

Article

Time Series Prediction of Key Indicators of Aquatic Ecological Health in the Xiaoqing River Based on Differential Attention Mechanism

Yonggang Li ^{1,2,†}, Yan Wang ^{3,†}, Zhiyuan Wang ^{1,2}, Ke Li ^{1,2}, Zhongzhu Zhu ^{1,2}, Shuaishuai Wang ^{1,2}, Ning Li ^{4,*} and Jinfeng Ma ^{5,*}

¹ Jinan Municipal Hydrology Center, Jinan 250014, China; sd-yonggang@163.com (Y.L.)

² Jinan Key Laboratory of Aquatic Ecosystem Monitoring and Conservation, Jinan 250014, China

³ State Key Laboratory of Regional Environment and Sustainability, Research Center for Eco-Environmental Sciences, Chinese Academy of Sciences, Beijing 100085, China

⁴ China Academy of Urban Planning and Design, Beijing 100044, China

⁵ State Key Laboratory of Regional and Urban Ecology, Research Center for Eco-Environmental Sciences, Chinese Academy of Sciences, Beijing 100085, China

* Correspondence: lining@mail.bnu.edu.cn (N.L.); jfma@rcees.ac.cn (J.M.)

† These authors contributed equally to this work.

Abstract

Accurate assessment of aquatic ecological health relies on the effective characterization of the temporal dynamics of key water quality and algal indicators. To address challenges such as strong non-stationarity, significant noise interference, and complex multivariate coupling in water environment monitoring time series, this paper introduces a differential attention mechanism (DIFF Transformer) and constructs a time series prediction model for key indicators of aquatic ecological health. Taking the Huangtaiqiao monitoring section of the Xiaoqing River in Jinan, China, as the study area, multivariate time series prediction research was conducted for key proxy indicators of aquatic ecological health, including blue-green algae, chlorophyll, dissolved oxygen, turbidity, total phosphorus, ammonia nitrogen, permanganate index, and TOC, based on hourly measured data from 2017 to 2024. The results show that the differential attention mechanism improves balanced forecasting across multiple water quality and algal indicators by reducing isolated and unstable high-frequency residual fluctuations and enhancing the reconstruction of dominant temporal patterns. Compared with LSTM, GRU, and the standard Transformer, DIFF Transformer achieved the highest macro-averaged Nash-Sutcliffe efficiency coefficient among the learning-based models, with a macro-averaged NSE of 0.93. Per-indicator evaluation further showed that DIFF Transformer achieved NSE values above 0.95 for 11 of the 12 water quality indicators. In addition, DIFF Transformer obtained lower MSE, MAE, and RMSE than LSTM, GRU, and the standard Transformer. These results indicate that DIFF Transformer can provide continuous, stable, and high-temporal-resolution predictions of key aquatic ecological proxy indicators, supporting routine water quality trend tracking and dynamic aquatic ecological assessment.

Keywords: time series prediction; aquatic ecological health; differential attention mechanism; key water quality indicators; algal indicators

Academic Editor: Jianjun Ni

Received: 23 June 2026

Revised: 3 August 2026

Accepted: 5 August 2026

Published: 7 August 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

A healthy water environment is essential for the sustainable development of both ecosystems and human society [1–3]. Accurately understanding the characteristics of cyanobacteria and water quality in rivers plays a crucial role in water environment protection, as they are closely related to aquatic ecological health [4–6]. Achieving precise predictions of multivariate water quality indicators and cyanobacteria in rivers is key to assessing aquatic ecological health status [7,8]. However, traditional modeling methods are often time-consuming and labor-intensive when handling large volumes of input data [9,10]. Machine learning-based time series forecasting techniques offer new solutions for predicting water quality or cyanobacteria in rivers [11–13].

One of the core challenges in time series forecasting is how to effectively capture complex dependencies within the data, especially over long-time spans or across multivariate data [14,15]. Traditional machine learning methods perform poorly under conditions of multi-indicator prediction and big data, as their recursive computation patterns cause time costs to increase linearly with sequence length, making it difficult to achieve accurate and efficient time series forecasting [16,17]. The introduction of deep learning for time series prediction has overcome the aforementioned shortcomings and has developed rapidly in water quality forecasting [18]. To date, deep learning technologies have been applied to forecast water quality in the River Lee (London, UK) [19], Malaysian rivers [20], and southern Iran [21], and the Dez River, Iran [22], using models such as recurrent neural networks (RNNs) and long short-term memory (LSTM) networks et al. Nevertheless, conventional deep learning models, such as RNNs and LSTM, often fall short when facing long-term dependencies and the vanishing gradient problem [23–25].

The Transformer model, with its unique self-attention mechanism, has overcome the limitations of traditional methods, demonstrating significant advantages in computational efficiency and the ability to model both short- and long-term dependencies [26,27]. This model has been applied in fields such as remote sensing image change detection and bearing fault diagnosis, and water quality forecasting [28–31]. However, when processing long-sequence or complex time series data, challenges such as noise interference and difficulty in focusing on key information still remain.

Recently, the differential attention mechanism (DIFF) has been proposed and applied to improve the Transformer model. By computing the difference between two independent soft attention distributions, it effectively eliminates attention noise, enabling the model to focus more on key information while ignoring irrelevant context, thereby optimizing the performance of the Transformer model [32]. This mechanism demonstrates significant advantages in handling long-term dependencies and multivariate time series. However, the DIFF Transformer has been studied only in a few areas, such as autoregressive prediction of new energy power and flood detection in SAR imagery [33,34], and has not yet been applied to the field of aquatic ecological health prediction.

This study focuses on key proxy indicators within the aquatic ecological health assessment system and employs a differential attention mechanism to construct a DIFF Transformer prediction model for the aquatic ecological health of the Xiaoqing River. Aquatic ecological health is a comprehensive concept, typically characterized indirectly through multiple water quality indicators such as algae, biological oxygen demand, and nutrients. Rather than directly predicting a comprehensive health index, this study centers on key proxy indicators, with an emphasis on exploring the applicability and advantages of the differential attention mechanism for multivariate time series prediction in complex aquatic environments.

The constructed model offers the following advantages: ① Long-term dependencies: The DIFF Transformer leverages the differential attention mechanism to focus on contextual information at critical time points, enhancing the model's ability to model long time

series, thereby significantly improving the processing of long-span data from the Huang-taiqiao station on the Xiaoqing River. ② Multivariate modeling: The DIFF Transformer effectively captures complex correlations among multivariate time series data, such as blue-green algae and various water quality indicators, avoiding the signal loss caused by variable mixing in traditional models, and thus improving the accuracy of multivariate time series prediction. ③ DIFF Transformer noise suppression and information extraction: By reducing isolated and unstable high-frequency attention responses, the DIFF Transformer emphasizes dominant temporal features and improves the stability and balance of multi-indicator forecasting without treating all local variability as noise.

2. The Principle of DIFF Transformer

To improve the performance of the attention mechanism, Tianzhu Ye et al. proposed the Differential Transformer in 2024 [32]. Its core idea is to separate the underlying noise embedded within the data through a differential operation, thereby optimizing the attention mechanism and enhancing the model's data modeling capability. The differential attention mechanism is illustrated in Figure 1.

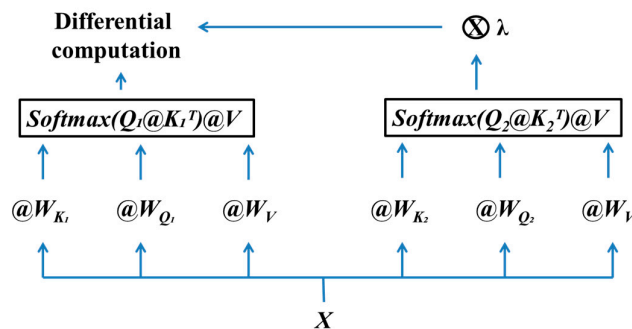


Figure 1. Schematic diagram of the differential attention mechanism.

Figure 1 illustrates the core architecture of the differential attention module, whose central structure is the differential multi-head attention mechanism, as expressed in Equation (1).

$$\text{DIFFAttn}(X) = (\text{softmax}(\frac{Q_1 K_1^T}{\sqrt{d}}) - \lambda \cdot \text{softmax}(\frac{Q_2 K_2^T}{\sqrt{d}}))V \quad (1)$$

In the formula, two independent Query inputs (Q_1 , Q_2) and Key input mappings (K_1 , K_2) are used to construct two parallel attention pathways, enabling the model to extract two distinct attention patterns from the data. These two patterns share the same set of value information, while their outputs are differentially integrated through a balancing coefficient λ , which is itself a learnable parameter. Its specific definition is shown in Equation (2):

$$\lambda = \exp(\lambda_{q_1} \cdot \lambda_{k_1}) - \exp(\lambda_{q_2} \cdot \lambda_{k_2}) + \lambda_{\text{init}} \quad (2)$$

where $\lambda_{q_1}, \lambda_{k_1}, \lambda_{q_2}, \lambda_{k_2} \in \mathbb{R}^d$ are all learnable parameters. In addition, to better align with the statistical characteristics of the data in specific application scenarios, some of these parameters can also be predefined as fixed values in order to achieve improved model performance.

3. Construction Process of the DIFF Transformer Time Series Prediction Model

3.1. Framework

Figure 2 presents the overall framework of the proposed multivariate forecasting model. Multi-year hourly observations from the Huangtaiqiao monitoring section,

including water quality, algal, rainfall, and discharge variables, were first subjected to temporal consistency checking, invalid value screening, missing value imputation, detection limit treatment, and standardization. The processed dataset was then divided into training, validation, and test sets, and sliding time windows were used to construct historical input sequences and future prediction targets. LSTM, GRU, the standard Transformer, and DIFF Transformer were trained under consistent experimental settings, while the key hyperparameters of DIFF Transformer, including input sequence length, prediction length, and number of attention layers, were calibrated using grid search based on validation performance. The final models were evaluated using NSE, MSE, MAE, and RMSE, together with residual and frequency domain analyses, to assess overall forecasting accuracy, cross-indicator performance balance, and the ability to reconstruct dominant temporal patterns.

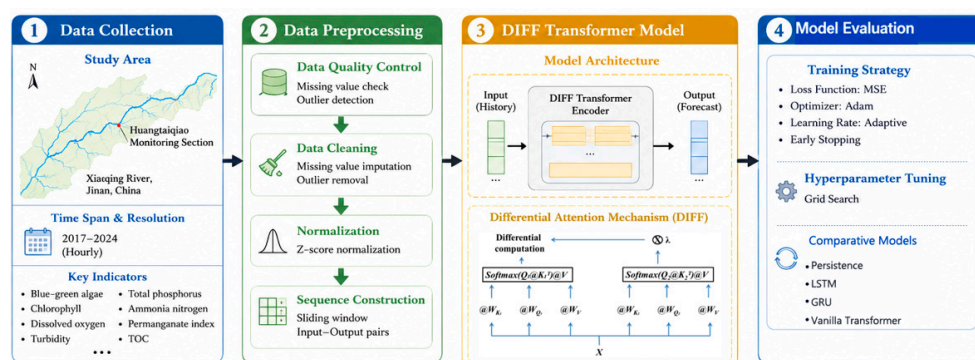


Figure 2. Framework of the DIFF Transformer-based multivariate forecasting model for key indicators of aquatic ecological health.

3.2. Case Study

Water quality data were collected at the Huangtaiqiao monitoring section (Figure 3) using a suite of online water quality analyzers and sampling systems. These included a Hach GLI 5500 five-parameter probe (Hach Company, Loveland, CO, USA, measuring temperature, pH, dissolved oxygen, conductivity, and turbidity), a Hach COD-203A permanganate index auto-analyzer (Hach Company, USA), a Hach NPTW-1600 nutrient auto-analyzer (Hach Company, USA, for ammonia nitrogen, total phosphorus, and total nitrogen), a Hach FP360 petroleum hydrocarbons monitor (Hach Company, USA), a Hach Astro TOCTM UV online total organic carbon (TOC) analyzer (Hach Company, USA), a Hach Hydrolab multi-parameter probe (Hach Company, USA, for blue-green algae and chlorophyll a), and a Greyspray FC-9624YL refrigerated autosampler (Beijing Grasp Technology Development Company, Beijing, China), supported by auxiliary modules (water intake/distribution, pre-treatment, and data transmission systems manufactured by ZTE Instruments), according to the Technical Specification for Surface Water Environmental Quality Monitoring (HJ 91.2–2022) [35]. Sensor calibration via low- and high-concentration standard sample verification was conducted monthly, and quality control followed the Technical Specification for Operation and Maintenance of Surface Water Quality Automatic Monitoring Stations (HJ 915.3–2024) [36]. Discharge and precipitation data were obtained from the Jinan Hydrology Center, with hourly rainfall records from catchment rain gauges and continuous discharge measurements at the Huangtaiqiao hydrological station adhering to the Specifications for Precipitation Observation (SL 21-2015) [37], the Code for Liquid Flow Measurement in Open Channels (GB 50179-2015) [38], and the Standard for Stage Observation (GB/T 50138-2010) [39]. The dataset consists of 62,013 hourly records from 3 May 2017 to 30 May 2024. Twelve water quality indicators and two

hydrometeorological variables were used. As shown in Table 1, the original water quality records had missing rates of 6.23–10.53%, depending on the indicator, while the final aligned modeling dataset contained no missing values. The indicators showed distinct dynamic ranges and seasonal patterns. Water temperature and precipitation peaked in summer, dissolved oxygen and pH were higher in winter, and nutrient-related variables showed variable-specific seasonality. These characteristics indicate strong multi-scale variability and cross-indicator heterogeneity, which motivated the use of multivariate sequence forecasting models.

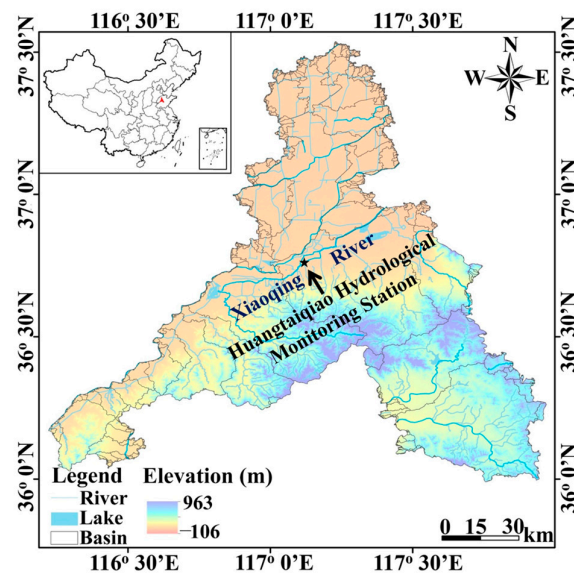


Figure 3. Location map of the Huangtaiqiao Hydrological Monitoring Station.

Table 1. Descriptive statistics of hourly water quality and hydrometeorological variables at the Huangtaiqiao monitoring section.

Variabl	Unit	Missing_Rate	Minimum	Maximum	Mean	Variance	Standard_Deviation	Seasonal_Mean
BLUEGREEN	µg/L	10.53%	0.00	743.25	309.46	18,423.37	135.73	65.2
CHA	µg/L	10.53%	0.00	141.95	11.32	531.60	23.06	16.48
DO	mg/L	6.23%	2.21	14.20	5.97	6.36	2.52	2.07
COD	mg/L	6.23%	0.25	12.48	5.36	3.61	1.90	1.39
ELE	µS/cm	6.23%	561.87	1795.42	1044.31	54,494.24	233.44	105.90
PH	unitless	6.23%	5.25	8.79	7.36	0.58	0.76	0.46
NH4	mg/L	7.81%	0.03	8.03	1.08	2.57	1.60	0.66
TN	mg/L	6.23%	0.05	29.99	8.10	24.25	4.92	0.99
TP	mg/L	6.23%	0.01	0.93	0.10	0.02	0.13	0.06
TUR	NTU	6.23%	0.00	59.90	7.77	60.62	7.79	1.75
TEMP	deg C	6.23%	−0.40	28.98	18.38	55.57	7.45	17.40
TOC	mg/L	6.23%	30.00	49.27	39.96	3.07	1.75	0.61
Flow	m ³ /s	97.89%	17.85	79.06	40.29	253.56	15.92	33.08
Prec	mm/h	97.82%	0.00	68.00	0.07	0.91	0.95	0.23

3.3. Data Preprocessing

To ensure temporal consistency and data quality, a unified preprocessing workflow was applied before model training. First, all monitoring records were aligned to an hourly time index. After temporal alignment, the dataset contained 62,013 hourly records. Continuous missing timestamps were identified in the original water quality series. Most indicators had 3863 missing hours, accounting for 6.23% of the full time series, while

chlorophyll and blue-green algae had 6531 missing hours, accounting for 10.53%. Short gaps were filled using causal forward filling with a maximum duration of six hours. Longer gaps were imputed using the median value from the same month and hour calculated from the training period, thereby reducing the risk of future information leakage.

Second, observations outside physically reasonable or instrument-defined ranges were treated as invalid values. The main screening ranges were as follows: pH, 5–10; dissolved oxygen, 2–15 mg/L; electrical conductivity, 500–3000 $\mu\text{S}/\text{cm}$; CODMn, 0–20 mg/L; ammonia nitrogen, 0–15 mg/L; total nitrogen, 0–40 mg/L; total phosphorus, 0–1 mg/L; turbidity, 0–1000 NTU; water temperature, -8 – 38 $^{\circ}\text{C}$; total organic carbon, 0–50 mg/L; chlorophyll, 0–1000 $\mu\text{g}/\text{L}$; and blue-green algae, ≤ 1000 $\mu\text{g}/\text{L}$. Values outside these ranges were marked as abnormal and subsequently imputed. In total, 48,475 water quality data cells were identified as range-based abnormalities, accounting for 6.51% of all water quality cells.

Third, values below the detection limit were processed using the half-detection-limit substitution method, which is commonly used in environmental monitoring data processing. Specifically, CODMn values below 0.5 mg/L were replaced with 0.25 mg/L, ammonia nitrogen values below 0.05 mg/L were replaced with 0.025 mg/L, total nitrogen values below 0.1 mg/L were replaced with 0.05 mg/L, and total phosphorus values below 0.01 mg/L were replaced with 0.005 mg/L. This procedure affected 25,819 data cells, accounting for 3.47% of the water quality data cells.

Fourth, isolated spikes were screened using a historical window robust criterion. If a value showed an excessively high robust score relative to the median of the previous 72 h and did not belong to a temporally continuous process variation, it was replaced by the corresponding 72 h historical median. Hydrometeorological variables were processed separately according to their data characteristics. Precipitation was treated as an event-based variable: hours without rainfall records were set to zero, while observed rainfall amounts were retained without smoothing. Discharge observations were relatively sparse; therefore, short gaps were filled by causal forward filling, and longer gaps were imputed using the training period month–hour median. Discharge and precipitation were used as auxiliary input variables but were excluded from the per-indicator water quality performance evaluation.

Before model training, all input variables were standardized using StandardScaler (scikit-learn 1.9.0). The scaler was fitted only on the training set and then applied to the validation and test sets. The dataset was split chronologically into training, validation, and test subsets with a ratio of 70%/10%/20%, avoiding random splitting and reducing the risk of temporal information leakage.

3.4. Model Hyperparameter Calibration

To improve the forecasting performance under varying water quality conditions and seasonal fluctuations, this study conducted a focused grid search analysis for the main temporal and structural parameters of the DIFF Transformer model. The calibration was designed around the practical forecasting question of how much historical information should be used and how far ahead the model can reliably predict. Therefore, the grid search focused on three key parameters: input sequence length, prediction length, and number of attention layers.

The input sequence length determines the historical information available to the model. Candidate values ranged from 96 h to 288 h. The prediction length controls the forecasting horizon and was evaluated within the range of 48 h to 120 h. The number of attention layers determines model depth and representation capacity, with a candidate range of one to eight layers. For each parameter combination, the model was independently trained and evaluated on the validation set using NSE as the main criterion, with the proportion of samples with $\text{NSE} > 0.5$ used only as an auxiliary stability indicator.

Other training and architectural hyperparameters were kept fixed across experiments to ensure controlled comparison and reduce confounding effects. Specifically, the learning rate was set to 1×10^{-4} , the batch size to 32, the embedding dimension d_{model} to 64, the number of attention heads to two, the feed-forward dimension to 64, and the dropout rate to 0.1. These settings were used consistently across the compared models to maintain comparable model capacity and training conditions. Thus, the grid search should be interpreted as a focused calibration of the temporal forecasting configuration, rather than an exhaustive optimization of all possible hyperparameters.

For the differential attention module, λ_{init} was not treated as an empirical calibration parameter. Following the original Differential Transformer design, λ_{init} is a layer-dependent initialization constant used to initialize the learnable differential balancing coefficient λ . The effective λ was learned during training through its re-parameterized learnable vectors. Therefore, λ_{init} was fixed according to the original formulation to preserve the intended gradient-flow behavior and avoid introducing an additional manually tuned degree of freedom.

As shown in Figure 4, the calibration results showed that longer input sequences generally improved validation performance, indicating that the model benefited from extended historical context. A 288 h input window provided the best balance between historical information and computational cost. Prediction lengths around 48–72 h achieved a practical compromise between short-term accuracy and medium-term forecasting capability. Increasing model depth beyond the optimal range did not consistently improve validation performance but increased computational cost. Therefore, the final experimental setting used an input length of 288 h, a prediction length of 48 h, and one attention layer for the comparative validation experiments.

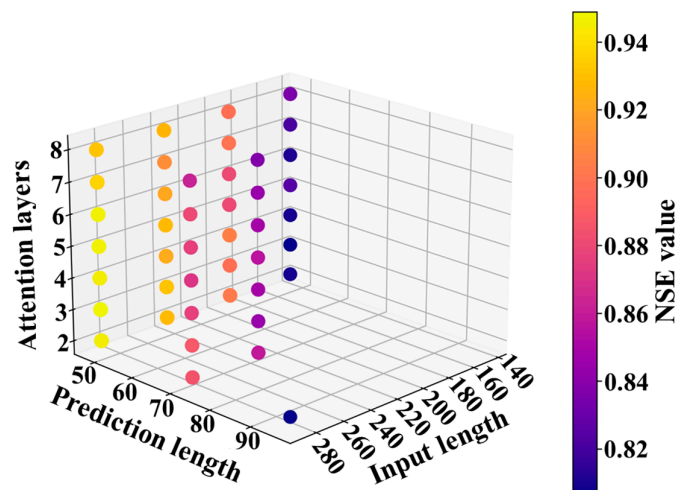


Figure 4. Performance evaluation of the DIFF Transformer for aquatic ecological health prediction in the Xiaoqing River.

4. Model Validation and Mechanism Interpretation

4.1. Comparative Experiment Design

To evaluate the predictive performance of the proposed model, five models were compared under the same data partition and forecasting horizon: Persistence, LSTM, GRU, Transformer, and DIFF Transformer. The dataset contained 14 input variables, including water quality indicators and hydrometeorological variables. As mentioned above, the data were split chronologically into training, validation, and test sets, with a ratio of 7:1:2,

respectively. The standardization parameters were fitted only on the training set to avoid information leakage.

All learning-based models used an input length of 288 h and predicted the following 48 h. The batch size was 32, the learning rate was 1×10^{-4} , and MSE was used as the training loss. For Transformer and DIFF Transformer, the main hyperparameters were set to $d_{\text{model}} = 64$, $n_{\text{heads}} = 2$, $e_{\text{layers}} = 1$, $d_{\text{layers}} = 1$, and $d_{\text{ff}} = 64$. These values were explicitly specified during the experiments. The LSTM and GRU baselines used one recurrent layer with hidden size 64, followed by a linear projection layer to produce the 48-step multi-variable forecast. The Persistence model was included as a naive baseline, where the future value was assumed to be equal to the last observed value. This baseline is necessary because hourly water quality series usually have strong short-term autocorrelation.

Model performance was evaluated using NSE, MSE, MAE, and RMSE. Global metrics were used to compare overall error levels, while per-indicator metrics were calculated to evaluate the stability of each model across different water quality variables. Flow and precipitation were treated as auxiliary inputs and were excluded from the per-indicator water quality comparison. Macro-averaged NSE was calculated by first computing NSE for each indicator and then averaging across indicators, so that each variable contributed equally to the final score.

4.2. Training and Validation Loss Analysis

To evaluate the convergence behavior and potential overfitting risk of the proposed model, we conducted a training and validation loss analysis for DIFF Transformer. As shown in Figure 5, the training loss decreases rapidly during the first few epochs and then converges to a stable level. The validation loss also decreases initially and reaches its minimum at epoch 11, after which it fluctuates within a narrow range. This indicates that the DIFF Transformer training is stable and convergent, with no evident severe overfitting, although a persistent gap between training and validation losses suggests moderate generalization discrepancy.

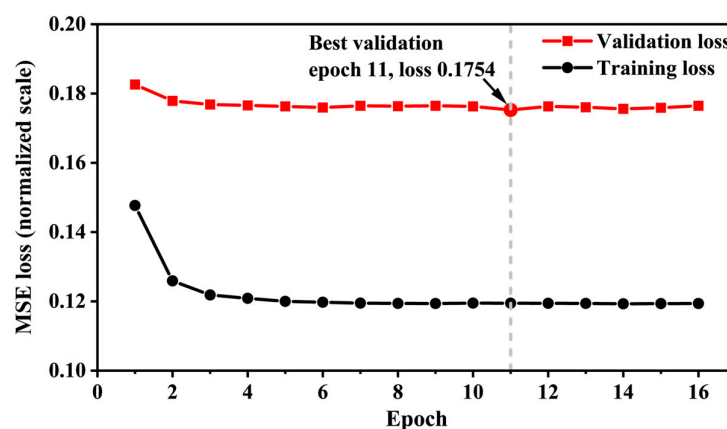


Figure 5. Training and validation loss curves of DIFF Transformer. The loss is computed as MSE in the normalized data space, and the best validation loss is obtained at epoch 11.

4.3. Global Performance Comparison

The global comparison shown in Figure 6 indicates that DIFF Transformer achieved the highest macro-average NSE, indicating the best balanced predictive skill across indicators. Persistence obtained the lowest MSE, MAE, and RMSE, confirming that the hourly series has strong short-term continuity and that the naive baseline is highly competitive.

However, among learning-based models, DIFF Transformer clearly outperformed LSTM, GRU, and Transformer in NSE and error-based metrics.

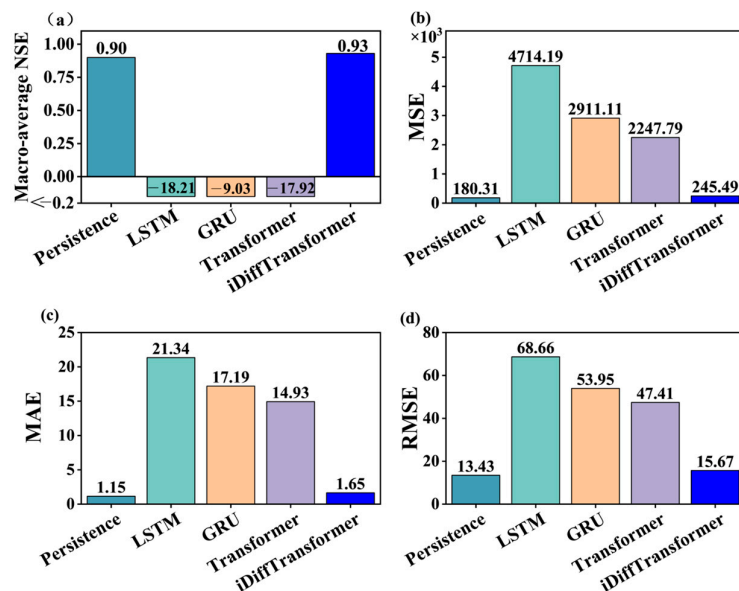


Figure 6. Global performance comparison of different forecasting models. The four subplots report (a) macro-average NSE, (b) MSE, (c) MAE, and (d) RMSE, respectively. For NSE, larger values indicate better predictive performance, whereas for MSE, MAE, and RMSE, smaller values are better. Negative NSE values are compressed in the plot to emphasize positive model performance, while the original values are retained as text annotations. The DIFF Transformer achieves the highest macro-average NSE among all models and shows substantially lower error than LSTM, GRU, and Transformer, although the persistence baseline remains competitive in absolute error metrics.

4.4. Variable-Wise Forecasting Performance Comparison

The per-indicator comparison shown in Figure 7 indicates that DIFF Transformer achieved NSE values above 0.95 for 11 of the 12 water quality indicators, with TOC being the main exception. The relatively low NSE of TOC is mainly caused by its low temporal variability and near-constant values in many samples after preprocessing. Since NSE normalizes the squared error by the variance of the observed series, a small observational variance makes NSE highly sensitive to even small absolute prediction errors. Therefore, the low TOC NSE does not necessarily indicate large absolute prediction errors, but rather reflects the limited variability of TOC in the test set. Electrical conductivity showed relatively large absolute errors because of its larger numerical scale, although its NSE remained high.

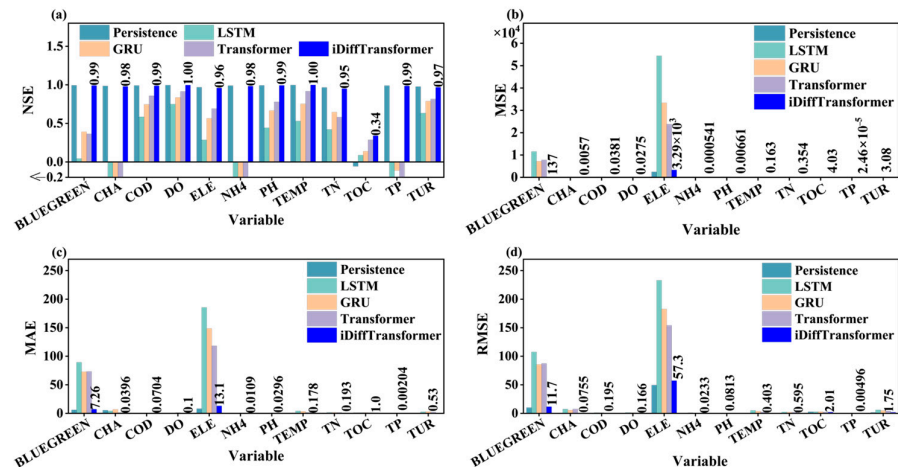


Figure 7. Variable-wise forecasting performance of different models. The four panels show (a) NSE, (b) MSE, (c) MAE, and (d) RMSE for each water quality variable, excluding precipitation and flow. NSE values below -0.2 are compressed for visualization. The blue bars denote DIFF Transformer, and their exact metric values are annotated in the MSE, MAE, and RMSE panels. Higher NSE and lower MSE, MAE, and RMSE indicate better performance.

Overall, these experiments support the conclusion that DIFF Transformer provides the best balanced performance among the learning-based models. At the same time, the strong Persistence baseline indicates that the results should be interpreted with multiple metrics rather than relying only on NSE threshold statistics.

4.5. Biological Indicators Time Series Prediction

Figure 8 presents the representative time series prediction results for blue-green algae, chlorophyll, dissolved oxygen, and turbidity. It can be observed that the DIFF Transformer effectively tracks the dynamic variation characteristics of each indicator on an hourly scale. For biological indicators such as blue-green algae and chlorophyll, the model not only accurately captures the overall trend but also maintains synchronous responses to peak periods during most fluctuation phases, demonstrating the model's strong ability to capture temporal features associated with biological processes. For physicochemical indicators such as dissolved oxygen and turbidity, which fluctuate frequently over short time scales, the prediction results still maintain good consistency with the observed sequences, indicating that the differential attention mechanism plays a positive role in suppressing high-frequency noise and enhancing effective signals.

It should be noted that during certain short-term periods of sharp fluctuations, the prediction results exhibit a certain degree of deviation from the observed values. Such deviations mainly occur during intervals with rapid changes in the indicator values, which may be related to sudden external inputs or local variations in hydrodynamic conditions. Nevertheless, the model as a whole maintains correct judgment regarding the direction of change and the main trends, demonstrating high reliability in characterizing the evolution of water quality status.

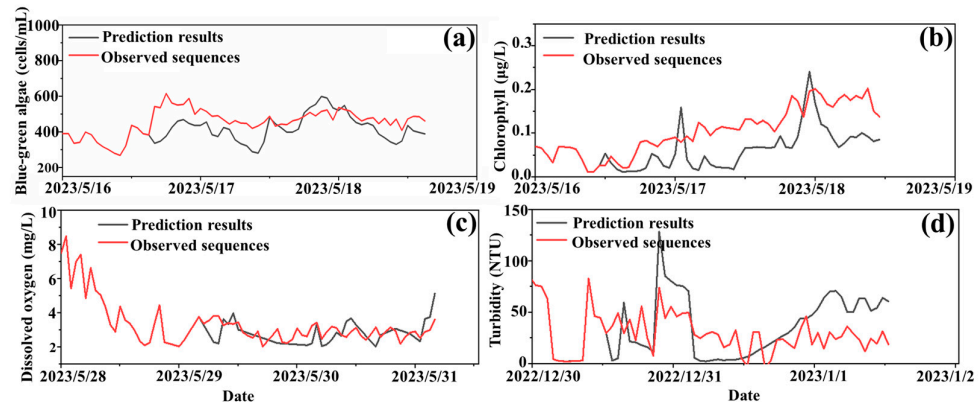


Figure 8. Representative time series prediction results for (a) blue-green algae, (b) chlorophyll, (c) dissolved oxygen, and (d) turbidity.

4.6. Nutrient and Organic Pollution Indicators Time Series Prediction

The prediction results for nutrient and organic pollution indicators, including total phosphorus, ammonia nitrogen, permanganate index, and TOC, are shown in Figure 9. The DIFF Transformer demonstrates relatively stable fitting performance on these indicators, with low overall error levels. These indicators typically exhibit relatively stable background variations in the monitoring sequence, with periodic increases occurring during specific time periods. The model can effectively identify and reproduce these variation characteristics, reflecting its efficient utilization of long-term dependency information.

During certain peak phases, the predicted curve exhibits a degree of smoothing compared to the observed values, indicating that the model tends to capture the overall trend rather than instantaneous extremes under conditions of extreme fluctuation. This characteristic helps, to some extent, to reduce the influence of random noise on the prediction results but also suggests that there remains room for further improvement in the refined characterization of extreme pollution events. Overall, the DIFF Transformer demonstrates good stability and consistency across most time periods, meeting the practical requirements for predicting key indicators of aquatic ecosystem health.

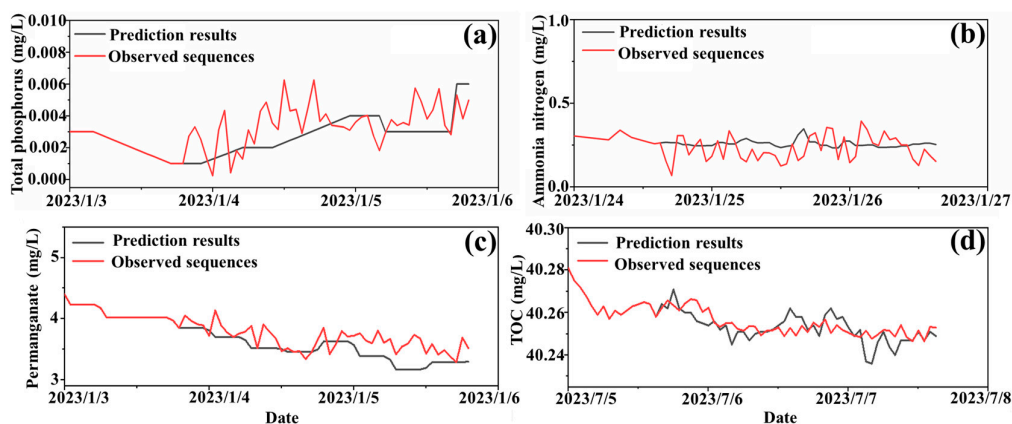


Figure 9. The representative time series prediction results for (a) total phosphorus, (b) ammonia nitrogen, (c) permanganate, and (d) TOC.

4.7. Interpretation of the Differential Attention Mechanism

To directly verify the contribution of the differential attention component, we conducted an architecture-controlled ablation experiment. Specifically, we compared two variants with identical Transformer backbones, training settings, input variables, and prediction horizons: one using standard FullAttention and the other using the proposed differential attention. As shown in Figure 10, the residual time series and power spectrum provide representative evidence that DiffAttention suppresses high-frequency residual fluctuations. Compared with FullAttention, DiffAttention yields smoother residual trajectories and substantially lower residual spectral power in the high-frequency band for both NH4 and TP samples.

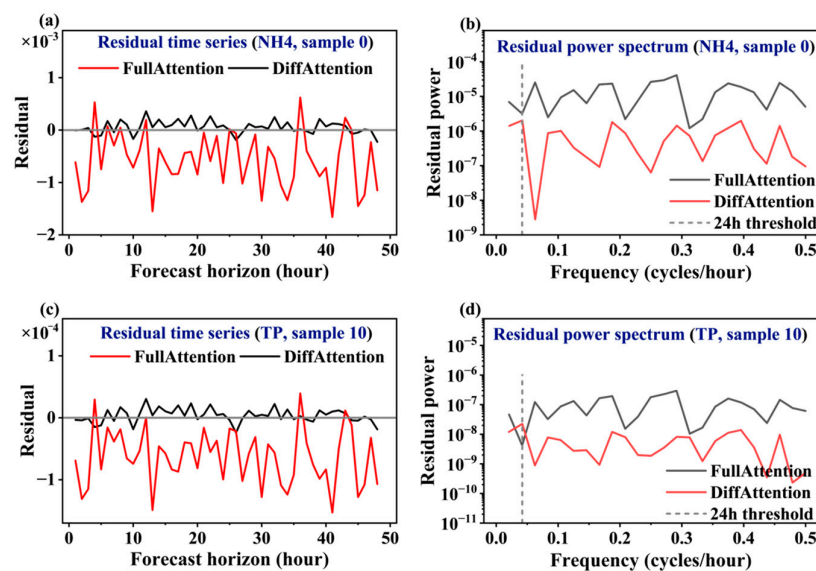


Figure 10. Residual-based noise suppression comparison between FullAttention and DiffAttention on representative test samples. Panels (a,c) show the residual time series for NH4 sample 0 and TP sample 10, respectively. Panels (b,d) show the corresponding residual power spectra. The vertical dashed line denotes the 24 h frequency threshold. Compared with FullAttention, DiffAttention produces residuals with smaller amplitudes and lower spectral power, especially in the high-frequency range, indicating reduced short-term fluctuation errors.

5. Discussion

5.1. Differential Attention Improves Balanced Multi-Indicator Forecasting

Hourly water quality monitoring data are challenging to forecast because they simultaneously reflect long-term seasonal variation, short-term hydrological disturbances, sensor noise, and complex interactions among biological, chemical, and physical variables. These characteristics make it difficult for a single model to maintain both high accuracy and stable performance across multiple indicators. To address this challenge, this study developed a DIFF Transformer-based multivariate forecasting framework for hourly water quality and algal indicators at the Huangtaiqiao monitoring section of the Xiaoqing River. The results show that DIFF Transformer effectively reconstructed the dominant temporal patterns of most indicators, including algal variables, nutrients, oxygen-related variables, physicochemical properties, and organic pollution indicators. Compared with LSTM, GRU, and the standard Transformer, DIFF Transformer achieved more balanced performance across variables and obtained the highest macro-averaged NSE among the learning-based models rather than being concentrated in only a few highly predictable

variables. Therefore, its practical value lies in improving balanced multi-indicator forecasting among learning-based models and providing a flexible framework for exploiting multivariate dependencies in complex hourly water quality time series.

5.2. Interpretation of the Differential Attention Mechanism

The differential attention mechanism is designed to subtract two attention distributions, thereby reducing redundant or unstable attention responses. In the residual analysis, DIFF Transformer produced smoother residual trajectories and lower residual power in high-frequency bands than the corresponding FullAttention model under the same architecture. This provides evidence that the differential mechanism can suppress isolated high-frequency residual fluctuations.

This finding is particularly relevant for automatic water quality monitoring data, which often contain measurement disturbances, sensor instability, and short-term irregular fluctuations. By reducing unstable high-frequency residual components, differential attention can improve the robustness of trend reconstruction and reduce the influence of irrelevant short-term perturbations.

However, the interpretation of “noise suppression” requires caution; short-term fluctuations in environmental monitoring data should not be automatically treated as noise. In environmental systems, not all high-frequency or local fluctuations are noise. Abrupt variations may represent real ecological or hydrological processes, such as storm runoff responses, intermittent pollutant inputs, rapid algal growth, or local hydrodynamic changes. Therefore, the role of differential attention should be understood as suppressing isolated and redundant fluctuations rather than eliminating all local variability. The model is more effective in reconstructing dominant temporal patterns than in reproducing all short-duration peaks.

5.3. Implications for Water Quality and Aquatic Ecological Assessment

Continuous prediction of key water quality and algal indicators can provide useful support for aquatic ecological assessment. Compared with low-frequency manual monitoring, hourly forecasting can help characterize short-term trends and provide more continuous information on ecological proxy indicators such as algal biomass, oxygen conditions, nutrient levels, turbidity, and organic pollution.

The results suggest that DIFF Transformer is suitable as a data-driven component for routine trend tracking and multi-indicator forecasting. Its strength lies in providing stable predictions across several indicators and reconstructing the main temporal evolution of water quality conditions. Such predictions may support dynamic assessment of aquatic ecological status and provide auxiliary information for water-environment management.

At the same time, its application to early warning should be carefully defined. The model may smooth short-duration peaks during abrupt fluctuations, which limits its direct use as a standalone warning tool for algal blooms or sudden pollution events. For operational warning, the model should be combined with event-oriented modules, hydrological forcing information, threshold-based detection, and expert quality control rules. In its current form, DIFF Transformer is better suited for continuous trend prediction than for independent extreme event detection.

5.4. Limitations and Future Work

Several limitations remain. First, the study is based on one monitoring section. Although the dataset spans multiple years, the spatial generalization ability of the model has not been fully evaluated. Multi-section validation and cross-basin testing are needed to determine whether the learned temporal dependencies are transferable to other hydrological and ecological settings.

Second, the model evaluation focuses mainly on general forecasting metrics, including NSE, MSE, MAE, and RMSE. These metrics are appropriate for overall predictive performance, but they do not fully characterize extreme event forecasting. Future studies should include event-oriented metrics, such as peak magnitude error, peak timing error, exceedance event hit rate, false alarm rate, and high-percentile prediction error.

Third, the preprocessing strategy may influence model behavior. Although invalid observations were screened using physical ranges, instrument limits, and temporal continuity criteria, the distinction between measurement artifacts and true environmental disturbances is not always straightforward. Future work should combine statistical anomaly detection with hydrological forcing, cross-variable consistency, and monitoring quality control records to better preserve ecologically meaningful abrupt changes.

6. Conclusions

Accurate forecasting of hourly water quality dynamics remains challenging because monitoring series simultaneously contain long-term temporal dependence, short-term disturbances, measurement uncertainty, and heterogeneous interactions among biological, chemical, and physical indicators. To address this problem, this study developed a DIFF Transformer-based multivariate forecasting framework and evaluated it using multi-year hourly observations from the Huangtaiqiao monitoring section of the Xiaoqing River. The principal finding is that differential attention improves balanced forecasting across multiple water quality and algal indicators rather than merely enhancing the prediction of a few highly predictable variables. Compared with LSTM, GRU, and the standard Transformer, DIFF Transformer achieved the highest macro-averaged NSE among the learning-based models and reproduced the dominant temporal patterns of algal, nutrient, oxygen-related, physicochemical, and organic pollution indicators with greater consistency across variables. Residual and frequency domain analyses further indicated that differential attention reduced isolated and unstable high-frequency residual components, thereby improving the robustness of temporal-pattern reconstruction. This effect should be interpreted as the suppression of redundant fluctuations rather than the removal of all short-term variability, because abrupt changes may represent genuine hydrological or ecological events. Overall, DIFF Transformer provides a flexible data-driven approach for continuous multi-indicator trend prediction and can support dynamic water quality assessment and routine aquatic ecological monitoring. However, its tendency to smooth some short-duration peaks limits its direct use as a standalone detector of algal blooms or sudden pollution events. Future research should therefore validate the framework across multiple monitoring sections and watersheds, evaluate extreme event performance using event-oriented metrics, examine sensitivity to alternative preprocessing strategies, and integrate hydrological forcing, quality control information, and physical or ecological constraints to improve spatial transferability and the representation of ecologically meaningful abrupt changes.

Author Contributions: Y.L. and Y.W.: investigation and writing—original draft. Z.W.: investigation. K.L.: formal analysis, data curation. Z.Z.: investigation. S.W.: investigation. N.L.: conceptualization and editing. J.M.: methodology, writing—review and editing. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (No. 52270083), Research Project on the Spa-tio-temporal Evolution Patterns of Aquatic Ecology in the Xiaoqing River Basin, Jinan City (No. 37000000025001720240149), and the National Key R&D Program of China (Grant No. 2024YFC3810900).

Data Availability Statement: The source code used in this study is publicly available on GitHub: https://github.com/jfmjfm/DIFFTransformer_WQ (accessed on 2 August 2026). The repository contains the implementation of the DIFF Transformer model, baseline models, data preprocessing scripts, model training scripts, and evaluation scripts. The data are subject to a confidentiality agreement. Further inquiries can be directed to the corresponding authors.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Kutyla, S.; Kolada, A.; LawniczakMalinska, A. Hydromorphological pressure explains the status of macrophytes and phytoplankton less effectively than eutrophication but contributes to water quality deterioration. *Water Res.* **2025**, *268*, 122669.
2. Barats, A.; Renac, C.; Garrido-Hoyos, S.; GonzalezPerez, B.; Garcia-Mendoza, K.; Esteller-Alberich, M.V.; Jara-Marini, M.E.; Aguilar-Chavez, A. Assessment of the water quality in the coastal Yaqui valley (Mexico): Implications for human health and ecological risks. *Environ. Res.* **2024**, *264*, 120275.
3. Aguilera, P.A.; Frenich, A.G.; Torres, J.A.; Castro, H.; Vidal, J.L.M.; Canton, M. Application of the kohonen neural network in coastal water management: Methodological development for the assessment and prediction of water quality. *Water Res.* **2001**, *35*, 40534062.
4. Sun, Y.; Wang, D.; Li, L.; Ning, R.; Yu, S.; Gao, N. Application of remote sensing technology in water quality monitoring: From traditional approaches to artificial intelligence. *Water Res.* **2024**, *267*, 122546.
5. Ismaiel, M.M.S.; Piercey Normore, M.D. Cooperative antioxidative defense of the blue green alga *Arthrospira* (Spirulina) platensis under oxidative stress imposed by exogenous application of hydrogen peroxide. *Environ. Pollut.* **2024**, *341*, 123002.
6. Imtiaz, F.; Rashid, J.; Kumar, R.; Eniola, J.O.; Barakat, M.A.E.; Xu, M. Recent advances in visible light driven inactivation of bloom forming blue green algae using novel nanocomposites: Mechanism, efficiency and fabrication approaches. *Environ. Res.* **2024**, *248*, 118251.
7. Pang, H.; Ben, Y.; Cao, Y.; Qu, S.; Hu, C. Time series based machine learning for forecasting multivariate water quality in full scale drinking water treatment with various reagent dosages. *Water Res.* **2025**, *268*, 122777.
8. Prasad, D.V.V.; Venkataramana, L.Y.; Kumar, P.S.; Prasannamedha, G.; Harshana, S.; Srividya, S.J.; Harrinei, K.; Indraganti, S. Analysis and prediction of water quality using deep learning and auto deep learning techniques. *Sci. Total Environ.* **2022**, *821*, 153311.
9. Ao, X.; Gong, Y.; He, A. A Review of Time Series Prediction Models Based on Deep Learning. *IEEE Access* **2025**, *13*, 153696–153712.
10. Granata, F.; Di Nunno, F.; Modoni, G. Hybrid Machine Learning Models for Soil Saturated Conductivity Prediction. *Water* **2022**, *14*, 1729.
11. Sha, J.; Li, X.; Zhang, M.; Wang, Z.-L. Comparison of Forecasting Models for Real-Time Monitoring of Water Quality Parameters Based on Hybrid Deep Learning Neural Networks. *Water* **2021**, *13*, 1547.
12. Ni, J.; Liu, R.; Li, Y.; Tang, G.; Shi, P. An Improved Transfer Learning Model for Cyanobacterial Bloom Concentration Prediction. *Water* **2022**, *14*, 1300.
13. Rathore, W.U.A.; Ni, J.; Ke, C.; Xie, Y. BloomSense: Integrating Automated Buoy Systems and AI to Monitor and Predict Harmful Algal Blooms. *Water* **2025**, *17*, 1691.
14. Huang, H.; Tian, Y.; Tao, Z. Multirule combination prediction of compositional data time series based on multivariate fuzzy time series model and its application. *Expert Syst. Appl.* **2024**, *238*, 121966.
15. Fang, H.; Li, T.; Xian, H. Comparative Analysis of Machine/Deep Learning Models for Single-Step and Multi-Step Forecasting in River Water Quality Time Series. *Water* **2025**, *17*, 1866.
16. Jiang, M.; Ding, K.; Chen, X.; Cui, L.; Zhang, J.; Yang, Z.; Cang, Y.; Cao, S. Research on timeseries based and similarity search based methods for PV power prediction. *Energy Convers. Manag.* **2024**, *308*, 118391.
17. Bui, D.T.; Khosravi, K.; Tiefenbacher, J.; Nguyen, H.; Kazakis, N. Improving prediction of water quality indices using novel hybrid machine learning algorithms. *Sci. Total Environ.* **2020**, *721*, 137612.
18. Niknam, A.R.R.; Goodarzi, M.R.; Barzegar, R. Integrating Entropy-based Water Quality Index and Deep Learning Forecasting for Uncertainty-aware Assessment. *Earth Syst. Environ.* **2026**, 1–24. <https://doi.org/10.1007/s41748-026-01110-w>.
19. Ali, A.J.; Ahmed, A.A. Advancing river water quality prediction: A comparative assessment of deep learning models for dissolved oxygen forecasting. *Int. J. River Basin Manag.* **2026**, 1–22. <https://doi.org/10.1080/15715124.2026.2626322>.

20. Wai, K.P.; Koo, C.H.; Huang, Y.F.; Chong, W.C. Enhancing water quality index forecasting: A novel approach integrating signal decomposition and deep learning models for multi-step ahead predictions. *Ain Shams Eng. J.* **2026**, *17*, 104133.
21. Saeidinia, M.; Hafshejani, L.D.; Shahsavari, M. A parsimonious hybrid model: Integrating wavelet neural networks and deep learning for water quality forecasting in Southern Iran. *J. Water Process Eng.* **2026**, *83*, 109478.
22. Niknam, A.R.R.; Sabaghzadeh, M.; Barzkar, A.; Shishebori, D. Comparing ARIMA and various deep learning models for long-term water quality index forecasting in Dez River, Iran. *Environ. Sci. Pollut. Res. Int.* **2025**, *32*, 10206–10222.
23. Lu, P.; Zhang, J.; Al-Mahmood, A.; Dung, H.Q.; Yunus, Z.M. Novel deep learning models for accurate time series prediction of phyococyanin concentration in water management. *J. Water Process Eng.* **2025**, *74*, 107783.
24. Memarzadeh, G.; Keynia, F. Solar power generation forecasting by a new hybrid cascaded extreme learning method with maximum relevance interaction gain feature selection. *Energy Convers. Manag.* **2023**, *298*, 117763.
25. Waqas, M.; Humphries, U.W. A critical review of RNN and LSTM variants in hydrological time series predictions. *MethodsX* **2024**, *13*, 102946.
26. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. In Proceedings of the 31st Annual Conference on Neural Information Processing Systems (NIPS), Long Beach, CA, USA, 4–9 December 2017.
27. Osman, E.G.A.; Otaibi, F.A. Integrating deep learning and econometrics for stock price prediction: A comprehensive comparison of LSTM, transformers, and traditional time series models. *Mach. Learn. Appl.* **2025**, *22*, 100730.
28. Wang, Z.; Shi, B.; Osmond, P.; Zhang, K. Multimodal deep-learning-driven water-quality forecasting and scenario assessment: Shifting from reactive monitoring to proactive basin management. *J. Hydrol.* **2026**, *675*, 135554.
29. Li, Q.; Zhong, R.; Du, X.; Du, Y. TransUNetCD: A Hybrid Transformer Network for Change Detection in Optical Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 119.
30. Chen, H.; Qi, Z.; Shi, Z. Remote Sensing Image Change Detection with Transformers. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 114.
31. Ding, Y.; Jia, M.; Miao, Q.; Cao, Y. A novel time–frequency Transformer based on self–attention mechanism and its application in fault diagnosis of rolling bearings. *Mech. Syst. Signal Process.* **2022**, *168*, 108616.
32. Ye, T.; Dong, L.; Xia, Y.; Sun, Y.; Zhu, Y.; Huang, G.; Wei, F. Differential Transformer. *arXiv* **2024**, arXiv:2410.05258. <https://doi.org/10.48550/arXiv.2410.05258>.
33. Zhou, C.; Che, C.; Wang, P.; Zhang, Q. Difformer: A dynamic self-differential transformer for new energy power autoregressive prediction. *Knowl.-Based Syst.* **2023**, *281*, 111061.
34. Saleh, T.; Weng, X.; Holail, S.; Hao, C.; Xia, G.S. DAMNet: Flood detection from SAR imagery using differential attention metric-based vision transformers. *ISPRS J. Photogramm. Remote Sens.* **2024**, *212*, 440453.
35. HJ 91.2—2022; Technical Specifications for Surface Water Environmental Quality Monitoring. Ministry of Ecology and Environment: Beijing, China, 2022.
36. HJ 915.3—2024; Technical Specifications for Operation and Maintenance of Surface Water Quality Automated Monitoring Station (Conventional Five Parameters, COD_{Mn}, NH₃-N, TP, TN). Ministry of Ecology and Environment: Beijing, China, 2024.
37. SL 21-2015; Specifications for Precipitation Observations. Ministry of Water Resources of the People’s Republic of China: Beijing, China, 2015.
38. GB 50179-2015; Code for Liquid Flow Measurement in Open Channels. China Planning Press: Beijing, China, 2015.
39. GB/T 50138-2010; Standard for Stage Observation. China Planning Press: Beijing, China, 2010.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.