GRU-based models can effectively capture long-term dependency characteristics in multivariate sequences through their gating mechanisms (Li et al., 2021). Nevertheless, although LSTM and GRU have advantages over traditional methods in modeling long-term dependencies, their global modeling capabilities are still limited when dealing with sequences with very long time spans (Li et al., 2024).

In recent years, deep learning models based on the Transformer architecture (Vaswani, 2017) have made significant progress in the field of time series forecasting. With the powerful representation ability of its core self-attention mechanism, such models can effectively capture the complex and dynamic dependencies between time points and show excellent performance when processing time series data with highly nonlinear characteristics. It is particularly worth noting that in long sequence prediction tasks, its inherent global modeling ability effectively breaks through the bottleneck of traditional methods that are limited by local dependencies. However, the computational complexity of the standard self-attention mechanism grows quadratically with the length of the sequence ($O(L^2)$), which constitutes the main constraint on its application in ultra-long sequence scenarios (Zeng et al., 2023). To this end, researchers have proposed a variety of efficient attention mechanisms, aiming to significantly reduce the computational cost while retaining its powerful sequence modeling advantages as much as possible. Among them, Sparse Attention (Arepalli et al., 2024) limits the scope of attention calculation so that it only focuses on key positions, which greatly reduces the computational overhead; Low-Rank Decomposition (Fan et al., 2021) and Kernel-Based Approaches (Gan et al., 2023) respectively achieve effective compression of computational complexity by performing low-rank approximation on the attention matrix or mapping it to a high-dimensional feature space; In addition, Segment-Based Attention (Du et al., 2023) and Window Mechanism (Tran and Xin, 2023) divide long sequences into multiple sub-intervals for separate modeling, which significantly reduces resource consumption while also improving the model's ability to balance local and global feature modeling.

Although Transformer-based optimization methods have made significant progress in long-sequence modeling, they still face many challenges when facing complex application scenarios such as aquaculture water quality parameter prediction, revealing the potential deficiencies of current mainstream methods in multi-scale feature fusion, multivariable coupling relationships, attention distribution identification, and robustness to concept drift. First, water quality parameter time series generally show significant multi-scale characteristics, and there are significant differences in their behavior patterns between short-term fluctuations and long-term trends. However, the existing Transformer architecture mostly relies on single-scale modeling and lacks effective scale decoupling and feature integration mechanisms, which can easily lead to key patterns (such as sudden drops in dissolved oxygen) being masked by redundant information or smooth trends, thereby weakening the stability of the prediction. Secondly, the dot product attention mechanism commonly used by mainstream methods tends to generate smooth and homogeneous attention distribution, making it difficult to focus on key time nodes such as sudden changes (such as sudden changes in turbidity caused by heavy rain events), resulting in insufficient response capabilities of the model to local abnormal events. Although models such as DECSF-Net (Song et al., 2025) have introduced cross-source data fusion strategies, their attention allocation mechanisms have not been effectively improved and have limited performance in sudden event prediction. Finally, due to factors such as seasonal changes and extreme weather, water quality data often experience distribution drift. However, current static model structures such as Transformer and LSTM lack adaptive adjustment capabilities and are difficult to cope with dynamic changes in data distribution. Although online learning methods (such as OneNet (Wen et al., 2023)) have improved the overall robustness of the model, their structure fails to effectively model the distribution differences of local data fragments and is difficult to solve the inconsistency problem between local features and global patterns. In view of this, this paper proposes a novel deep learning framework with the following main contributions:

Table 1
Prediction results of different models for the BafferCreek dataset for the next 7 days.

Model	Metrics	Temp	pH	Turbidity	Chl-a	DO
Proposed	MAE	0.1084	0.0085	0.1563	0.071	0.0543
	RMSE	0.1375	0.0115	0.1941	0.0946	0.0699
	R^2	0.9945	0.9716	0.9459	0.9788	0.9686
	KGE	0.9939	0.9824	0.9652	0.9718	0.9839
TimeDART	MAE	0.1274	0.0089	0.1563	0.1275	0.0589
	RMSE	0.1756	0.0121	0.2015	0.1475	0.0808
	R^2	0.9911	0.9685	0.9416	0.9485	0.958
	KGE	0.9931	0.9794	0.9275	0.9546	0.9656
MSGnet	MAE	0.2924	0.0289	0.453	0.3703	0.1426
	RMSE	0.3739	0.0357	0.5611	0.478	0.1806
	R^2	0.9596	0.7252	0.5476	0.4587	0.79
	KGE	0.9798	0.7891	0.6241	0.6286	0.8927
FourierGNN	MAE	0.1755	0.1069	0.4842	0.302	0.4517
	RMSE	0.2202	0.1391	0.6162	0.3608	0.4852
	R^2	0.986	−3.1819	0.4545	0.6917	−0.5158
	KGE	0.9907	0.0154	0.7232	0.8398	0.7986
TimeMixer	MAE	0.159	0.0136	0.2356	0.1102	0.0699
	RMSE	0.205	0.0162	0.293	0.1412	0.0922
	R^2	0.9879	0.9436	0.8766	0.9528	0.9453
	KGE	0.9788	0.9463	0.935	0.9693	0.9698
PatchTST	MAE	0.2006	0.011	0.2082	0.1045	0.0751
	RMSE	0.2276	0.0131	0.2455	0.137	0.0926
	R^2	0.985	0.9627	0.9134	0.9556	0.9448
	KGE	0.9893	0.9673	0.9076	0.9442	0.9626
iTransformer	MAE	0.4634	0.04	0.5197	0.3342	0.1572
	RMSE	0.5995	0.0492	0.6557	0.4272	0.197
	R^2	0.8961	0.4776	0.3822	0.5676	0.7502
	KGE	0.9571	0.7035	0.6974	0.7387	0.862

(1) A time–frequency enhanced multi-scale decomposable fusion strategy is proposed to eliminate redundant information in multi-scale time series data and balance local and global key features. Through time–frequency domain enhancement technology, the global trend and local detail characteristics in the time series are highlighted, and different time patterns are extracted using the improved moving average method. The sequence is decomposed into multiple scales by selecting an appropriate kernel size to ensure the diversity and independence of features at each scale; and redundant information is eliminated through residual connections to aggregate various time patterns.

(2) An adaptive sequence-aware attention mechanism is proposed to solve the problem of failing to capture key time points, local features, and multivariate dependencies due to the row homogeneity phenomenon caused by the traditional attention mechanism. By combining the dynamic changes in the time domain and the periodic characteristics in the frequency domain, the key time points and their characteristics are accurately captured, and the efficiency of attention allocation and feature extraction is optimized. At the same time, a graph structure framework is introduced to model the complex dependencies between multiple variables through graph representation and graph aggregation of time series.

(3) A GRU-MoE model is proposed to solve the problem of inconsistency between local and overall distributions caused by distribution drift. It avoids the model from overfitting local features and ignoring global trends, which in turn affects short-term predictions and long-term trend identification. This is a set of specially designed expert models, each of which is customized and optimized for the specific distribution of each patch in the input time series data, and automatically adjusts the expert's weights and strategies to achieve more accurate and adaptive predictions.

This study not only expands the theoretical framework of multi-scale time series modeling at the methodological level, but also responds to the urgent need for high-precision water quality prediction models in intelligent management of aquaculture at the application level. The proposed multi-scale decomposition strategy, adaptive attention mechanism and hybrid expert structure synergistically improve the model's ability to identify key features and adapt to complex environmental changes, showing good prediction stability and generalization performance. The research results can provide effective technical support for dynamic early warning, water quality regulation and ecological risk prevention and control in aquaculture scenarios, and promote the transformation of water quality management from experience-driven to data-driven. At the same time, it also provides a solid data foundation and theoretical support for ecological carrying capacity assessment, sustainable use of marine resources and related policy formulation in the context of climate change, which has important scientific significance and practical value.

The rest of this paper is organized as follows: Section 2 reports the principal experimental results and offers a detailed discussion. Section 3 concludes this paper and gives future work. Section 4 introduces the framework of our model.

2. Results and discussion

2.1. Results and analyses

To fully demonstrate the advancement of the proposed model in water quality prediction tasks, six current mainstream deep learning baseline models were selected for comparison, and experiments were conducted on multiple datasets. These baseline models cover different network architectures and modeling strategies, representing the typical methods of current deep learning in the field of water quality prediction. In the experiment, multiple evaluation indicators were used to quantitatively evaluate the prediction performance of the model from different dimensions to ensure comprehensiveness and fairness of the results.

2.1.1. Model comparison and analysis

The experimental results in Table 1 indicate that all evaluation indicators in this study are optimal on all datasets. Therefore, this model maintains a high prediction accuracy in most cases, fully proving its adaptability to diverse time series tasks and strong robustness.

Whether in terms of error measurement or trend fitting, the model in this study is superior to other comparison models, demonstrating leading overall performance.

Our model significantly outperforms existing approaches by effectively capturing both local and global dependencies in aquaculture water quality time series. Compared to TimeDART, which struggles to balance global dependencies and local details despite its self-supervised learning capabilities, our multiscale decomposable fusion approach with time–frequency enhancement adapts to varying time granularities and maintains high prediction accuracy, reducing MAE and RMSE by 14.31% and 15.94%, while improving R^2 by 0.0103 and KGE by 0.0.0154. MSGnet, though leveraging frequency domain analysis and adaptive graph convolution, suffers from redundant features across scales, increasing errors and obscuring patterns. Our model eliminates these redundancies via residual connections, reducing MAE and RMSE by 68.35% and 67.58%, while improving R^2 by 0.2757 and KGE by 0.1966. FourierGNN, which integrates spatiotemporal dynamics using hypervariable graphs, lacks fine decomposition and multiscale modeling, limiting its ability to capture trends and anomalies. Our model overcomes this by enhancing decomposition, reducing MAE and RMSE by 72.49% and 71.43%, while boosting R^2 by 1.2850 and KGE by 0.3059. TimeMixer, although effective in extracting key past information, loses dependencies as the prediction horizon increases. Our two-stage parallel sequence-aware attention mechanism dynamically captures time-domain changes and frequency-domain periodic features, improving feature extraction and attention allocation, while our graph-based learning framework models multivariate interactions to uncover hidden patterns. Compared to TimeMixer, our model reduces MAE and RMSE by 32.17% and 30.58%, while increasing R^2 by 0.0306 and KGE by 0.0196. PatchTST, which employs time series patch segmentation to retain local semantics, struggles with global dependencies due to its channel-independent strategy. Our model addresses this limitation, reducing MAE and RMSE by 30.67% and 25.64%, while improving R^2 by 0.0196 and KGE by 0.0252. Finally, iTransformer, which encodes time points into variable tokens, suffers from row homogeneity, leading to uniform attention weights and diminished differentiation across time points. Our model introduces a novel attention computation method integrating dynamic time-domain changes with periodic frequency-domain characteristics, precisely identifying key time points. Compared to iTransformer, our model reduces MAE and RMSE by 73.90% and 73.29%, while increasing R^2 by 0.3571 and KGE by 0.1877.

2.1.2. Visualization of prediction results

The results of the proposed model were compared with the prediction results of six mainstream deep learning baseline models on BaffleCreek datasets to more intuitively demonstrate its prediction ability. Fig. 1 presents the long-term prediction results of the proposed model for the next 7 days on the BaffleCreek dataset and the six comparison models.

In Fig. 1, the left panel shows the time series of each variable, where the blue and orange lines represent observed and predicted values of the proposed model, respectively, while dotted lines denote baseline models. The right panel presents the Taylor diagram for performance evaluation. Results show that the proposed model outperforms all baselines across multiple variables. For seawater temperature, its predictions align closest with observations, excelling in standard deviation, correlation coefficient, and CRMSE. Although PatchTST and TimeMixer also perform well, their standard deviation differences suggest lower accuracy in certain scenarios, particularly for fluctuating data. For seawater pH, the proposed model achieved the highest consistency with observed values, whereas TimeDART and FourierGNN struggled with capturing fluctuation amplitude and frequency. The model's time–frequency-enhanced multiscale fusion mechanism effectively highlights key information while reducing noise. In seawater turbidity prediction, the proposed model again showed the best alignment with observations. While iTransformer and TimeMixer performed reasonably well, their standard deviation discrepancies indicate lower reliability in some cases. For chlorophyll a and DO, the proposed model consistently demonstrated superior performance in key indicators. Although PatchTST and TimeMixer were relatively close in the Taylor diagram, differences in standard deviation suggest that their predictions may be less reliable under significant data variations, reinforcing the advantage of the proposed model.

Fig. 2 presents the model's multivariate long-term prediction results for the next 7 days on three different datasets, aiming to verify the generalization ability of the proposed model. To more intuitively evaluate the model performance, the observed and predicted values of each dataset were visualized and compared using line graphs and scatter plots, respectively. The line graph shows that the trends of the predicted values of the proposed model are highly consistent with the actual observed values, whether it is a public dataset greatly affected by natural environmental changes, such as the Mumford dataset, or a relatively stable water quality dataset in Shandong Peninsula, China. Especially in key peak and trough areas, the model can accurately capture the changing trends of water quality parameters. The scatter plot further confirms the accuracy and stability of the model.

The fitting line between the predicted and true values almost completely coincides with the Y=X line, suggesting that the prediction performance of the model on different datasets is consistent and has extremely high accuracy. In addition, the prediction error is evenly distributed on both sides of the Y=X line, indicating that the error exhibits a Gaussian distribution characteristic, which further proves the reliability of the model's prediction results. In general, whether from the perspective of the fitting degree of the prediction trend or the analysis from the perspective of error distribution, the model has demonstrated excellent stability and effectiveness in the long-term water quality prediction task. Compared with other models, the proposed model exhibited stronger generalization ability and prediction performance.

2.2. Ablation experiment

This study conducted ablation experiments on four aquaculture water quality parameter datasets to evaluate the effectiveness of each component in the proposed model. Different ablation variants were designed for comparative analysis by replacing or removing TF-MDM, the adaptive sequence-aware attention mechanism, and GRU-MoE. Specifically, In Ablation Variant 1 removed TF-MDM, while in Ablation Variants 2 and 3 retained only the time-domain and frequency-domain enhancements of TF-MDM, respectively. In Ablation Variant 4 adopted the self-attention mechanism from the original Transformer. In Ablation Variants 5 and 6 preserved only the time-domain and frequency-domain branches of the first stage of the adaptive sequence-aware attention mechanism, respectively, whereas in Ablation Variant 7 removed its graph learning component. In Ablation Variant 8, GRU-MoE was replaced with a simple linear flatten layer, and in Ablation Variant 9, a linear layer was used as the expert network. These ablation studies provide insights into the contributions of different model components and their impact on performance.

The experimental results in Table 2 demonstrate that the proposed full model outperforms all other models across all test datasets. The ablation results from variants 1, 2, and 3 clearly indicate that time–frequency enhanced multi-scale decomposition and fusion (TF-MDM) significantly enhance the model's predictive capabilities. In particular, time and frequency enhancements play complementary roles in capturing both temporal trends and frequency characteristics. By integrating time–frequency enhancement with multi-scale decomposition, the full model effectively improves prediction accuracy across various datasets. The results from ablation variants 4, 5, 6, and 7 further validate the effectiveness of the Ada-MSA module. This fully adaptive sequence perception mechanism integrates dynamic sequence modeling with time–frequency feature extraction, substantially enhancing the

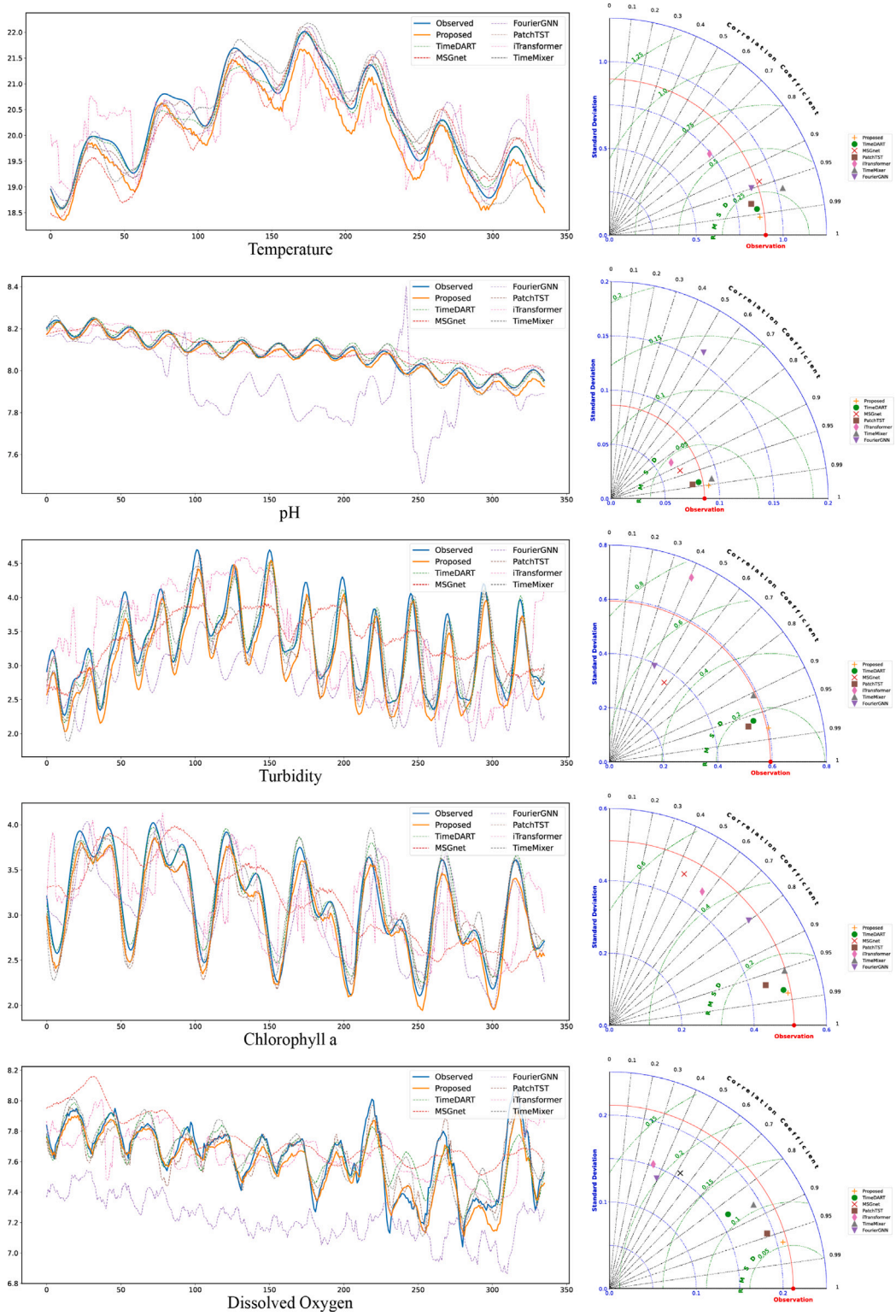

Fig. 1. Comparison and analysis of the prediction results with other models for the next 7 days on the BaffleCreek dataset.

model's ability to capture temporal patterns. As a result, it achieves the highest prediction accuracy across multiple datasets, highlighting its exceptional performance in time series modeling. Additionally, the results from ablation variants 8 and 9 underscore the crucial role of the GRU-MoE module in time series data modeling. The GRU expert excels at capturing temporal dependencies through its strong dynamic modeling capabilities, while the MoE mechanism enhances the model's adaptability and flexibility through dynamic task allocation. The synergy between these components enables the complete GRU-MoE model to achieve significant performance gains across diverse and complex datasets, demonstrating its strong potential in the field of time series modeling.

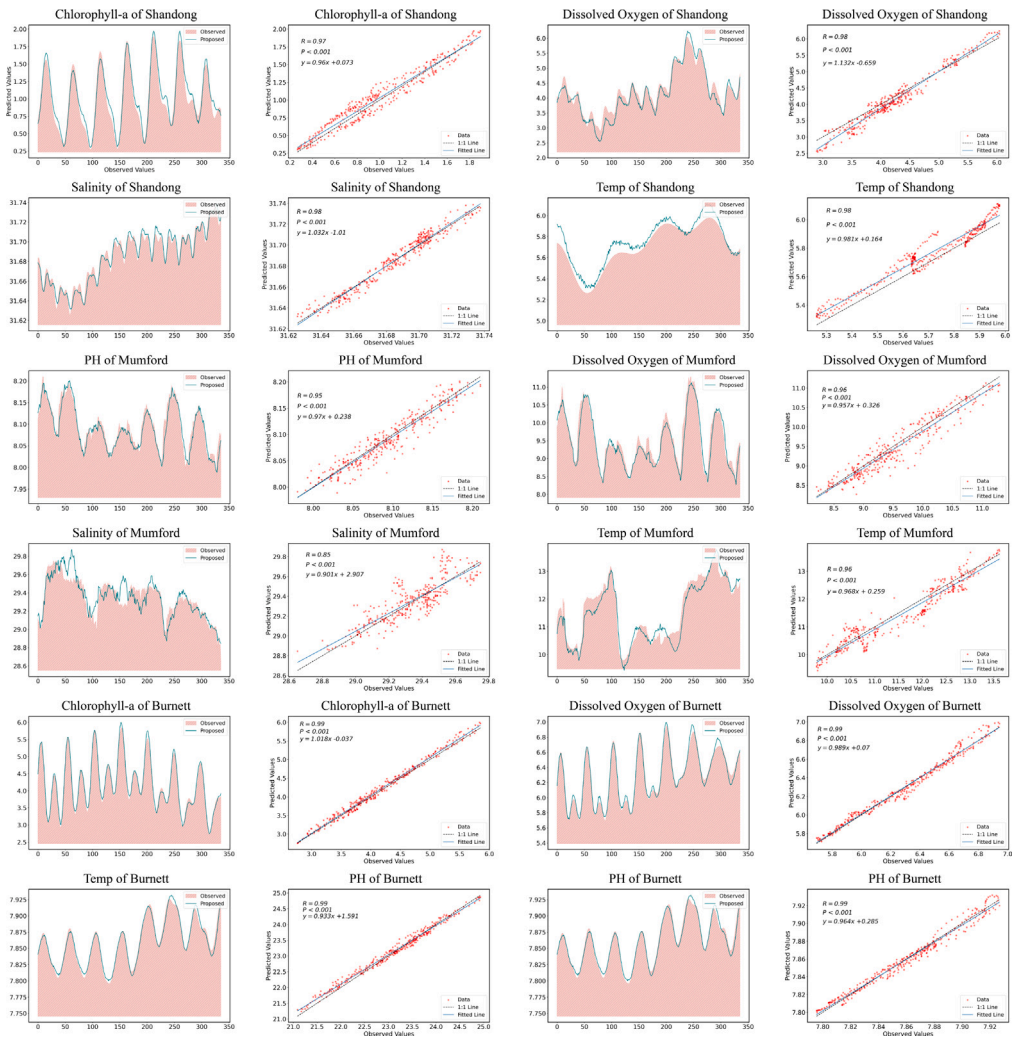

Fig. 2. Visualization of the model's 7-day prediction results across three datasets.

Table 2
Module ablation study of the proposed model.

Model	BaffleCreek		Mumford		Burnett		Shandong Peninsula	
	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
Proposed	0.4431	0.6948	0.4967	0.7379	0.6955	1.1608	0.4042	0.6111
Ablation 1	0.5117	0.7803	0.5342	0.7871	0.8351	1.3069	0.4227	0.6285
Ablation 2	0.4535	0.7035	0.5173	0.7739	0.7539	1.2225	0.4039	0.6101
Ablation 3	0.4534	0.7059	0.5258	0.7809	0.7352	1.2014	0.4209	0.6363
Ablation 4	0.4634	0.7526	0.5649	0.8530	0.7475	1.2720	0.4721	0.7391
Ablation 5	0.4545	0.7272	0.5088	0.7661	0.6974	1.1706	0.4334	0.6679
Ablation 6	0.4561	0.7460	0.5395	0.8077	0.7349	1.2462	0.4708	0.7371
Ablation 7	0.4522	0.7242	0.5582	0.8244	0.6958	1.2178	0.4301	0.6598
Ablation 8	0.4550	0.7360	0.5704	0.8518	0.7630	1.2756	0.4744	0.7417
Ablation 9	0.4537	0.7335	0.5984	0.8852	0.7277	1.2030	0.4724	0.7378

2.3. Discussion

2.3.1. Advantages of proposed model in predicting aquaculture water quality parameters

In order to cope with the challenges of multi-scale feature coupling, key dynamic identification and distribution drift in the prediction of aquaculture water quality parameters, this paper proposes an integrated deep neural network model. The core of the model includes three innovative designs: (1) Multi-scale decomposable fusion module, which can adaptively decompose and integrate multi-scale features in time series, significantly improving the model's ability to collaboratively model short-term disturbances and long-term trends; (2) Adaptive sequence-aware attention mechanism, which effectively alleviates the problem of weight distribution homogeneity in traditional dot-product attention and enhances the ability to identify key time point events (such as sudden water quality changes) and complex nonlinear dependencies; (3) GRU-MoE (Gated Recurrent Unit-Mixture of Experts) module, through a dynamic expert selection strategy based on input features, significantly improves the adaptability and robustness of the model under distribution drift and environmental disturbances. This integrated architecture constructs a collaborative optimization mechanism from multi-scale feature modeling, dynamic dependency

to distribution drift response, providing an effective solution for high-precision, high-stability and long-term reliable prediction of water quality parameters.

In recent years, a variety of modeling methods have been proposed from different perspectives to predict aquaculture water quality. For example, Nagaraju et al. proposed a model combining wavelet analysis and soft computing to predict ammonia nitrogen pollution in aquaculture ponds, emphasizing the importance of frequency domain decomposition in capturing complex dynamic processes (Nagaraju et al., 2023). Subsequently, Nagaraju et al. modeled the biochemical oxygen demand (BOD) of inland water bodies and revealed the complex coupling mechanism between its physical, chemical and ecological processes (Nagaraju et al., 2024a). Furthermore, Nagaraju et al. developed a comprehensive assessment tool for the Godavari Delta region in Andhra Pradesh, India, to address the management and assessment challenges of inland aquaculture environments (Nagaraju et al., 2024b). In addition, Gottumukkala et al. proposed a water quality index construction method based on machine learning, aiming to balance the dynamic relationship between aquaculture activities and estuarine ecosystems (Gottumukkala et al., 2024). Although these studies are representative in local parameter modeling or regional management, most of them focus on specific indicators, adopt static modeling or rule-driven methods, and have not fully considered core issues such as multivariate coupling dependence, multi-scale time series feature fusion, and distribution drift. In contrast, the method proposed in this paper has more advantages in complex dynamic structure modeling, generalization ability enhancement, and prediction robustness improvement, especially for long-term water quality prediction scenarios with multiple parameters and multiple time scales.

This paper also conducts comparative analysis with a variety of baseline models. The FourierGNN model based on graph neural network (GNN) performs well in mining variable correlation and time information, ensuring good prediction performance, but its insufficient use of multi-scale information limits further improvement. Although the TimeMixer model based on multi-layer perceptron (MLP) can achieve good results, it ignores the redundant information of scale data and fails to fully model the correlation between variables, which affects the prediction accuracy. The PatchTST and TimeDART models use channel independence strategies to reduce potential noise interference between variables and achieve good results, but they also ignore the dependency between variables. Crucially, all baseline models do not effectively deal with the problem of time series distribution drift, which further restricts their prediction capabilities.

2.3.2. Limitations and future research goals

While the proposed method has been validated on four real-world water quality datasets, it still exhibits several limitations. First, the available datasets span only short periods (6–9 months) and fail to capture complete annual cycles (for instance, the Mumford dataset lacks summer observations). This constrains the model's capacity to learn long-term water quality dynamics and extreme phenomena such as interannual variability. Second, the data are spatially limited, relying primarily on a single or a small number of monitoring sites, which may hinder the model's generalization to water bodies with different hydrological or ecological conditions, or to other geographic regions. Moreover, the absence of critical external drivers (e.g., high-resolution meteorological, hydrological, and anthropogenic data) reduces the model's ability to adapt to sudden environmental changes, such as rapid declines in dissolved oxygen. Collectively, these factors may cause the model to overfit specific spatiotemporal conditions during training, thereby compromising its robustness when applied to new time periods, unfamiliar locations, or interference-prone scenarios.

In addition, the model's complex architecture and high computational demands present challenges for deployment on resource-constrained edge devices, such as embedded monitoring terminals in aquaculture farms. In practical applications, real-time monitoring and decision-making must be achieved within limited hardware capacity, requiring a careful balance between predictive accuracy and inference efficiency. Future research should investigate optimization strategies tailored to edge computing environments while maintaining satisfactory performance. Potential approaches include lightweight network designs (e.g., depthwise separable convolutions, low-rank factorization), structured pruning to reduce redundant parameters and computation, and knowledge distillation to transfer predictive capabilities from large models to smaller, more efficient student models. These techniques are expected to substantially lower inference latency and energy consumption, enabling adaptation to low-power hardware platforms while ensuring scalability in data-scarce and resource-limited contexts.

Future work should also emphasize the incorporation of more representative spatiotemporal datasets and external driving variables, the development of adaptive transfer strategies across diverse environments, and the integration of lightweight modeling techniques into both training and deployment. Such efforts will facilitate the broader application of intelligent water quality forecasting in aquaculture management.

3. Conclusion

This study proposes a deep learning model for accurately predicting aquaculture water quality parameters. This model utilizes time–frequency enhancement techniques to highlight local and global features and employs multi-scale decomposable fusion techniques to reduce data redundancy through residual concatenation. The model combines Transformer networks with graph learning to effectively handle temporal dependencies and variable correlations. Furthermore, the model integrates a hybrid expert network to automatically select optimal experts based on data patterns, thereby improving its adaptability and robustness in complex environments. Notably, the model maintains high accuracy even in the presence of nonstationary dynamics and shifting data distributions. An innovative KS test analysis method demonstrates that the model accurately captures dynamic changes and reproduces the realistic data distribution, outperforming existing baseline models in long-term prediction tasks. This research provides an advanced solution for water quality prediction and supports sustainable aquaculture management by improving water quality anomaly detection, optimizing feeding and oxygenation strategies, and reducing resource waste and ecological impact. In the future, this model has the potential to be integrated with traditional expert knowledge systems for water quality monitoring. By integrating data-driven deep learning with expert rule-based knowledge, the system enhances the interpretability and reliability of prediction results, enabling more effective anomaly detection and decision support. Especially under rare or extreme conditions (such as sudden turbidity caused by continuous heavy rainfall, a sharp drop in dissolved oxygen due to summer heat, or an abnormal surge in chlorophyll a caused by a red tide outbreak), the expert system, drawing on years of accumulated domain experience and mechanistic knowledge, provides context and appropriate thresholds, helping the model make more robust judgments in the absence of similar historical data. This fusion approach enables the construction of an intelligent water quality prediction framework that is both data-adaptive and domain-interpretable, significantly enhancing the system's practical value in complex and dynamic environments. Future research will continue to expand the model's applicability and scalability, including validating it under a wider range of environmental and data conditions, improving its deployment capabilities in resource-constrained settings, and further enhancing its robustness and generalization performance in extreme event response and long-term forecasting tasks.

Table 3
Basic statistics of all datasets.

Datasets	Input variables	Output variables	Time span	Total samples
BaffleCreek	Temp, pH, Turbidity, Chl-a, DO	Temp, pH, Turbidity, Chl-a, DO	1 year	13 000
Mumford	Temp, pH, DO, Salinity	Temp, pH, DO, Salinity	4 months	5000
Burnett	Chl-a, DO, pH, Temp, Turbidity	Chl-a, DO, pH, Temp, Turbidity	5 months	6000
Shandong Peninsula	Temp, Salinity, Chl-a,DO	Temp, Salinity, Chl-a,DO	5 months	6000

4. Materials and methods

4.1. Overview of the study area

In order to comprehensively evaluate the performance of the proposed multivariate prediction model, four real water quality monitoring data sets were selected in this paper, and the relevant statistical information is shown in Table 3. Considering the complex coupling and dynamic correlation between water quality parameters in aquaculture environments, we selected five representative key variables for modeling and analysis, namely chlorophyll a, temperature, turbidity, salinity and pH. The above parameters are not only widely used in existing studies, but also have important reference value in actual aquaculture monitoring. Specifically, chlorophyll a reflects algal biomass and is an important indicator of nutrient level and algal bloom risk; temperature affects gas solubility and metabolic rate, and is the core factor driving a variety of physical and biochemical processes; turbidity represents suspended matter concentration, which is related to pollutant input and water stability; salinity regulates osmotic pressure and water density, which cannot be ignored especially in marine aquaculture; pH affects chemical reactions in water and the physiological state of aquatic organisms. The joint modeling of these parameters helps to capture the nonlinear dependence and co-evolution laws between multiple variables in the water quality system, thereby improving the prediction accuracy and application breadth of the model.

The four datasets exhibit significant differences in hydrological and ecological characteristics, which may affect model performance. For example, the dataset collected from semi-enclosed aquaculture ponds exhibits relatively stable parameter variations and low noise levels, enabling more consistent model performance. In contrast, the dataset from open coastal areas is more susceptible to external disturbances such as tides, heavy rainfall, and human activities, resulting in greater variability and more frequent extreme events, which challenges the model's ability to capture sudden changes. Furthermore, differences in the mean, standard deviation, and seasonal patterns of the selected parameters across datasets affect the learned temporal dependencies, leading to differences in model generalization performance when transferring across different water bodies.

Baffle Creek (Australia), located in a subtropical climate zone, remains largely unaffected by industrial pollution, with its ecosystem retaining a relatively pristine state. Water quality in this region is primarily shaped by seasonal rainfall fluctuations, making it an ideal site for studying the natural evolution of non-polluted water bodies.

Mumford (USA), positioned in a temperate marine climate zone, is subject to water quality fluctuations driven by snowmelt runoff and precipitation during the winter-to-spring transition, particularly affecting pH, salinity, and dissolved oxygen levels. Research in this region provides valuable insights into the effects of external environmental stressors on freshwater aquaculture systems.

Burnett River Basin (Australia), spanning subtropical to tropical climate zones, is significantly influenced by agricultural activities. Agricultural runoff contributes to increased nutrient loads, which impact dissolved oxygen levels and plankton growth. Additionally, seasonal variations in precipitation and river discharge further regulate water quality dynamics.

Shandong Peninsula(China) is located in the temperate monsoon climate zone with four distinct seasons. Under the influence of ocean and atmospheric circulation, the water quality shows significant seasonal fluctuations. The superposition of factors such as seawater exchange, wind and wave action, and aquaculture activities further shapes the regional water quality dynamics, providing key data support for the study of water quality changes in marine aquaculture environments.

Understanding the temporal and spatial variations in water quality across these diverse ecological environments is essential for ensuring the stability and sustainability of aquaculture systems. Therefore, comprehensive research on these regions will enhance the adaptability of water quality prediction models and provide a scientific foundation for the optimal management of aquaculture environments.

4.2. Framework of the proposed forecasting system

The framework of the aquaculture water quality parameter prediction model is shown in Fig. 3, which consists of a time–frequency-enhanced multiscale decomposable fusion module, a Transformer architecture with an adaptive series-aware attention mechanism, and a mixture of experts module. This model is an end-to-end dynamic prediction model. The overall design focuses on capturing the non-stationarity, periodicity and mutation in time series. It can dynamically adjust the modeling strategy under different time scales, variable structures and data distribution conditions. First, the multivariate data enters the time–frequency-enhanced multiscale decomposable fusion module for dual-domain enhancement and multiscale decomposable fusion. After processing, the data is passed to the Transformer module, which first combines the time and frequency domain information to mine the intrinsic relationship between variables and then models the relationship between variables through graph learning. Finally, the Transformer output enters the GRU-MoE module for expert selection, expert network processing, and temporal pattern projection.

4.2.1. Time–frequency-enhanced multiscale decomposable fusion

Compared with the general time series, time series of aquaculture water quality parameters have significant multi-scale characteristics. Their short-term fluctuations and long-term trends behave differently on different time scales and are affected by the complex interactions of physical, chemical, and biological processes. Due to the large amount of redundant information between different time scales, it is difficult to effectively balance the short- and long-term key characteristics, demonstrating the characteristics of nonstationarity and the coexistence of local and global characteristics. Therefore, this study proposes a time–frequency-enhanced multiscale decomposable fusion method(TF-MDM) for the significant multiscale characteristics of aquaculture water quality parameters time series and the complex manifestations of short-term fluctuations and long-term trends on different time scales to achieve a more comprehensive capture and effectively fuse multiscale feature information. The specific modeling steps are as follows:

Step 1: The input feature of this module is $X \in \mathbb{R}^{L\times C}$, where L denotes the size of the lookback window and C denotes the number of variables. First, the global trend and local detail characteristics in the time series are highlighted through the time–frequency domain enhancement technology to better reflect the dynamic change characteristics of time series data at different scales. This study transforms the time series from the time domain to the frequency domain to enhance the globality of the time series. The input X is decomposed into Fourier basis, and the $X_{f_{k_n}}$ basis with the largest amplitude is selected to

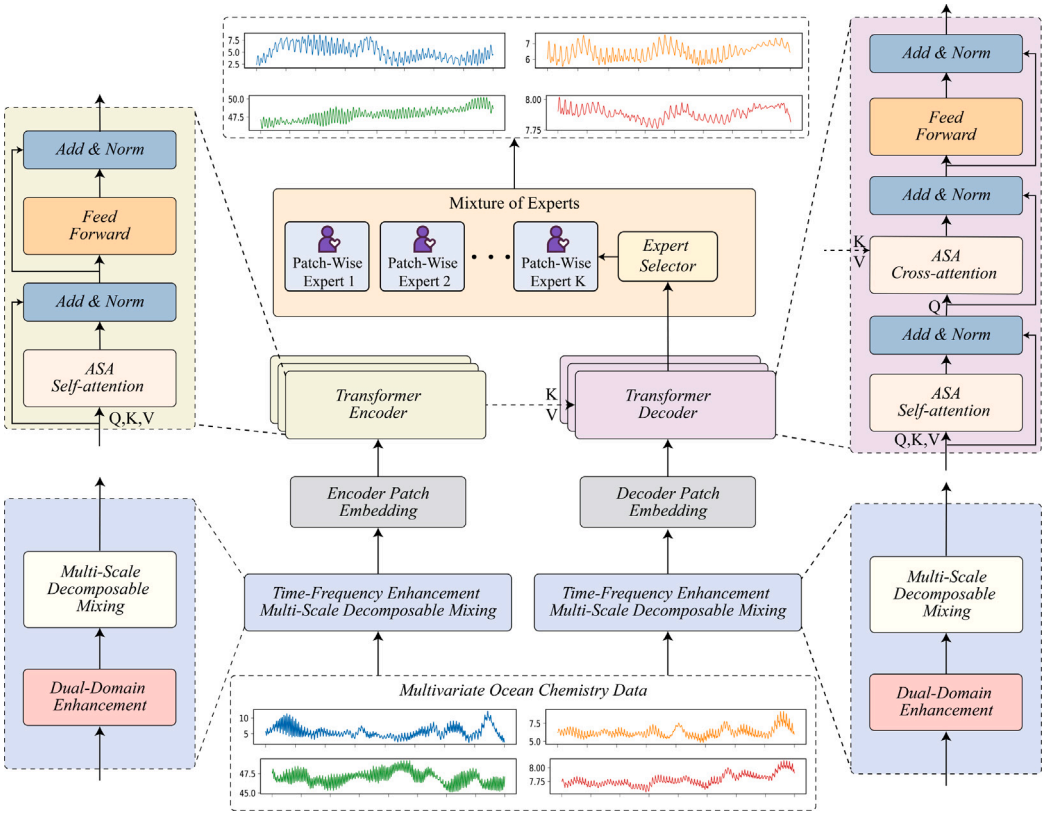

Fig. 3. The overall framework of the proposed model.

maintain the sparsity of the frequency domain and ensure the globality of the time series, $n \in [1, \ldots, N]$. As shown in Eqs. (1)–(3):

$$X_{f_n} = FFT(x_n) \tag{1}$$

$$X_{f_{k_n}} = TopK(X_{f_n}, K) \tag{2}$$

$$x_{if_n} = \mathrm{IFFT}(X_{f_{k_n}}) \tag{3}$$

Step 2: Then, in the time domain, the window function is used to perform operations, which helps enhance the local information of the time series and reduce the discontinuity caused by spectrum leakage. This study first defines a Hamming window $w[n] = 0.54 - 0.46\cos\left(\frac{2\pi n}{w}\right), \quad n = 1, \ldots, w$ with a window size of w (even number), where n indexes the sample point in the window. Then, x_{if_n} is reflected and padded to ensure that its length matches the window size w, as expressed in Eq. (4):

$$x_p[n] = \begin{cases} x_{if}\left[\frac{w}{2} - n\right], & 1 \le n \le \frac{w}{2} \\ x_{if}[n - \frac{w}{2}], & \frac{w}{2} < n \le N + \frac{w}{2} \\ x_{if}[N + w - n], & N + \frac{w}{2} < n \le N + w \end{cases} \tag{4}$$

The reflected and padded sequence x_p is convolved with the Hamming window w_n, as expressed in Eq. (5):

$$x_h[t] = \frac{\sum_{n=1}^{w} x_p[t+n] \cdot w[n]}{\sum_{n=1}^{w} w[n]}, \quad t = 1, \ldots, L \tag{5}$$

After the dual-domain enhancement operation, the global and local features of the time series are highlighted.

Step 3: $X_h \in \mathbb{R}^{L \times C}$ is decomposed into individual time patterns and then aggregated to enhance time series data. First, kernels of different sizes are used to obtain sequences containing information of different scales$(X_h^1, \ldots, X_h^S)$, as expressed in Eq. (6):

$$X_h^i = AvgPooling(X_h, c_i) \tag{6}$$

where c_i denotes the size of the ith scale information kernel. Then, different temporal patterns are mixed from the coarse scale X_h^S to the fine scale X_h^1 through a feedforward residual network. The mixing of the ith layer temporal pattern can be expressed as in Eq. (7):

$$X_h^i = X_h^i + MLP(X_h^{i+1}) \tag{7}$$

where X_h^i denotes the output of the ith layer of temporal mode mixing. Finally, after completing the temporal mode mixing of S scales, the mixed scale information $X^1 \in \mathbb{R}^{L \times C}$ is obtained.

4.2.2. Adaptive dynamic series-aware attention mechanism

The traditional attention mechanism employs dot product operations to assign weights, which easily leads to row homogeneity caused by the smooth and uniform distribution of attention weights. This phenomenon is manifested as the uniform distribution of weights at all time points and the lack of differentiation. This phenomenon weakens the importance of the model in capturing key time points, resulting in reduced sensitivity to local salient features, particularly in complex dynamic systems, where the key role of local anomalies or emergencies in overall prediction may be obscured. Moreover, in multivariate time series prediction, traditional methods cannot effectively simulate the dependencies between variables, thereby limiting the efficiency of information extraction. To this end, this study proposes an adaptive sequence-aware attention mechanism. The main modeling steps of this module are as follows:

In the first stage, a method for calculating attention weights that integrates time domain and frequency domain information was designed. This method effectively combines the dynamic changes of time series in the time domain with the periodic characteristics in the frequency

domain. As a result, the attention mechanism can more accurately identify key time points and their corresponding important features, thereby achieving more efficient attention allocation and feature extraction. The principle of time-aware attention in the first stage is to dynamically redirect and scale $\mathbf{Q} \in \mathbb{R}^{M\times H\times E}$ and $\mathbf{K} \in \mathbb{R}^{L\times H\times E}$, where M and L denote the sequence lengths; H, the number of heads; and E, the dimension of each attention head. The attention score of each head is expressed as shown in Eqs. (8)–(10):

$$\phi_p = f_p(tanh(x)) \tag{8}$$

$$f_p(x) = x \cdot w_{\text{dir}} \cdot (\text{std}(x))^{-p} \cdot \lambda_{\text{dyn}} \tag{9}$$

$$Score(\mathbf{Q}_i, \mathbf{K}_j) = \phi_p(\mathbf{Q}_i)\phi_p(\mathbf{K}_j)^T \tag{10}$$

where ϕ_p denotes a specially designed function applied to $\mathbf{Q}$ and $\mathbf{K}$ and w_{dir} and λ_{dyn} represent the learnable direction matrix and dynamic parameters, respectively. These two parameters help achieve the "dynamic" effect proposed in this study, and $std(x)$ represents the standard deviation of the input x. In addition, a dynamic scaling factor τ is introduced to calculate the attention weight as shown in Eq. (11):

$$A = Softmax(\frac{Score \cdot scale}{\tau}) \tag{11}$$

Where them, $\tau = \sqrt{\text{var}(Score)}$ dynamically normalizes the score, $var(Score)$ calculates the variance of the score, and scale is the scaling factor. Finally, the output of this attention is as shown in Eq. (12):

$$O_t = \sum_s A \cdot \mathbf{V} \tag{12}$$

Frequency domain-aware attention can effectively reflect the frequency characteristics of time series. As shown in Eq. (13):

$$O_f = \mathcal{F}^{-1}(Softmax(\frac{\mathcal{F}(\mathbf{Q}) \cdot \mathcal{F}(\mathbf{K})}{scale}) \cdot \mathcal{F}(\mathbf{V})) \tag{13}$$

Where $\mathcal{F}(\cdot)$ denotes the fast Fourier transform, which converts the time series from the time domain to the frequency domain.

In the second stage, graph structure learning is used to model cross-sequences. Specifically, the entire graph structure learning module can be described as shown in Eqs. (14)–(16):

$$\mathbf{M}_1 = \arctan(\mathbf{E}\mathbf{\Theta}_1) \tag{14}$$

$$\mathbf{M}_2 = \arctan(\mathbf{E}\mathbf{\Theta}_2) \tag{15}$$

$$\mathbf{A}' = \text{Relu}(\mathbf{M}_1\mathbf{M}_2^T - \mathbf{M}_2\mathbf{M}_1^T) \tag{16}$$

The node embedding of the sequence is generated by randomly initializing the matrix $\mathbf{E} \in \mathbb{R}^{N\times D_g}$, where D_g denotes the feature dimension of the node embedding. $\mathbf{E}$ is transformed using trainable parameters $\mathbf{\Theta}_1, \mathbf{\Theta}_2 \in \mathbb{R}^{D_g\times D_g}$ and a nonlinear activation function 'act' to obtain a new representation matrix $\mathbf{M} \in \mathbb{R}^{N\times D_g}$. Subsequently, for each node, the first K-nearest neighbor nodes are identified from the adjacency relationship $\mathbf{A}'$ as its neighbors, and the weights of unconnected nodes are set to zero. Through this process, the sparse adjacency matrix $\mathbf{A} \in \mathbb{R}^{N\times N}$ of the unidirectional graph $\mathbf{G}$ is finally generated.

The core goal of graph aggregation is to fuse the information of each sequence with that of its neighbors via iterative message passing to capture dependencies between sequences, thereby strengthening the representation of each sequence on related patterns. In the lth layer, each sequence is represented as a global tokenized sequence matrix $\mathbf{G}^{(l)} \in \mathbb{R}^{N\times D}$. Then, the tokens of all sequences in the lth layer are collected and input into the graph neural network, and cross-sequence information interaction and joint modeling are achieved through the graph aggregation mechanism. As shown in Eq. (17):

$$\hat{\mathbf{G}}_i = \sum_{d=0}^{D} \tilde{\mathbf{A}}^d \mathbf{G}_i \mathbf{W}_d \tag{17}$$

This equation represents multihop information fusion on the graph, where D denotes the depth of graph aggregation and $\hat{A}$ is the graph Laplacian matrix. Each embedding $\hat{\mathbf{G}}_i$ is assigned to its original sequence connection, generating a graph-enhanced embedding $\hat{X}^{(l)} \in \mathbb{R}^{N\times C\times d}$.

4.2.3. GRU-MoE

After patching the aquaculture water quality parameter data, the distribution of local patches often deviates from the original pattern owing to distribution drift and external factors. This inconsistency may lead to the destruction of the original pattern. Such a distribution drift can easily cause the model to overfit local features and ignore the overall trend. This poses a challenge to dealing with short-term mutations or identifying long-term trends. To this end, this study proposes the GRU-MoE model.

This model is a set of specially designed expert models, each of which is customized and optimized for the specific distribution of each patch in the input time series data to achieve more accurate and adaptive predictions. The main modeling steps are as follows:

Step 1: An expert selector is built. According to the gating network Gate, the gating weight of each expert on the patch is calculated and the top k experts are selected. The gating weight is calculated as shown in Eqs. (18)–(19):

$$R(\mathbf{X}^i) = \mathbf{X}^i + \psi \cdot \text{Softplus}(\mathbf{X}^i) \cdot \mathbf{W}_{\text{noise}} \tag{18}$$

$$Gate(\mathbf{X}^i) = \text{Softmax}(\text{Topk}(R(\mathbf{X}^i), k)) \tag{19}$$

$\mathbf{X}^i \in \mathbb{R}^{C\times d}$, where k denotes the number of selected experts; $\psi \in \mathbb{N}(0, 1)$, standard Gaussian noise; and $\mathbf{W}_{\text{noise}} \in \mathbb{R}^{d\times d}$, a learnable weight that controls the noise value. The sum of the weight parameters of each expert is one. Our TopK method is as shown in Eq. (20):

$$\text{TopK}(\mathbf{u}, k) = \begin{cases} \alpha \cdot \log(\mathbf{u} + 1), & \text{if} \mathbf{u} < u_k \\ \alpha \cdot \exp(\mathbf{u}) - 1, & \text{if} \mathbf{u} \geq u_k \end{cases} \tag{20}$$

where u_k denotes the kth largest value in u and α is a constant used to adjust the selector weight.

Step 2: Each patch is input into the corresponding selected expert network. GRU-MoE contains K expert networks, denoted as $E_1, \ldots, E_K$. Each expert network consists of two layers of GRU neural networks. Given a patch input, each expert network E_k processes the input to generate its own output. As shown in Eqs. (21)–(25):

$$r_t = sigmoid(W_r * [h_{t-1}, x_t] + b_r) \tag{21}$$

$$z_t = \text{sigmoid}(W_t * [h_{t-1}, x_t] + b_z) \tag{22}$$

$$\widetilde{h}_t = tanh(W_{\widetilde{h}} * [r_t \cdot h_{t-1}, x_t] + b_{\widetilde{h}}) \tag{23}$$

$$h_t = (1 - z_t) \cdot h_{t-1} + z_t \cdot \widetilde{h}_t \tag{24}$$

$$y_t = sigmoid(W_o * h_t + b_y) \tag{25}$$

Step 3: Temporal pattern projection is performed on the output of the expert network to obtain the final prediction result. The final output of GRU-MoE is the weighted sum of the outputs of all selected experts, and the weights are provided by the gating network. As shown in Eq. (26):

$$\hat{Y} = Flatten(\sum_{k=1}^{K} Gate(X^i)E_k(X^i)) \tag{26}$$

4.3. Evaluation metrics

It is the core technology to evaluate the overall performance of all models used in this study. Four time series evaluation metrics, Mean absolute error(MAE), Root mean squared error(RMSE), Coefficient of determination(R^2), and Kling–Gupta efficiency(KGE), were used to analyze this model and other baselines. The performance metrics are calculated as shown in Eqs. (27)–(30):

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_{obs} - y_{pred}| \tag{27}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_{obs} - y_{pred})^2} \tag{28}$$

$$R^2 = 1 - \frac{\sum_i (y_{pred}^{(i)} - y_{obs}^{(i)})^2}{\sum_i (\bar{y}_{obs} - y_{obs}^{(i)})^2} \tag{29}$$

$$KGE = 1 - \sqrt{(r-1)^2 + \left(\frac{\sigma_{pred}}{\sigma_{obs}} - 1\right)^2 + \left(\frac{\mu_{pred}}{\mu_{obs}} - 1\right)^2} \tag{30}$$

where y_{obs} denotes the observed values, y_{pred} denotes the predicted values, N is the total number of observations, r is the Pearson's correlation coefficient between the predicted values and observations, σ_{pred} and σ_{obs} are the standard deviations of predicted values and observations, and μ_{pred} and μ_{obs} are the mean values of predicted values and observations respectively. μ_{pred} and μ_{obs} are the mean of the predicted and observed values, respectively. MAE and RMSE range from 0 to $+\infty$, with values closer to 0 indicating better predictive performance. R^2 and KGE range from $-\infty$ to 1, with values closer to 1 indicating better predictive performance.

4.4. Experimental settings

To ensure a comprehensive and fair comparison, the experiments were conducted as follows: (1) The four datasets (Shandong Peninsula, BaffleCreek, Mumford, Burnett) were divided into training sets (70%), validation sets (10%), and test sets (20%). Each dataset contains water quality parameters (e.g., temperature, dissolved oxygen, chlorophyll a) sampled every 30 min. (2) In order to be consistent with the modeling configuration of baseline models such as TimeMixer and iTransformer. (3) In the proposed model, a 7-day historical observation window is used as input features and the next 7 days are predicted to ensure experimental consistency. (4) The baseline models (such as TimeMixer and iTransformer) are trained under the same conditions (batch size = 24, AdamW optimizer, learning rate = 1e-4), and their best results are compared with the results of the proposed model. In addition, to alleviate overfitting, the experiment adopts dropout (rate = 0.1) and early stopping (patience = 3).

Datapreprocessing: During data preprocessing, all water quality parameters were normalized using z-scores to ensure consistency across numerical scales. Outliers were detected using sliding window statistical analysis (mean, standard deviation, and z-score), and further validated by combining multivariate correlations. For example, the temporal correlation and variation patterns of anomalies with other water quality parameters were examined to distinguish true environmental responses from noise. Anomalous data confirmed to be environmentally significant were retained and annotated, enabling the model to learn water quality characteristics under different scenarios. Missing values due to sensor failure or communication interruptions were filled using time series interpolation or multivariate regression. High-frequency noise unrelated to the environment was suppressed using time–frequency smoothing. This processing effectively preserves key environmental information while minimizing the impact of noise on model stability, thereby improving the model's adaptability to real-world aquaculture scenarios.

Baseline: To demonstrate the performance of the proposed model, this study selects baselines from the following five aspects: CNN-based (MSGnet(AAAI2024) (Cai et al., 2024)), GNN-based (FourierGNN(NIPS2023) (Yi et al., 2024)), diffusion-based (TimeDART (ICML2025) (Anonymous, 2025)), MLP-based (TimeMixer(ICLR2024) (Wang et al., 2024b)), and Transformer-based (PatchTST(ICLR2023) (Nie et al., 2023) and iTransformer(ICLR2024) (Liu et al., 2024)) models.

CRediT authorship contribution statement

Xiaodong Ji: Writing – original draft, Visualization, Validation, Software, Methodology, Conceptualization. **Lu Liu:** Validation, Software, Data curation. **Bentao Duan:** Validation, Funding acquisition, Data curation. **Ying Li:** Software, Project administration, Investigation, Data curation. **Haoran Xing:** Investigation, Funding acquisition, Formal analysis, Data curation. **Bin Wang:** Software, Investigation, Data curation. **Dashe Li:** Resources, Project administration, Investigation, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

All authors thank the editors and reviewers for the attention paid to this paper. We kindly request that the editors and reviewers provide valuable comments and corrections for this study. This work is supported by the National Natural Science Foundation of China under grant no. 71973106 and the Yantai Science and Technology Innovation Development Program Project under grant nos. 2021XDHZ062 and 2022JCYJ032.

Data availability

Data will be made available on request.

References

Ahmed, M.H., Lin, L.S., 2021. Dissolved oxygen concentration predictions for running waters with different land use land cover using a quantile regression forest machine learning technique. J. Hydrol. 597, 126213.

Anonymous, 2025. Diffusion auto-regressive transformer for effective self-supervised time series forecasting. URL: https://openreview.net/forum?id=yGv5GzlBwr.

Arepalli, P.G., Naik, K.J., Rout, J.K., 2024. Aquaculture water quality classification with sparse attention transformers: Leveraging water and environmental parameters. In: Proceedings of the 2024 13th International Conference on Software and Computer Applications. pp. 318–325.

Cai, W., Liang, Y., Liu, X., Feng, J., Wu, Y., 2024. Msgnet: Learning multi-scale inter-series correlations for multivariate time series forecasting. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, (10), pp. 11141–11149.

Dai, T., Wu, B., Liu, P., Li, N., Bao, J., Jiang, Y., Xia, S.T., 2024. Periodicity decoupling framework for long-term series forecasting. In: The Twelfth International Conference on Learning Representations.

Derot, J., Yajima, H., Schmitt, F.G., 2020. Benefits of machine learning and sampling frequency on phytoplankton bloom forecasts in coastal areas. Ecol. Informatics 60, 101174.

Du, D., Su, B., Wei, Z., 2023. Preformer: predictive transformer with multi-scale segment-wise correlations for long-term time series forecasting. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, IEEE, pp. 1–5.

Fan, X., Liu, Z., Lian, J., Zhao, W.X., Xie, X., Wen, J.R., 2021. Lighter and better: low-rank decomposed self-attention networks for next-item recommendation. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 1733–1737.

Gan, Y., Fu, Y., Wang, D., Li, Y., 2023. A novel approach to attention mechanism using kernel functions: Kerformer. Front. Neurorobotics 17, 1214203.

Gottumukkala, S.B., Thotakura, V.N., Gvr, S.R., Chinta, D.P., Park, R., 2024. Balancing aquaculture and estuarine ecosystems: Machine learning–based water quality indices for effective management. Environ. Sci. Pollut. Res. 1–17.

Jayasiri, M., Yadav, S., Dayawansa, N., Propper, C.R., Kumar, V., Singleton, G.R., 2022. Spatio-temporal analysis of water quality for pesticides and other agricultural pollutants in Deduru Oya river basin of Sri Lanka. J. Clean. Prod. 330, 129897.

Kim, S., Kim, S., Hwang, S., Lee, H., Kwak, J., Song, J.H., Jun, S.M., Kang, M.S., 2023. Impact assessment of water-level management on water quality in an estuary reservoir using a watershed-reservoir linkage model. Agricult. Water. Manag. (ISSN: 0378-3774) 280, 108234. http://dx.doi.org/10.1016/j.agwat.2023.108234.

Leong, W.C., Bahadori, A., Zhang, J., Ahmad, Z., 2021. Prediction of water quality index (WQI) using support vector machine (SVM) and least square-support vector machine (LS-SVM). Int. J. River Basin Manag. 19 (2), 149–156.

Li, W., Fang, H., Qin, G., Tan, X., Huang, Z., Zeng, F., Du, H., Li, S., 2020. Concentration estimation of dissolved oxygen in pearl river basin using input variable selection and machine learning techniques. Sci. Total Environ. 731, 139099.

Li, K., Huang, G., Wang, S., Razavi, S., 2022. Development of a physics-informed data-driven model for gaining insights into hydrological processes in irrigated watersheds. J. Hydrol. 613, 128323.

Li, W., Liu, C., Xu, Y., Niu, C., Li, R., Li, M., Hu, C., Tian, L., 2024. An interpretable hybrid deep learning model for flood forecasting based on transformer and LSTM. J. Hydrol.: Reg. Stud. 54, 101873.

Li, D., Sun, Y., Ruan, J., 2023a. Time series forecasting algorithm based on GOSSA and HMM. Electron. J. 51 (9), 2492–2503.

Li, W., Wu, H., Zhu, N., Jiang, Y., Tan, J., Guo, Y., 2021. Prediction of dissolved oxygen in a fishery pond based on gated recurrent unit (GRU). Inf. Process. Agric. 8 (1), 185–193.

Li, D., Zhang, X., Yang, Y., Yang, H., Liu, S., 2023b. An interpretable hierarchical neural network insight for long-term water quality forecast: A study in marine ranches of Eastern China. Ecol. Indic. 146, 109771.

Liu, Y., Hu, T., Zhang, H., Wu, H., Wang, S., Ma, L., Long, M., 2024. Itransformer: Inverted transformers are effective for time series forecasting. In: The Twelfth International Conference on Learning Representations. URL: https://openreview.net/forum?id=JePfAI8fah.

Nagaraju, T.V., Bala, G.S., Bonthu, S., Mantena, S., 2024a. Modelling biochemical oxygen demand in a large inland aquaculture zone of India: Implications and insights. Sci. Total Environ. 906, 167386.

Nagaraju, T.V., Malegole, S.B., Chaudhary, B., Ravindran, G., Chitturi, P., Chinta, D.P., 2024b. Novel assessment tools for inland aquaculture in the western godavari delta region of andhra pradesh. Environ. Sci. Pollut. Res. 31 (25), 36275–36290.

Nagaraju, T.V., Sunil, B., Chaudhary, B., Prasad, C.D., Gobinath, R., 2023. Prediction of ammonia contaminants in the aquaculture ponds using soft computing coupled with wavelet analysis. Environ. Pollut. 331, 121924.

Nie, Y., Nguyen, N.H., Sinthong, P., Kalagnanam, J., 2023. A time series is worth 64 words: Long-term forecasting with transformers. In: The Eleventh International Conference on Learning Representations. URL: https://openreview.net/forum?id=Jbdc0vTOcol.

Noori, N., Kalin, L., Isik, S., 2020. Water quality prediction using SWAT-ANN coupled approach. J. Hydrol. 590, 125220.

Quevedo-Castro, A., Bustos-Terrones, Y.A., Bandala, E.R., Loaiza, J.G., Rangel-Peraza, J.G., 2022. Modeling the effect of climate change scenarios on water quality for tropical reservoirs. J. Environ. Manag. 322, 116137.

Rezaie-Balf, M., Attar, N.F., Mohammadzadeh, A., Murti, M.A., Ahmed, A.N., Fai, C.M., Nabipour, N., Alaghmand, S., El-Shafie, A., 2020. Physicochemical parameters data assimilation for efficient improvement of water quality index prediction: Comparative assessment of a noise suppression hybridization approach. J. Clean. Prod. 271, 122576.

Seifi, A., Pourebrahim, S., Ehteram, M., Shabanian, H., 2024. A robust multi-model framework for groundwater level prediction: The BFSA-MVMD-GRU-RVM model. Results Eng. 24, 103250.

Sheng, S., Lin, K., Zhou, Y., Chen, H., Luo, Y., Guo, S., Xu, C.Y., 2023. Exploring a multi-output temporal convolutional network driven encoder-decoder framework for ammonia nitrogen forecasting. J. Environ. Manag. 342, 118232.

Song, L., Song, Y., Tian, Y., Quan, J., 2025. DECSF-net: A multi-variable prediction method for pond aquaculture water quality based on cross-source feedback fusion. Aquac. Int. 33 (4), 1–25.

Tran, N.T., Xin, J., 2023. Fourier-mixed window attention: Accelerating informer for long sequence time-series forecasting. arXiv preprint arXiv:2307.00493.

Uddin, M.G., Nash, S., Rahman, A., Olbert, A.I., 2022. A comprehensive method for improvement of water quality index (WQI) models for coastal water quality assessment. Water Res. 219, 118532.

Vaswani, A., 2017. Attention is all you need. Adv. Neural Inf. Process. Syst.

Wan, H., Mao, Y., Cai, Y., Li, R., Feng, J., Yang, H., 2022a. An SPH-based mass transfer model for simulating hydraulic characteristics and mass transfer process of dammed rivers. Eng. Comput. 1–16.

Wan, H., Xu, R., Zhang, M., Cai, Y., Li, J., Shen, X., 2022b. A novel model for water quality prediction caused by non-point sources pollution based on deep learning and feature extraction methods. J. Hydrol. 612, 128081.

Wang, Y., Long, H., Zheng, L., Shang, J., 2024a. Graphformer: Adaptive graph correlation transformer for multivariate long sequence time series forecasting. Knowl.-Based Syst. 285, 111321.

Wang, H., Peng, J., Huang, F., Wang, J., Chen, J., Xiao, Y., 2023. Micn: Multi-scale local and global context modeling for long-term series forecasting. In: The Eleventh International Conference on Learning Representations.

Wang, S., Wu, H., Shi, X., Hu, T., Luo, H., Ma, L., Zhang, J.Y., Zhou, J., 2024b. TimeMixer: Decomposable multiscale mixing for time series forecasting. In: The Twelfth International Conference on Learning Representations. URL: https://openreview.net/forum?id=7oLshfEIC2.

Wen, Q., Chen, W., Sun, L., Zhang, Z., Wang, L., Jin, R., Tan, T., et al., 2023. Onenet: Enhancing time series forecasting models under concept drift by online ensembling. Adv. Neural Inf. Process. Syst. 36, 69949–69980.

Yi, K., Zhang, Q., Fan, W., He, H., Hu, L., Wang, P., An, N., Cao, L., Niu, Z., 2024. FourierGNN: Rethinking multivariate time series forecasting from a pure graph perspective. Adv. Neural Inf. Process. Syst. 36.

Zeng, A., Chen, M., Zhang, L., Xu, Q., 2023. Are transformers effective for time series forecasting? In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, (9), pp. 11121–11128.

Zhou, L., Zhao, C., Liu, N., Yao, X., Cheng, Z., 2023. Improved LSTM-based deep learning model for COVID-19 prediction using optimized approach. Eng. Appl. Artif. Intell. 122, 106157.