Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Application of PatchTST Model in Time Series Prediction of Dissolved Oxygen in Stratified Coastal Waters of Hong Kong

Yuhang Wang
Shandong University of Science and Technology
Jing Lv
jissrain@163.com

Shandong University of Science and Technology
Zhengbao Li
Shandong University of Science and Technology
Xiangfeng Kong
Qilu University of Technology (Shandong Academy of Sciences)

Article

Keywords: Hong Kong coastal waters, dissolved oxygen, PatchTST, time series forecasting, stratified water body modeling

Posted Date: March 4th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8806649/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Application of PatchTST Model in Time Series Prediction of Dissolved Oxygen in Stratified Coastal Waters of Hong Kong

Yuhang Wang[1], Jing Lv[1,*], Zhengbao Li[1], Xiangfeng Kong[2]

[1]College of Marine Science and Engineering, Shandong University of Science and Technology, Qingdao 266000, China
[2]State Key Laboratory of Physical Oceanography, Institute of Oceanographic Instrumentation, Qilu University of Technology (Shandong Academy of Sciences)， Jinan 250353, China
[*]corresponding.author: jissrain@163.com

Abstract

Dissolved oxygen (DO) is a critical water quality indicator for assessing the health of marine ecosystems, and its accurate prediction is of great significance for environmental protection and risk early-warning in coastal waters. Based on 20 years of monitoring data from Hong Kong waters, this study constructs a DO time series prediction model using PatchTST and, for the first time, proposes an independent modeling strategy stratified by water depth (middle layer / bottom layer) to enhance predictive adaptability and model generalization capability. The results show that at the representative station SM3, the PatchTST model achieves $R^2 = 0.77$ and RMSE = 0.85 in middle-layer prediction, and $R^2 = 0.80$ in bottom-layer prediction. Overall performance improves by approximately 20%–40% compared to traditional machine learning models, with greater stability observed under the stratified modeling framework. Subsequently, model interpretation via Permutation Importance reveals that water temperature, salinity, historical DO, and seasonal features are the core drivers of DO dynamics, with bottom-layer water showing significantly stronger dependence on historical DO. This study validates the cross-spatial prediction capability of PatchTST across multiple water layers and sites, and proposes a DO prediction framework integrating stratified modeling and feature interpretability, providing technical reference for intelligent monitoring and ecological early-warning in coastal waters.

Keywords: Hong Kong coastal waters, dissolved oxygen, PatchTST, time series forecasting, stratified water body modeling

Introduction

Dissolved Oxygen (DO) is a key indicator for evaluating the health of marine ecosystems, habitat suitability for aquatic organisms, and the self-purification capacity of water bodies. In nearshore marine environments, DO concentration is governed by a combination of physical transport, biogeochemical processes, and air-water gas exchange. With global climate warming and intensified coastal urbanization, marine eutrophication and seasonal hypoxia have become increasingly severe, posing serious threats to fish survival and ecological balance[1].

Hong Kong is a typical subtropical coastal city whose nearshore waters are influenced by monsoon climate, tidal dynamics, Pearl River discharge, and anthropogenic activities, resulting in pronounced seasonality and vertical stratification in water quality evolution. Therefore, establishing a high-accuracy

DO prediction model holds significant application value for intelligent monitoring and ecological early-warning in Hong Kong's coastal waters.

In recent years, data-driven approaches—owing to their robustness in capturing nonlinear water quality dynamics—have gradually become the mainstream for DO prediction. In the era of traditional machine learning, Liu et al. (2024)[2] employed input variable selection combined with hybrid machine learning algorithms to accelerate the identification of key influencing factors. However, such models struggle to capture deep spatiotemporal dependencies in long-sequence, complex-trend water quality data. With the advancement of deep learning, Recurrent Neural Networks (RNNs) and their variant Long Short-Term Memory networks (LSTMs) have been widely applied, yet they still face limitations in gradient stability under ultra-long sequences. Othman and Wu (2025)[3] proposed an RF-LSTM hybrid model that optimized feature inputs using Random Forest; however, its ability to capture long-term dependencies remains challenged in the face of highly dynamic nearshore environments. Shadkani et al. (2024) investigated a novel hybrid temporal pattern attention deep neural network, leveraging attention mechanisms to better capture DO fluctuations in river systems—a sign that models are evolving from pure time-series modeling toward spatiotemporal feature fusion[4].

For highly dynamic estuarine environments, Chen et al. (2025)[5] used an advanced interpretable ensemble machine learning model to study the driving mechanisms of water quality parameters on DO. Although this work achieved breakthroughs in interpretability, there remains room for improvement in evaluating model stability across multi-site and multi-layer tasks. Additionally, Ali et al. (2025)[6] demonstrated the application of multi-dimensional data fusion in coastal DO inversion through an XGB-2DCNN model. At national or continental scales, Zhi et al. (2021)[7] proved the feasibility of deep learning models in complex cross-regional environments. However, for regions like Hong Kong with pronounced vertical stratification, most existing studies focus only on surface or single-layer waters, neglecting the essential mechanistic differences between middle and bottom layers[7]. Li et al. (2024)[8] noted in their Yangtze River Estuary study that different water depths experience distinct proportions of physical transport versus biochemical oxygen consumption—providing crucial theoretical support for our proposed stratified independent modeling strategy.

To address efficiency and accuracy challenges in long-sequence prediction, Transformer architectures based on self-attention mechanisms have become a research hotspot. Wang et al. (2023) enhanced DO concentration prediction by integrating Temporal Transformers with graph embedding techniques using multi-site meteorological data[9]. Subsequently, Nie et al. (2023)[10] proposed the PatchTST model, which slices time series into patches, significantly reducing computational complexity while preserving local semantic information to strengthen long-dependency modeling. Liu et al. (2024) argued that innovative prediction-fusion models better simulate complex biochemical reactions[11].

Improving model interpretability has become a major requirement in the intersection of environmental science and artificial intelligence. Wang et al. (2024)[12] developed an interpretable deep learning model based on multi-source data fusion, primarily addressing uncertainty in multi-step predictions. From the perspective of feature attribution, Khan et al. (2025) showed that using XAI methods for feature selection enhances model transparency compared to traditional algorithms. To tackle the "black-box" nature of deep neural networks[13], Petrovska et al. (2025)[14] explored the performance of SHAP and ensemble explainers in complex prediction tasks. Meanwhile, Integrated Gradients (IG), proposed by Sundararajan et al., quantifies the cumulative contribution of input features to model outputs, offering rigorous mathematical axiomatic support for understanding complex ecosystem mechanisms[15].

In summary, despite significant progress in data-driven algorithms, research on stratified modeling and stability evaluation remains scarce for regions like Hong Kong's coastal waters—characterized by multiple river inflows and clear vertical heterogeneity. This study leverages 20 years of long-term monitoring data from the Hong Kong Environmental Protection Department (EPD) to establish a stratified prediction framework based on PatchTST. The innovations include: (1) adopting a water-depth-stratified (middle/bottom layer) independent modeling approach; (2) combining Permutation Importance and Integrated Gradients to identify dominant DO drivers at each layer; and (3) conducting cross-spatial validation across open sea, enclosed bays, and other site types to systematically evaluate the robustness and generalization of the PatchTST framework in complex hydrodynamic environments.

Research Methodology

Data Source and Preprocessing

Data Source

The data used in this study originate from the automatic water quality monitoring system of the Hong Kong Environmental Protection Department (EPD), covering monthly observations from January 2005 to December 2024 at three representative monitoring stations (SM3, VM1, WM1) in Hong Kong's coastal waters, totaling 2,121 valid samples. These stations represent three typical aquatic environments: open sea, nearshore estuary, and enclosed bay, encompassing varying hydrodynamic conditions and gradients of human influence, thus ensuring strong regional representativeness.

Each observation record includes 12 environmental variables and 3 temporal features, with a sampling frequency of once per month. The environmental variables are: water temperature (°C), suspended solids (SS, mg/L), silica (SiO_2, mg/L), Secchi depth (SD, m), salinity (‰), pH, dissolved oxygen (DO, mg/L—target variable), chlorophyll-a (Chl-a, μg/L), five-day biochemical oxygen demand (BOD_5, mg/L), turbidity (NTU), total phosphorus (TP, mg/L), and total nitrogen (TN, mg/L). Temporal features include: seasonal index (1–4 for spring, summer, autumn, winter) and periodic encodings of month (1–12) using sine and cosine functions (i.e., $\sin(2\pi m/12)$ and $\cos(2\pi m/12)$, where mm is the month), effectively capturing seasonal patterns in water quality parameters.

DO, as a key indicator of water self-purification capacity and ecosystem health, is designated as the primary prediction target. This multi-dimensional, long-term dataset provides a solid foundation for building high-accuracy DO prediction models and enables in-depth analysis of water quality dynamics under diverse environmental conditions.

Data Cleaning and Missing Value Handling

During preprocessing, missing values were filled using feature-wise median imputation, which avoids bias from extreme values and ensures data robustness. After imputation, all data were sorted chronologically to maintain time series continuity. Non-numeric entries were converted to NaN and appropriately filled. All observational data underwent manual review and validation to ensure accuracy and reliability.

Data Standardization

To eliminate scale differences among features and accelerate model training, all input features were standardized using StandardScaler, transforming each feature to zero mean and unit variance. The target variable (DO) was standardized separately and de-standardized during inference to recover its original physical units.

Dataset Splitting

The dataset was split chronologically into 80% training and 20% validation sets per station. The training set was used for model learning and parameter optimization, while the validation set assessed predictive performance and mitigated overfitting.

Stratification by Water Depth

To avoid distributional confounding from differing water conditions, data were grouped by monitoring station and water depth layer (middle and bottom), with each layer treated as an independent training unit [3,16]. This strategy minimizes environmental heterogeneity-induced data bias across layers.

Notably, surface (epipelagic) layer data were excluded because surface waters are heavily influenced by external climatic factors (e.g., precipitation, solar radiation), leading to dynamics markedly different from deeper layers. Middle and bottom layers exhibit more stable, periodic temporal patterns, better reflecting long-term trends. Given the absence of comprehensive climate data, this study focuses solely on middle and bottom layers to prevent inter-layer interference. This grouping enhances data quality and improves model generalization across diverse aquatic environments.

Site Selection and Dataset Overview

Three representative stations were selected to test model generalizability under varying hydrodynamic regimes:

SM3: Located in central open waters, minimally affected by terrestrial inputs, serving as the baseline for regional background conditions.

VM1: Situated in the northeastern nearshore Sai Kung area, strongly influenced by river discharge and coastal human activities.

WM1: Located in a southern enclosed bay with weak water exchange and pronounced seasonal stratification.

These sites collectively represent open sea, nearshore estuary, and semi-enclosed bay environments, enabling comprehensive evaluation of PatchTST's spatial transferability and layer-specific performance.



Fig. 1. Monitoring station locations.

Correlation Analysis and Feature Selection

Feature Correlation Analysis

Using SM3 data (2005–2024, middle + bottom layers combined), we conducted multi-scale association analyses between 12 environmental variables and DO:

Pearson correlation coefficients to assess linear relationships;

Cross-Correlation Function (CCF) to examine lagged dynamics between temperature and DO;

Scatter plots and hexagonal density plots to visualize nonlinear responses.

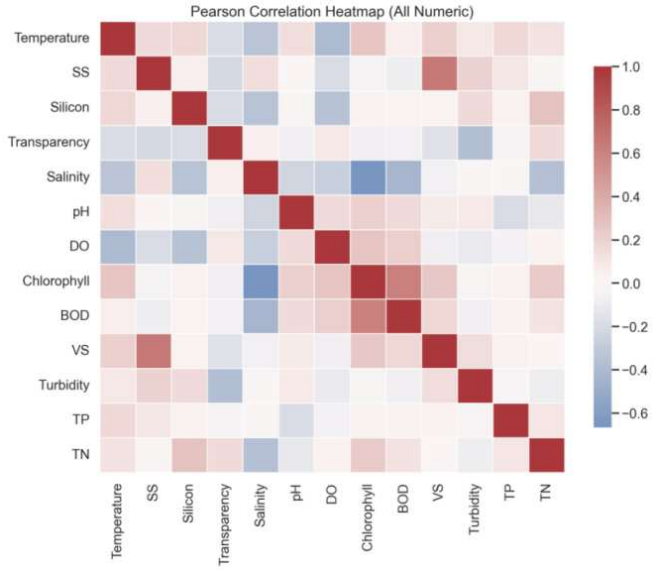


Fig. 2. Heatmap of Pearson correlation coefficients among key physicochemical parameters in the aquatic environment.

The Pearson heatmap (Fig. 2) reveals strong negative correlation between temperature and DO. TN, TP, and Chl-a show positive correlations, indicating a "nutrient–primary production–DO" cascade, especially prominent in enclosed bays.

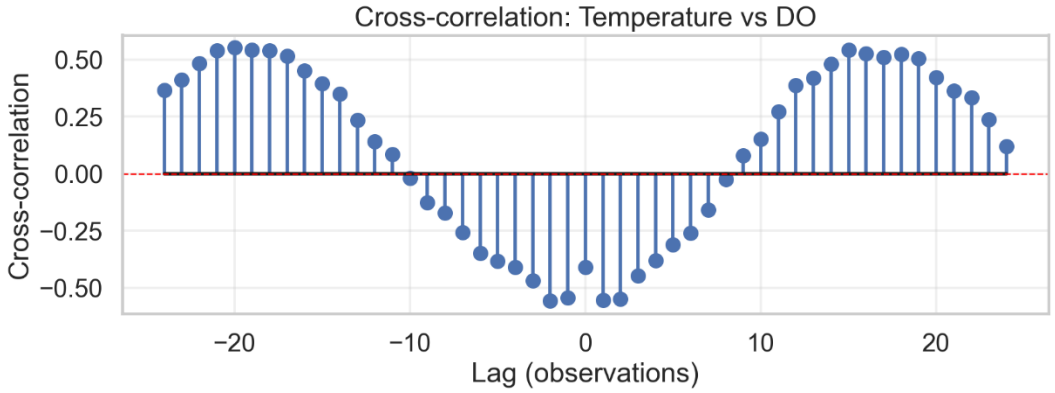


Fig. 3. Cross-correlation analysis between Water Temperature and Dissolved Oxygen (DO).

CCF analysis (Fig. 3) shows persistent negative correlation between temperature and DO across lags, peaking at 0–2 months, suggesting a sustained thermal response effect linked to mixing processes.

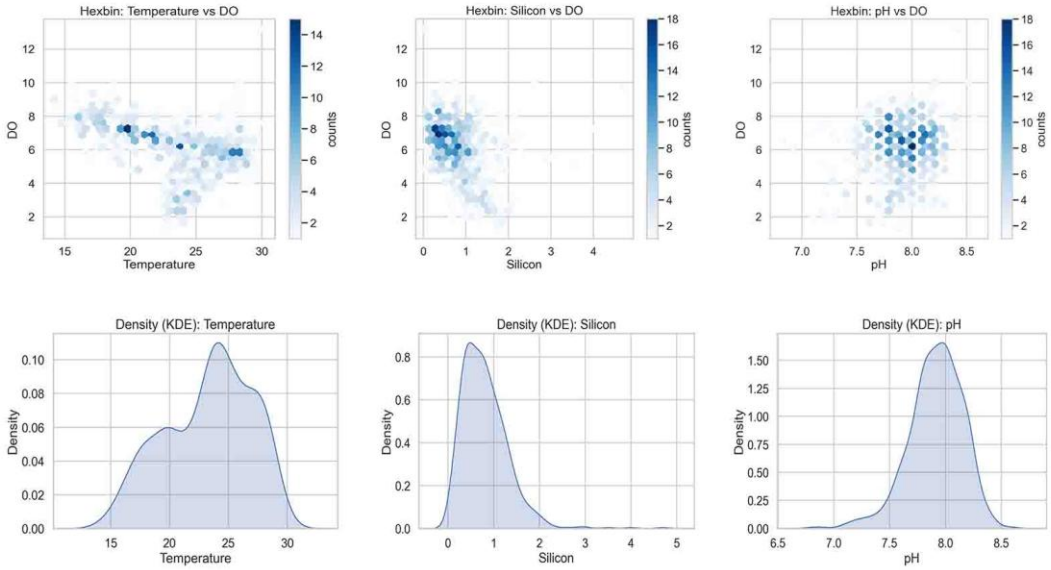


Fig. 4. Multivariate analysis of water quality parameters depicting bivariate relationships and univariate distributions.

Scatter plots (Fig. 4) indicate weak linear correlations ($|r| < 0.4$) between most variables and DO, confirming that DO is regulated by multiple interacting processes: thermal solubility control, algal photosynthesis/respiration, eutrophication feedback, and hydrodynamic exchange. For example:

Temperature vs. DO shows a bimodal seasonal pattern;

Chl-a exhibits a threshold effect: positive at low levels (enhancing DO via photosynthesis), but diminishing or negative at high levels (due to respiration/decomposition).

These nonlinear, heteroscedastic characteristics justify the use of PatchTST, capable of capturing complex temporal dependencies.

Feature Selection

Initial screening retained variables with $|r| \geq 0.1$, including 11 physicochemical features plus temporal encodings (Month_sin/cos, Season). Despite weak linear correlations, many variables may exert nonlinear or lagged effects.

Following stratified modeling principles, feature sets were tailored per layer:

Feature	Physical / chemical meaning	Expected relationship with DO	Middle layer (Base Val Loss ≈ 0.313)	Bottom layer (Base Val Loss ≈ 0.275)	Interpretation / selection decision

Temperature (°C)	Controls oxygen solubility and water density	Negative (↑T → ↓DO)	*Retained*	*Retained*	Primary physical driver of DO variation in both layers
Month_sin / Month_cos	Represents annual seasonal cycle	Periodic / cyclic	*Retained* (dominant)	*Retained* (dominant)	Captures long-term seasonal rhythm of thermal and mixing processes
Season	Seasonal categorical variable	Periodic	*Retained*	*Retained*	Secondary seasonal component correlated with temperature
Salinity (psu)	Density and stratification strength	Negative	*Retained (weak)*	*Retained*	Stronger influence near the bottom due to stable stratification
DO (mg/L) (lag term)	Historical oxygen persistence	Positive	*Weak*	*Retained*	Bottom layer exhibits stronger oxygen storage inertia
pH	Carbonate balance / biological activity	Slightly positive	*Retained (minor)*	*Retained (minor)*	Minor chemical contribution, correlated with temperature and photosynthesis
Transparency (m)	Light penetration	Positive	*Removed*	*Removed*	Little relevance below photic zone
Turbidity (NTU)	Suspended particles limiting light	Negative	*Removed*	*Removed*	Weak and redundant with suspended solids
Suspended Solids (mg/L)	Particle concentration	Negative	*Removed*	*Removed*	Similar to turbidity; limited predictive value
Chlorophyll-a (µg/L)	Algal biomass / photosynthesis	Positive in surface	*Removed*	*Removed*	Biological effect disappears at these depths
BOD_5 (mg/L)	Organic matter oxidation	Negative	*Removed*	*Removed*	Weak correlation, noisy monthly variability
Total Nitrogen (mg/L)	Nutrient enrichment	Ambiguous	*Optional*	*Removed*	Minimal influence; may indicate eutrophication trend

Total Phosphorus (mg/L)	Limiting nutrient	Ambiguous	*Optional*	*Retained (minor)*	Consistent but weak contribution to bottom oxygen variation
Silicon (mg/L)	Diatom nutrient	Weak positive	*Removed*	*Removed*	Insignificant across layers

Table 1. Summary of feature importance and selection results for dissolved oxygen (DO) prediction models in the middle and bottom water layers.

Key insight: Physical drivers (temperature, seasonality) dominate in both layers. Bottom layer additionally relies on salinity (stratification) and historical DO (oxygen storage inertia), while biological/nutrient variables show negligible importance.

Final selected features: Temperature, Month_sin/cos, Season, Salinity, and lagged DO.

Feature Contribution Analysis

Permutation Importance is a model-agnostic method for evaluating feature importance. Its core idea is to randomly shuffle the values of each input feature one at a time and observe the resulting change in model performance (e.g., validation loss), denoted as ΔLoss. A larger ΔLoss indicates that the feature contributes more significantly to the model's predictive capability. To mitigate the influence of randomness, each feature is typically permuted multiple times, and the mean ΔLoss is reported along with its standard deviation as a measure of uncertainty, thereby quantifying each feature's contribution to prediction [13,17,18]. This approach is widely applied across various machine learning models, effectively identifying the most influential features while maintaining computational efficiency and model independence [17,18](Khan et al., 2025; Mi et al., 2021). Furthermore, to address interpretation bias caused by dependencies among features, researchers have proposed improved methods such as conditional permutation to yield more accurate importance estimates [19](Molnar et al., 2020). Permutation Importance has also been leveraged to enhance both model performance and interpretability, aiding researchers in feature selection and model optimization for complex tasks[13,17,18].

To quantitatively assess the relative influence of each water quality feature on dissolved oxygen (DO) prediction, we sequentially shuffled the values of each input feature and computed the corresponding change in validation loss (ΔLoss). A larger ΔLoss implies that disrupting the feature leads to a greater degradation in model performance, indicating higher predictive relevance. To reduce stochastic variability, each feature was permuted five times, and the average ΔLoss was calculated, with the standard deviation serving as an indicator of uncertainty. This method is model-agnostic and has been extensively used for interpreting neural network predictions.

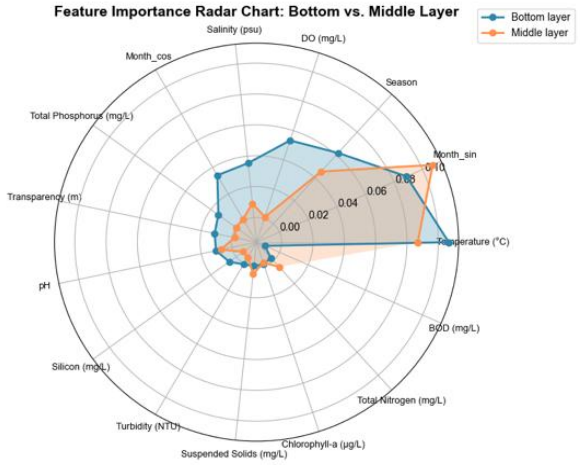


Fig. 5. Feature importance radar chart comparing bottom and middle water layers. The blue and orange polygons represent the permutation importance (ΔLoss) of each feature in the bottom and middle layers, respectively. Higher values indicate greater contribution to dissolved oxygen prediction.

Figure 5 (radar chart) intuitively illustrates the distinct patterns of feature importance between bottom and middle water layers. The blue polygon (bottom layer) extends notably outward along axes such as temperature, Month_sin, and season, indicating that these features dominate DO prediction in the bottom layer. In contrast, the orange polygon (middle layer) reaches a similar height for Month_sin but contracts significantly along dimensions like salinity and historical DO, reflecting weaker dependence on vertical stratification and memory effects in the middle layer. This visualization effectively reveals the structural differences in the controlling mechanisms between water layers.

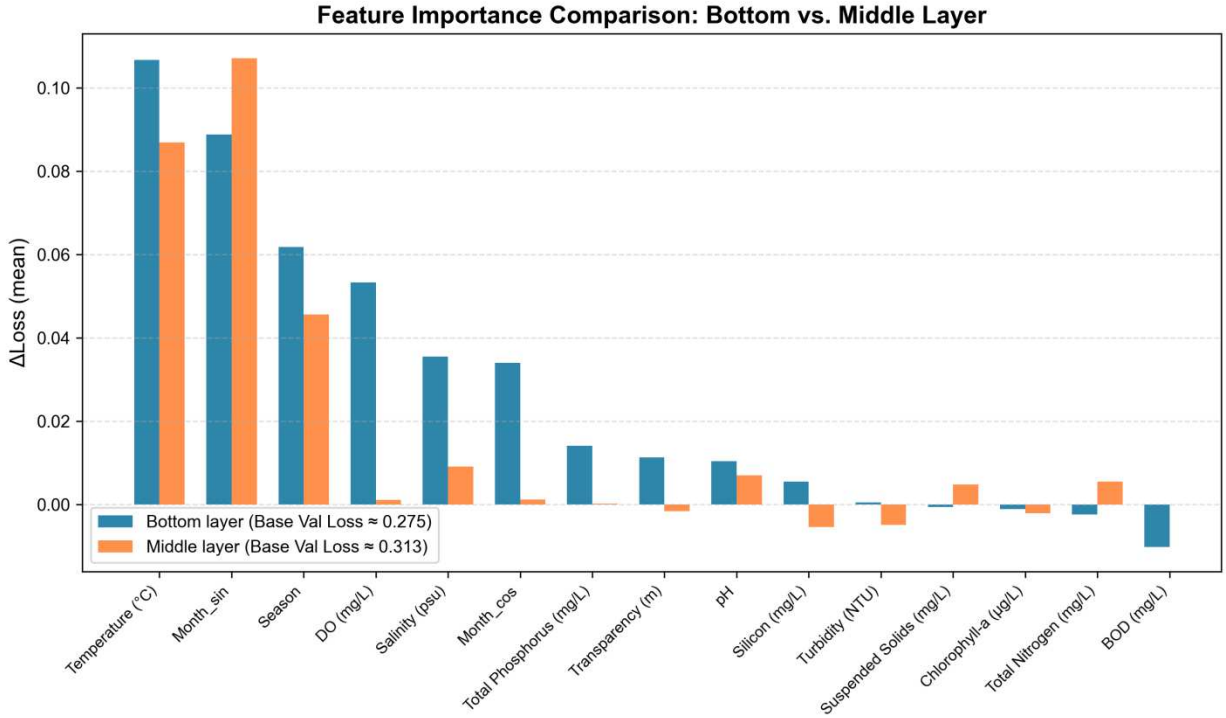


Fig. 6. Comparative analysis of feature importance between bottom and middle water layers. The bar chart shows the permutation importance (ΔLoss) of each feature in the bottom (blue) and middle (orange) layers, sorted by their average importance across both layers. Error bars represent the standard deviation over five shuffles.

Figure 6 provides a quantitative comparison of permutation importance across features. Features are sorted along the x-axis by their average importance in both layers, clearly showing that temperature (ΔLoss ≈ 0.107) and Month_sin (≈0.107) are the strongest shared drivers. Notably, salinity and historical DO exhibit significant contributions in the bottom layer (blue bars, ΔLoss > 0.035) but are nearly zero or even slightly negative in the middle layer (orange bars). This confirms that bottom-layer DO is governed by water column stratification and oxygen storage inertia, whereas the middle layer responds more strongly to short-term thermal dynamics and periodic climatic signals. The low standard deviations reflected in the error bars further indicate high result stability.

Figures 5 and 6, together with Tables 2 and 3, present the main feature contribution results for bottom and middle layer samples at station SM3. In the bottom layer (Base Validation Loss ≈ 0.275), the dominant drivers are temperature (0.107), Month_sin (0.089), and season (0.062), highlighting the control of DO dynamics by thermal and seasonal mechanisms. Additionally, historical DO (0.053) and salinity (0.036) show significant contributions, suggesting that bottom-layer DO exhibits clear oxygen storage inertia and stratification dependence. In contrast, in the middle layer (Base Validation Loss ≈ 0.313), although Month_sin (0.107) and temperature (0.087) remain primary, the influence of salinity and historical DO is markedly reduced. This indicates that middle-layer oxygen distribution is primarily driven by short-term thermal processes and annual cycles, while vertical mixing and ventilation processes are more complex, leading to higher prediction uncertainty.

Moreover, to verify the robustness of these findings, we also performed gradient-based attribution using Integrated Gradients. The resulting feature importance rankings were largely consistent with those from permutation importance, with discrepancies mainly observed in the contribution values at specific time steps. The convergence of results from these two complementary methods corroborates that the model is highly sensitive to physical features (e.g., temperature, seasonal terms) but shows weak dependence on nutrient and biological indicators (e.g., chlorophyll-a, BOD_5).

Feature	ΔLoss (mean)	ΔLoss (std)	Interpretation
Temperature (°C)	0.1067	0.0170	Thermal control of oxygen solubility
Month_sin	0.0888	0.0130	Annual seasonal cycle (phase signal)
Season	0.0618	0.0175	Seasonal stratification and mixing strength
DO (mg/L)	0.0533	0.0210	Historical oxygen inertia (storage memory)
Salinity (psu)	0.0355	0.0177	Density stratification strength
Month_cos	0.0340	0.0131	Complementary seasonal phase term
Total Phosphorus (mg/L)	0.0141	0.0066	Weak nutrient influence
Transparency (m)	0.0113	0.0065	Light penetration proxy

Feature	ΔLoss (mean)	ΔLoss (std)	Interpretation
pH	0.0104	0.0080	Weak chemical balance effect
Silicon (mg/L)	0.0055	0.0098	Minor or noise feature
Turbidity (NTU)	0.0005	0.0035	Negligible contribution
Suspended Solids (mg/L)	–0.0006	0.0063	Weak/negative signal
Chlorophyll-a (μg/L)	–0.0011	0.0030	Biological indicator with little effect at this depth
Total Nitrogen (mg/L)	–0.0024	0.0044	Nutrient signal not significant
BOD_5 (mg/L)	–0.0102	0.0048	Possible noise or redundant feature

Table 2. Feature importance for the bottom water layer (Base Val Loss ≈ 0.275).

Feature	ΔLoss (mean)	ΔLoss (std)	Interpretation
Month_sin	0.1071	0.0223	Annual cycle, dominant driver
Temperature (°C)	0.0869	0.0343	Thermal influence on oxygen solubility
Season	0.0456	0.0207	Seasonal pattern, moderate impact
Salinity (psu)	0.0091	0.0047	Weak stratification effect
pH	0.0070	0.0048	Minor chemical variation
Total Nitrogen (mg/L)	0.0055	0.0027	Weak nutrient contribution
Suspended Solids (mg/L)	0.0048	0.0043	Slight particulate influence
Month_cos	0.0012	0.0100	Minor seasonal complement
DO (mg/L)	0.0011	0.0129	Limited temporal persistence
Total Phosphorus (mg/L)	0.0002	0.0017	Negligible
Transparency (m)	–0.0016	0.0018	Possible noise
Chlorophyll-a (μg/L)	–0.0021	0.0021	Weak biological signal
BOD_5 (mg/L)	–0.0041	0.0014	Slight negative contribution
Turbidity (NTU)	–0.0049	0.0038	Possible redundant indicator
Silicon (mg/L)	–0.0054	0.0056	Insignificant

Table 3. Feature importance for the middle water layer (Base Val Loss ≈ 0.313).

Model Architecture

This study employs the PatchTST model, which integrates the Transformer architecture with patching (slicing) techniques to effectively capture both long-term dependencies and short-term dynamics in time series data[2,20].

Forward Process

The forward propagation process of the PatchTST model starts with the input data. The data first passes through the Patch Embedding layer for slicing, dividing the time series into multiple temporal segments:

$$X_{patch}:$$

$$X_{patch}=\text{Unfold}(X,\text{patch_len},\text{stride}) \quad (1)$$

where $X \in R^{B\times L\times C}$ is the input time series data, B is the batch size, Lis the sequence length, C is the feature dimension, patch_len is the length of each temporal segment, and stride is the sliding step size. Each temporal segment is embedded into a low-dimensional space via a linear transformation and added with positional encoding to preserve temporal order information[15]:

$$X^{'}=X_{patch}+P \quad (2)$$

Subsequently, the embedded temporal segments are fed into the Transformer encoder for processing. After computation through the self-attention mechanism and feed-forward network, the output representation is transformed via a linear mapping to generate future prediction values $\hat{y}$[21-23]:

$$\hat{y}=W_h\cdot H+b_h \quad (3)$$

where W_h is the weight matrix, b_h is the bias term, and $\hat{y}$ is the final output of the model.

Patching

In the PatchTST model, the time series data is first divided into multiple temporal segments. Assuming the input sequence has length L , it is sliced by a sliding window into several segments of length patch_len, with a sliding step size of stride. Each temporal segment contains features from patch_lenpatch_len consecutive time steps, denoted as:

$$X_{patch}=[x_1,x_2,...,x_n] \quad (4)$$

where $n=\frac{L\text{-patch_len}}{\text{stride}}+1$is the number of segments. Each temporal segment $R^{d_{model}}$ is mapped into an embedding space and processed via a linear transformation:

$$x_i=\text{Linear}(x_i)\forall\, i \quad (5)$$

This step, by embedding each temporal segment, enables the model to extract local features within a small time window while also capturing global temporal dependencies among segments through the Transformer encoder.

Transformer Encoder

The Transformer encoder captures global dependencies in time series data through multi-head self-attention mechanisms and feed-forward neural networks. The self-attention mechanism allows the model to establish connections between arbitrary positions in the sequence, thereby capturing long-term

dependencies. By stacking multiple self-attention layers, the encoder enables the model to learn more complex temporal patterns.

Loss Function

The model uses Mean Squared Error (MSE) as the loss function[6] to measure the difference between predicted and actual values. MSE optimizes model parameters by minimizing prediction error, ensuring that the prediction results are as close as possible to the true observed values.

$$\mathrm{MSE}=\frac{1}{N}\sum_{i=1}^{N}(y_i-\hat{y}_i)^2 \quad (6)$$

where y_i is the true value, $\hat{y}_i$ is the predicted value, and N is the number of samples.

Instance Normalization

Instance normalization standardizes the features of each temporal segment so that their mean is zero and variance is one. This helps improve training stability and accelerate convergence. Through instance normalization, the model can reduce bias among features and ensure a more balanced learning process.

$$\hat{x}_i=\frac{x_i-\mu}{\sigma} \quad (7)$$

where μ and σ are the mean and standard deviation of the sample, respectively.

Evaluation Metrics

In this study, we use commonly adopted regression metrics—including Mean Squared Error (MSE), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE), and the coefficient of determination (R^2)—to evaluate the predictive performance of the PatchTST model. These metrics comprehensively reflect the model's prediction accuracy and generalization capability from multiple perspectives, including error magnitude, relative deviation, and goodness-of-fit.

Integrated Gradients (IG) is a gradient attribution method for deep neural networks that quantifies the contribution of each input feature to the model's prediction by measuring "how gradients accumulate during the transition from a baseline input to the actual input" [24]. Its basic approach is to select a reference input (baseline) that contains almost no information, construct a path between the baseline and the input sample to be explained (the classical IG uses a straight-line path), and then perform multi-point sampling and integration of the gradients of the model output with respect to the input along this path. This yields an attribution score for each feature dimension, such that the sum of all feature attributions exactly equals the change in the model output relative to the baseline. This satisfies a set of important axioms, including sensitivity and implementation invariance, and thus possesses strong theoretical soundness among path-integral methods[24,25]. Compared with traditional methods that compute gradients at only a single point, Integrated Gradients accumulates distributed gradients along the path, which can to some extent alleviate the instability in interpretation caused by gradient saturation. Moreover, it requires no modification to the original network architecture and can be implemented using only multiple standard forward/backward passes. Therefore, it has been widely applied in various tasks such as image, text, point cloud, and credit risk modeling[10,15,24].

Results and Discussion

Model Performance and Generalization Analysis

To systematically evaluate the effectiveness and robustness of the constructed PatchTST dissolved oxygen (DO) prediction model, this study conducted comprehensive tests on mid-layer and bottom-layer water data from three representative nearshore monitoring stations in Hong Kong: SM3, VM1, and WM1. Based on the depth-stratified independent modeling strategy and feature selection results described in Chapter 2, we first compared the predictive performance of PatchTST against multiple mainstream machine learning models at the core station SM3. Subsequently, we validated the model's cross-site generalization capability across all three stations to examine its adaptability and stability under different hydrodynamic environments—open sea, nearshore estuary, and enclosed bay. The detailed performance comparisons and generalization analyses are presented below.

Mid-Layer Model Results (Results for SM3 – Middle Layer)

Table 4 shows the performance comparison of various models on the mid-layer data at the SM3 monitoring station. The PatchTST model demonstrates superior performance across all evaluation metrics.

Model	MAE	RMSE	MAPE(%)	R^2	Correlation coeffcient
PatchTST	0.6745	0.8461	14.97	0.7686	0.8844
Lasso	0.8343	1.1004	16.21	0.4477	0.6878
Random Forest	0.8033	1.1177	15.78	0.4301	0.6826
LightGBM	0.8652	1.1496	16.58	0.3971	0.6685
Gradient Boosting	0.8079	1.1552	14.70	0.3913	0.6737
XGBoost	0.8331	1.1576	15.32	0.3888	0.6634
KNN	0.8839	1.1617	18.90	0.3844	0.6655
SVR	1.0335	1.3657	23.44	0.1491	0.5558

Table 4. SM3 Performance Comparison of Mid-level Models at Monitoring Stations.

As shown in Table 4, PatchTST outperforms all other models across all evaluation metrics, with particularly notable advantages in MAE, RMSE, and R^2. The R^2 value of PatchTST is 0.7686, and its correlation coefficient reaches 0.8844, indicating that it effectively captures the temporal trends in water quality data. In contrast, traditional regression methods such as Lasso and SVR perform poorly on these two metrics—especially SVR, whose R^2 is only 0.1491.

Bottom-Layer Model Results

Table 5 presents the performance of various models on the bottom-layer data at the SM3 monitoring station. PatchTST again exhibits outstanding results.

Model	MAE	RMSE	MAPE(%)	R^2	Correlation coeffcient
PatchTST	0.6982	0.9085	15.82	0.8034	0.8979

Model	MAE	RMSE	MAPE(%)	R^2	Correlation coeffcient
Lasso	0.8292	1.0268	20.51	0.6328	0.8259
Random Forest	0.7958	1.0429	20.08	0.6212	0.8167
XGBoost	0.8457	1.1260	20.86	0.5585	0.7722
LightGBM	0.8921	1.1261	21.07	0.5584	0.7907
Gradient Boosting	0.8540	1.1470	21.42	0.5419	0.7565
KNN	0.9345	1.2199	24.53	0.4818	0.7147
Decision Tree	1.0798	1.4318	27.70	0.2861	0.6228
SVR	1.0933	1.4412	31.00	0.2767	0.6729

Table 5. SM3 Performance comparison of underlying models at monitoring stations.

As seen in Table 5, PatchTST achieves the best performance in all metrics for bottom-layer data: MAE = 0.6982, RMSE = 0.9085, MAPE = 15.82%, R^2 = 0.8034, and correlation = 0.8979. Notably, its R^2 and correlation coefficient—0.8034 and 0.8979, respectively—demonstrate exceptional capability in predicting bottom-layer water quality dynamics. In comparison, Lasso and Random Forest achieve R^2 values of 0.6328 and 0.6212, respectively, both significantly lower than PatchTST.

Generalization Performance Analysis

To verify the applicability and generalization capability of the PatchTST model under varying monitoring stations and water layer conditions, this study selected several representative stations—including WM1, SM3, and VM1—and predicted DO for both mid- and bottom-layers using the same model framework. By comparing model performance across these sites, we comprehensively assessed its stability and robustness under diverse hydrological and water quality environments (see Table 6).

Station	Layer	MAE	RMSE	Correlation coefficient	Data std.	R^2	Correlation
WM1	Middle	0.5669	0.7748	0.8697	1.5558	0.752	0.8697
WM1	Bottom	0.6961	0.842	0.9044	1.9587	0.8152	0.9044
SM3	Middle	0.6745	0.8461	0.8844	1.7589	0.7686	0.8844
SM3	Bottom	0.6982	0.9085	0.8979	2.049	0.8034	0.8979
VM1	Middle	0.5184	0.6922	0.8667	1.3695	0.7446	0.8667
VM1	Bottom	0.5813	0.8247	0.874	1.6901	0.7619	0.874

Table 6. Model performance metrics for dissolved oxygen (DO) prediction across different monitoring stations and water layers.

Overall Model Performance Summary

The overall results indicate that the PatchTST model achieves high prediction accuracy and stability at most monitoring stations. Particularly in bottom layers, the coefficient of determination (R^2) consistently exceeds 0.70, reaching a maximum of 0.8152 (WM1 bottom), with correlation coefficients concentrated in the range of 0.85–0.90, demonstrating the model's ability to effectively capture the temporal variation patterns of dissolved oxygen (DO). For mid-layers, R^2 is slightly lower (0.74–0.77) but remains stable overall, reflecting the model's strong adaptability to different depth conditions. Cross-site experiments further confirm PatchTST's generalization capability—it maintains consistent prediction accuracy across diverse aquatic environments, especially in regions with relatively stable hydrodynamics and pronounced periodic signals. Compared to traditional RNN or LSTM models, PatchTST leverages time series patch embedding and Transformer-based encoding mechanisms to more effectively capture global dependencies across multiple temporal scales, thereby enhancing predictive performance in cross-site and cross-layer tasks. In summary, PatchTST demonstrates excellent stability and generalization under multi-site, multi-layer conditions, providing reliable technical support for regional-scale water quality time series forecasting and intelligent monitoring.

Analysis and Discussion

Experimental results show that the PatchTST model exhibits significant advantages in water quality time series prediction, consistently outperforming traditional models in terms of R^2, correlation coefficient, MAE, and RMSE for both mid- and bottom-layer data. This superiority primarily stems from its Transformer-based self-attention mechanism, which enables the model to effectively capture long-term dependencies and complex nonlinear trends, and extract key features across multiple temporal scales—thereby improving prediction accuracy and reliability.

In contrast, traditional machine learning models such as Lasso, SVR, Random Forest, and XGBoost show limitations when handling water quality time series characterized by long-term dependencies and complex trends. Linear regression models are constrained by their linear assumptions and struggle to represent nonlinear relationships; tree-based models, while capable of nonlinear modeling, still fall short of PatchTST in capturing long-range temporal dependencies. This indicates that the intrinsic characteristics of time series forecasting tasks decisively influence model selection. In complex water quality time series, PatchTST uniquely satisfies the dual modeling demands of long-term dependency and nonlinear feature representation, leading to significantly better performance than conventional approaches.

Conclusion

This study systematically evaluated the PatchTST model on mid- and bottom-layer water data from multiple monitoring stations (WM1, SM3, VM1) to verify its cross-site generalization capability. Results show that PatchTST maintains stable and high prediction accuracy across diverse hydrological and water quality backgrounds, demonstrating strong robustness and transferability. In bottom layers, the coefficient of determination (R^2) at most stations exceeds 0.70, with a peak of 0.8152 (WM1 bottom), and correlation coefficients cluster within 0.85–0.90. This high predictability primarily arises because the bottom environment is relatively stable, and DO variations are dominated by long-term thermal processes, seasonal stratification, and oxygen storage inertia, resulting in strong regularity and predictability. For mid-layers, R^2 is slightly lower (0.74–0.77), likely influenced by intermittent vertical mixing and spatiotemporal variability in hydrodynamic conditions, yet cross-site performance remains stable with minimal fluctuation.

Compared to traditional RNN or LSTM models, PatchTST—leveraging its Transformer architecture, self-attention mechanism, and multi-scale modeling via patch embedding—demonstrates stronger capabilities in capturing long-term dependencies and modeling global temporal patterns in cross-site tasks, significantly enhancing generalization performance. This highlights its promising potential for application in regional-scale water quality monitoring and prediction systems.

Contributions of This Study

In summary, this study makes four main contributions:
First, we constructed a high-accuracy DO prediction framework based on PatchTST, achieving the first systematic application of PatchTST to monthly-scale DO forecasting in coastal bays. We designed a complete workflow encompassing preprocessing, feature engineering, and model validation. At the core station SM3, R^2 reached 0.77 (mid-layer) and 0.80 (bottom-layer), significantly outperforming traditional methods such as Lasso, Random Forest, XGBoost, and LightGBM. RMSE and MAE were reduced by 20%–35% on average, validating the advantage of Transformers in modeling complex temporal dependencies.

Second, this study systematically revealed the dominant drivers and ecological mechanisms of DO variation in mid- and bottom-layers. We found that bottom-layer DO is primarily controlled by temperature, seasonal terms, and historical DO storage, whereas mid-layer DO is more influenced by short-term thermal processes and annual cyclic modulation. Nutrients and biological processes show low contribution at the monthly scale, indicating that physical processes dominate DO dynamics in semi-enclosed bays.

Third, we proposed a depth-stratified independent modeling strategy, treating data from different depths at the same station as independent training units and applying differentiated feature configurations based on mechanistic differences. This effectively reduces data confounding caused by vertical environmental heterogeneity, enabling consistently high prediction accuracy ($R^2 > 0.74$) across multiple stations and layers.

Fourth, we verified the cross-site generalization capability and robustness of PatchTST. Its performance remains consistently stable across multi-site tasks at WM1, SM3, and VM1, with bottom-layer prediction achieving a peak R^2 of 0.8152. This demonstrates that PatchTST can maintain performance integrity across diverse hydrodynamic scenarios, confirming its potential for deployment in regional-scale water quality forecasting and intelligent monitoring systems.

Summary

This study successfully developed a high-precision dissolved oxygen time series prediction framework based on the PatchTST model and validated its superior predictive performance and strong generalization capability across multiple monitoring stations and water depths in Hong Kong's coastal bays. Through feature importance analysis, we systematically uncovered the relative contributions of key factors—including water temperature, seasonal variables, salinity, and historical DO—to DO dynamics at different depths, deepening our understanding of the spatiotemporal evolution patterns and driving mechanisms of DO in coastal bays.

Moreover, the proposed depth-stratified independent modeling strategy provides an effective method for time series prediction in water bodies exhibiting pronounced vertical stratification. The results further demonstrate that Transformer-based deep learning models hold significant advantages in water quality time series forecasting, effectively capturing long-term dependencies and complex nonlinear dynamics.

Specifically, the PatchTST model not only excels in retrospective prediction at single sites but also exhibits strong robustness and transferability in cross-site validation, laying a solid technical foundation for the development of regional-scale intelligent water quality monitoring and early-warning systems.

Data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Competing interests

The authors declare no competing interests.

Author contributions

Yuhang Wang: Conceptualization, Methodology, Formal analysis, Writing—original draft. Jing Lv: Writing—review & editing, Supervision. Zhengbao Li: Writing—review & editing. Xiangfeng Kong: Data curation, Project administration, Funding acquisition. All authors have read and agreed to the published version of the manuscript.

Funding

No funding.

References

1 Gambín, Á. F., Angelats, E., González, J. S., Miozzo, M. & Dini, P. Sustainable Marine Ecosystems: Deep Learning for Water Quality Assessment and Forecasting. *IEEE Access* **9**, 121344-121365, https://doi.org/10.1109/ACCESS.2021.3109216 (2021).

2 Liu, W. *et al.* Analysis of dissolved oxygen influencing factors and concentration prediction using input variable selection technique: A hybrid machine learning approach. *Journal of Environmental Management* **357**, 120777, https://doi.org/10.1016/j.jenvman.2024.120777 (2024).

3 Othman, Z. & Wu, Y. Dissolved oxygen prediction in rivers based on RF-LSTM model: A case study of Dujiangyan. (Preprints, 2025).

4 Shadkani, S., Hemmatzadeh, Y., Saber, A. & Mohammadi Sergini, M. Enhanced predictive modeling of dissolved oxygen concentrations in riverine systems using novel hybrid temporal pattern attention deep neural networks. *Environ. Res.* **263**, 120015, https://doi.org/10.1016/j.envres.2024.120015 (2024).

5 Chen, X. *et al.* Water quality parameters-based prediction of dissolved oxygen in estuaries using advanced explainable ensemble machine learning. *J. Environ. Manage.* **380**, 125146, https://doi.org/10.1016/j.jenvman.2025.125146 (2025).

6 Ali, A. *et al.* XGB-2D CNN-Based Dissolved Oxygen Inversion for Coastal Water. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **18**, 21289-21304, https://doi.org/10.1109/JSTARS.2025.3598875 (2025).

7 Zhi, W. *et al.* From Hydrometeorology to River Water Quality: Can a Deep Learning Model Predict Dissolved Oxygen at the Continental Scale? *Environmental science & technology* **55**, 2357-2368, https://doi.org/10.1021/acs.est.0c06783 (2021).

8 Li, X. Y., Wang, H., Wang, Y. Q., Zhang, L. J. & Wu, Y. Machine Learning-based Dissolved Oxygen Prediction Modeling and Evaluation in the Yangtze River Estuary. *Huan jing ke xue= Huanjing kexue* **45**, 7123-7133, https://doi.org/10.13227/j.hjkx.202312111 (2024).
9 Wang, H., Zhang, L., Wu, R. & Zhao, H. Enhancing Dissolved Oxygen Concentrations Prediction in Water Bodies: A Temporal Transformer Approach with Multi-Site Meteorological Data Graph Embedding. *Water* **15**, 3029, https://doi.org/10.3390/w15173029 (2023).
10 Alam, M. S., Boardman, J., Huang, X. & Turner, M. in *2022 IEEE International Conference on Big Data (Big Data)*. 4997-5005.
11 Liu, J., Zhang, C., An, D. & Wei, Y. Development and application of an innovative dissolved oxygen prediction fusion model. *Computers and Electronics in Agriculture* **227**, 109496, https://doi.org/10.1016/j.compag.2024.109496 (2024).
12 Wang, Z., Wang, Q., Liu, Z. & Wu, T. A deep learning interpretable model for river dissolved oxygen multi-step and interval prediction based on multi-source data fusion. *Journal of Hydrology* **629**, 130637, https://doi.org/10.1016/j.jhydrol.2024.130637 (2024).
13 Khan, A., Ali, A., Khan, J., Ullah, F. & Faheem, M. Exploring Consistent Feature Selection for Software Fault Prediction: An XAI-Based Model-Agnostic Approach. *IEEE Access* **13**, 75493-75524, https://doi.org/10.1109/ACCESS.2025.3558913 (2025).
14 Petrovska, N., Larrondo-Petrie, M. & Pavlovic, M. in *2025 International Conference on Advanced Machine Learning and Data Science (AMLDS)*. 175-180.
15 Schwegler, M., Müller, C. & Reiterer, A. Integrated Gradients for Feature Assessment in Point Cloud-Based Data Sets. *Algorithms* **16**, 316, https://doi.org/10.3390/a16070316 (2023).
16 Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. (arXiv, 2014).
17 Khan, A., Ali, A., Khan, J., Ullah, F. & Faheem, M. Using Permutation-Based Feature Importance for Improved Machine Learning Model Performance at Reduced Costs. *IEEE Access* **13**, 36421-36435, https://doi.org/10.1109/ACCESS.2025.3544625 (2025).
18 Mi, X., Zou, B., Zou, F. & Hu, J. Permutation-based identification of important biomarkers for complex diseases via machine learning models. *Nat. Commun.* **12**, 3008, https://doi.org/10.1038/s41467-021-22756-2 (2021).
19 Molnar, C., König, G., Bischl, B. & Casalicchio, G. Model-agnostic feature importance and effects with dependent features: a conditional subgroup approach. *Data Min. Knowl. Discovery* **38**, 1-39, https://doi.org/10.1007/s10618-022-00901-9 (2023).
20 Rabalais, N. & Turner, R. Gulf of Mexico hypoxia, A.K.A. "the dead zone". *Annual Review of Ecology and Systematics* **33**, 235-263, https://doi.org/10.1146/annurev.ecolsys.33.010802.150513 (2003).
21 D, G., Rabalais, N., Diaz, R. & J, Z. Evidence for greater oxygen decline rates in the coastal ocean than in the open ocean. *Biogeosci. Discuss.* **7**, https://doi.org/10.5194/bgd-6-9127-2009 (2009).
22 Fennel, K. *et al.* Nitrogen cycling in the Middle Atlantic Bight: Results from a three-dimensional model and implications for the North Atlantic nitrogen budget. *Global Biogeochem. Cycles* **20**, https://doi.org/10.1029/2005GB002456 (2006).
23 Lee, J. H. W., Wu, R. S. S. & Cheung, Y. K. Forecasting of Dissolved Oxygen in Marine Fish Culture Zone. *J. Environ. Eng.* **117**, 816-833, https://doi.org/10.1061/(ASCE)0733-9372(1991)117:6(816) (1991).
24 Sundararajan, M., Taly, A. & Yan, Q. in *International Conference on Machine Learning.*
25 Lundstrom, D. & Razaviyayn, M. Four Axiomatic Characterizations of the Integrated Gradients Attribution Method. *ArXiv* **abs/2306.13753**, https://doi.org/10.48550/arXiv.2306.13753 (2023).