Journal of Water Process Engineering 77 (2025) 108341

Contents lists available at ScienceDirect

Journal of Water Process Engineering

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

Designing advanced feature selection and uncertainty quantification-based deep learning approach to predict chlorophyll-a and water bloom risks in dam reservoir

Akram Seifi [a], Hossien Riahi Madvar [a], Rouhollah Davarpanah [b], Mumtaz Ali [c,e,*], Abdul-Wahab Mashat [d]

[a] Department of Water Science & Engineering, Vali-e-Asr University of Rafsanjan, Iran
[b] Department of Hydrology, Meteorology and Water Management, Institute of Environmental Engineering, Warsaw University of Life Sciences, Nowoursynowska 166, 02-787 Warsaw, Poland
[c] UniSQ College, University of Southern Queensland, QLD 4350, Australia
[d] Department of Meteorology, King Abdulaziz University, Jeddah 21589, Saudi Arabia
[e] Scientific Research Centre, Al-Ayen University, Nasiriyah 64001, Thi-Qar, Iraq

ARTICLE INFO

Editor: Guangming Jiang

Keywords:
Feature selection
Hybrid modeling
Reservoir management
Spatial analysis
Bloom risk detection

ABSTRACT

Predicting water quality indicators accurately is vital for the sustainable management of aquatic ecosystems, particularly in dam reservoirs that are highly vulnerable to environmental phenomena. Dissolved oxygen (DO) and chlorophyll-a (Chl-a) are essential indicators for evaluating ecosystem stability and water quality. In this study, an innovative and robust intelligent framework is designed using integrated uncertainty quantification and feature selection to predict DO, Chl-a, and bloom risk evaluation of dams. First, the individual machine learning and deep learning models, including Extreme Gradient Boosting (XGBoost), Convolutional Neural Networks (CNNs), Gated Recurrent Units (GRUs), Least Square Support Vector Regression (LSSVR), and Multi-Layer Perceptron (MLP) were assessed. Subsequently, the most effective models are then integrated to enhance predictive accuracy. The Boruta Feature Selection Approach (BFSA), Gamma Test, and Shapley Additive Explanations (SHAP) are used to select the most suitable and relevant features. Then the Monte-Carlo simulation is implemented for uncertainty analysis to evaluate the reliability of models' prediction by determining probability distribution functions. The hybrid XGBoost-CNNs achieved the highest performance in terms of $R^2 = 0.923$, RMSE = 0.547 μg/l for Chl-a prediction, and CNNs obtained $R^2 = 0.995$, RMSE = 0.143 ppm for DO prediction. The 95 % Prediction Uncertainty (95PPU) varied from 79.37 to 100, which shows strong predictive reliability. Also, d-factor values lower than 0.77 confirmed the model uncertainty is low. Furthermore, water bloom risk was assessed using the predicted Chl-a concentration. The analysis indicated no risk levels at reservoir depths of 0–5.5 m and 13.5–32 m, while low-risk levels were identified between 5.5 and 13.5 m. The maximum risk probability was 20.66 % when Chl-a concentrations were below 40 μg/l. The results highlight the effectiveness of hybrid artificial intelligence frameworks in enabling real-time water quality monitoring, early detection of harmful algal blooms, and promoting sustainable reservoir management.

1. Introduction

Managing water quality in dam reservoirs is essential for ecological preservation, ensuring public water supply security, and promoting economic and recreational opportunities. These water bodies are essential sources of drinking water and provide habitats for various aquatic species that enhance and sustain ecosystem services and contribute to the economic stability [1,2]. Freshwater ecosystems are facing significant environmental challenges, including algal blooms caused by global warming, industrialization, increasing pollutant trends, and various human activities [3–6]. Dissolved oxygen (DO) and chlorophyll-a (Chl-a) are two key indicators of water quality,

* Corresponding author.
E-mail addresses: a.seifi@vru.ac.ir (A. Seifi), h.riahi@vru.ac.ir (H. Riahi Madvar), Rouhollah_davarpanah@sggw.edu.pl (R. Davarpanah), mumtaz.ali@unisq.edu.au (M. Ali), amashat@kau.edu.sa (A.-W. Mashat).

https://doi.org/10.1016/j.jwpe.2025.108341
Received 19 May 2025; Received in revised form 2 July 2025; Accepted 14 July 2025
Available online 25 July 2025

representing the amount of oxygen available for aquatic life and reflecting the abundance of algae and cyanobacteria, reflecting the reservoir's trophic status and potential eutrophication [7,8]. The occurrence of blooms can pose significant ecological risks, such as lowering oxygen levels in water bodies, and can lead to the death of aquatic organisms, including fish [9].

Monitoring DO and Chl-a is crucial, yet challenging due to their temporal and spatial variability driven by complex interactions among physical, chemical, and biological processes [10]. Conventional monitoring techniques require manual sampling and laboratory testing, which are resource-intensive, time-consuming, and expensive. Moreover, these methods often lack the resolution and scalability required for comprehensive monitoring, particularly in large reservoirs with limited access [11]. Additionally, these conventional approaches fail to capture rapid changes in water quality caused by sudden environmental events, such as storms or nutrient influxes, which can significantly impact DO and Chl-a levels [12]. Therefore, accurate prediction of these water quality parameters is necessary for effective reservoir management, early detection of harmful algal blooms, and mitigation of eutrophication risks [13]. Over the years, the empirical models have been developed to predict Chl-a concentrations and analyze phytoplankton dynamics in natural ecosystems [14]. However, the interactions within algal communities, which are significantly affected by nutrient influx and temperature variations, complicate efforts to make precise approximations [15,16]. Process-based models often face challenges due to their intricate structure, suboptimal predictive accuracy, extensive calibration demands, and limited capacity to integrate dynamic input variables influencing the target outcome [17].

Recent developments in artificial intelligence (AI), particularly machine learning (ML) and deep learning (DL), have significantly enhanced the ability to address these challenges. These AI driven methods are beneficial for examining extensive datasets to detect the complex and nonlinear relationships in water quality parameters. Several studies demonstrated the effectiveness of integrating traditional water quality index (WQI) models with ML techniques for real-time monitoring and management of aquatic ecosystems. Kim et al. [18] suggested that the British Columbia Water Quality Index (BCWQI)-based assessment framework can be effectively replaced by long short-term memory (LSTM) models driven by in-situ measurable parameters. Zare et al. [19] analyzed extensive water quality datasets using ML algorithms to identify key water quality variables (WQVs). The model's robustness was confirmed to support data-driven decision-making in reservoir management. In addition, by applying ML/DL techniques, the DO and Chl-a variables can be predicted with greater accuracy and efficiency as compared to the other computational models, which provides a cost-effective solution for managing water quality [20,21]. For example, Kim et al. [7] showed that Random Forest (RF) is effective in predicting Chl-a concentrations in the Han River basin, highlighting the significance of feature selection for improving model accuracy. Likewise, Tian et al. [22] utilized Convolutional Neural Networks (CNNs) to model temporal variations of Chl-a, demonstrating the capability of DL models to capture complex patterns in time-series data. Huang and Zhang [23] found that nonlinear models, such as Multi-Layer Perceptron (MLP) and Support Vector Regression (SVR), are significantly accurate in predicting Chl-a concentrations compared to linear models. These studies highlight the increasing dependence on data-driven methods for monitoring and managing water quality.

Over the past few decades, remote sensing has become a powerful tool for monitoring water quality and eutrophication in inland and coastal waters. For instance, Mozafari et al. [24] employed MODIS-Aqua Level 3 chlorophyll-a (Chl-a) data to assess the trophic state of the Caspian Sea. Their findings showed acceptable agreement with in situ measurements, reinforcing the utility of satellite data for large-scale water quality assessment. Further advancements have been made through machine learning techniques. Shamloo and Sima [25] evaluated the performance of Landsat-8 (L8) and Sentinel-2 (S2) data, combined with multiple linear regression (MLR) and ANN models, to predict Chl-a in Lake Urmia. Their results indicated that L8-based models outperformed those using S2, with ANN achieving the highest accuracy. Similarly, Mozafari et al. [26] integrated MODIS-Aqua and ERA5 climate data into a generalized additive model (GAM) to predict Chl-a in the Caspian Sea, identifying photosynthetically active radiation (PAR) and sea surface temperature (SST) as key drivers of phytoplankton biomass. Despite these advancements, challenges remain in applying remote sensing and ML/DL methods across different water depths and optical conditions. Fooladi et al. [27] highlighted persistent uncertainties in predicting water quality parameters at varying reservoir depths, emphasizing the need for further refinement of algorithms and validation with in situ data. Expanding on these studies, this research aims to address existing gaps by predicting Chl-a and DO concentrations in the depth of the reservoir.

Hybrid modeling frameworks that integrate various ML and DL algorithms provide a promising opportunity to further enhance prediction accuracy and reliability. Hybrid approaches can leverage the strengths of individual models to overcome their limitations and provide more reliable predictions. A relevant study by Fooladi et al. [27] presented a hybrid clustering technique for predicting DO and Chl-a concentrations in the Wadi Dayqah Dam in Oman. The study used advanced clustering methods to segment the dataset into uniform groups, allowing for precise model training and validation for each cluster. By combining genetic algorithms (GA) with Bayesian Ridge Regression in a hybrid framework, the researchers enhanced prediction accuracy and included uncertainty quantification to evaluate model reliability. This method highlighted the significance of integrating clustering with advanced hybrid models to effectively tackle the spatial and temporal variability present in water quality data. Another study by Abbas et al. [20] introduced a hybrid model combining Long Short-Term Memory (LSTM) networks and CNNs, effectively integrating temporal and spatial features. These studies highlight the evolving nature of ML and DL applications in water quality modeling, with an increasing focus on hybrid approaches that combine the strengths of individual models.

Feature selection is a vital process in predictive modeling, as it ensures that only the most relevant variables are included in the prediction, thereby enhancing its interpretability, accuracy, and computational efficiency. In the context of DO and Chl-a prediction, selecting optimal input variables is essential to accurately represent the underlying environmental and water quality variables that influence these parameters [28]. Advanced feature selection techniques, such as the Boruta Feature Selection Approach (BFSA) and Shapley Additive Explanations (SHAP), have shown great potential in identifying key variables that influence predictions. The BFSA is a robust algorithm that determines the importance of features by comparing them to randomly shuffled shadow features, ensuring only statistically significant variables are retained for predictive modeling [29]. The SHAP framework offers a clear interpretation of how each variable contributes to model predictions, helping to understand the factors influencing DO and Chl-a levels [30,31]. Zhou et al. [31] compared traditional methods like Principal Component Analysis (PCA) with modern approaches such as SHAP and concluded that SHAP offers superior interpretability and performance in identifying critical variables. SHAP provides model-agnostic, interpretable feature importance scores, even for nonlinear interactions. While PCA reduces dimensionality by capturing maximum variance, it assumes linearity and does not inherently rank features by predictive importance [32]. Furthermore, Park et al. [33] highlighted the limitations of Tree-based Feature Importance (Tree-FI) and recommended incorporating SHAP to mitigate variable correlation challenges. The Gamma Test (GT) directly quantify the predictability of input features, identify those with the strongest nonlinear influence on the target variable. Unlike mutual information (MI), BFSA and SHAP account for feature dependencies and model-specific contexts. MI is scalable and robust to noise but evaluates features independently, potentially overlooking synergistic interactions [34]. The selection of BFSA, SHAP, and

GT was based on their strengths in handling nonlinear relationships, interpretability, and robust feature importance assessment, which are key requirements for water quality prediction. These methods separately applied and compared their selected features using XGBoost performance, ensuring the most robust subset was retained for modeling. In the current study, the selection of BFSA, SHAP, and GT was based on their strengths in handling nonlinear relationships, interpretability, and robust feature importance assessment, which are key requirements for water quality prediction. These methods were applied separately and compared their selected features using XGBoost performance, ensuring the most robust subset was retained for modeling.

Uncertainty analysis is essential in environmental modeling for evaluating the reliability of predictions and facilitating informed decision-making. Analyzing the uncertainty of predictions is essential for ensuring the robustness of water quality models, particularly when they are used to guide management interventions [106]. Busari et al. [106] applied different sampling periods to develop LSTM models enhanced with a Monte-Carlo dropout technique for predicting Chl-a concentrations in a freshwater lake. Their findings indicated that hourly data supported 7-day ahead predictions, surpassing the accuracy of daily data. In the current study, the uncertainty analysis using the Monte-Carlo simulation technique is applied to predictive models for providing probabilistic insights into the reliability of DO and Chl-a predictions.

In addition, assessing the risk level of water blooms using the best-selected model for predicting Chl-a is crucial for effective lake and reservoir management. Bloom risk assessment involves evaluating algal blooms' likelihood and potential impact on a water body. This comprehensive approach ensures that the model provides accurate predictions and supports effective and timely decision-making for water quality management [23]. By integrating risk assessment into the modeling framework, the study provides actionable insights for reservoir management to reduce environmental and ecological risks.

1.1. Research gap and motivation

Despite recent advancements in predictive modeling, research gaps in DO and Chl-a prediction still exist especially for bloom risk evaluations. First, most studies have focused on rivers, with limited exploration of shallow or deep dam reservoirs. These reservoirs pose distinct challenges due to their climatic variability and hydrological dynamics, which require strong and innovative modeling approaches. Second, predicting DO and Chl-a at different depths of dam reservoirs in arid and semi-arid regions, such as Iran, is currently limited. Third, despite the proven benefits of advanced feature selection techniques like Boruta and SHAP, many studies rely on traditional feature selection methods for finding input variables of ML/DL models. Fourth, hybrid models that combine the strengths of tree-based ML (e.g., XGBoost) and DL (e.g., CNNs, GRUs) are still underexplored for DO and Chl-a prediction which can effectively capture spatial patterns and nonlinear interactions. Fifth, integrating advanced feature selection techniques (BFSA and SHAP) and hybrid ML/DL models remains underutilized for predicting water quality parameters of DO and Chl-a. Finally, few studies have incorporated comprehensive uncertainty analysis and bloom risk assessment to help decision-makers for managing dam reservoirs.

1.2. Research objectives

This work aims to address the aforementioned gaps by (1) identifying the optimal input variables for predicting DO and Chl-a using advanced feature selection techniques, including BFSA, GT, and SHAP; (2) developing and evaluating state-of-the-art ML and DL models, such as XGBoost, MLP, GRUs, and CNNs, for accurate water quality prediction; (3) incorporating Monte Carlo-based uncertainty analysis and risk assessment to quantify the confidence in model predictions and support decision-making; and (4) applying the proposed framework to a dam reservoir in Iran for contributing to the limited body of research in this region.

2. Material and methods

2.1. Study area

The studied Dam reservoir is situated in southeastern Iran in a mountainous area with a semi-arid climate. Water resources in this region are limited, and the landscape is characterized by uneven formations. Agriculture significantly contributes to the economy of this region by cultivating crops like wheat, barley, pomegranates, and almonds that are suited to the cooler mountain climate. Livestock farming, particularly with sheep and goats, is a crucial economic activity in this region. The mountainous setting provides a distinctive natural environment, including valleys and streams nourished by seasonal rains and snowmelt from higher elevations. Climatic data indicate an average yearly temperature of 14.2 °C, a humidity level of 33 %, and a total precipitation of 256 mm.

The dam reservoir (Fig. 1), with a 40 million cubic meters storage capacity is the crucial source of drinking water around cities. The dam operation was started in 2009 and is responsible for downstream water demands, including 120 l/s for agriculture, 30 l/s for environmental needs, and 10 l/s for industry activities. The reservoir of dam has unique characteristics that impact its water quality. These include the inflow of floodwaters from rural areas and agricultural lands carrying organic materials, wastewater, plant residues, leaf litter, fertilizers, and pesticides. Additionally, factors such as local temperature fluctuations, the lake's depth, evaporation, and prolonged residence time due to carry-over contribute to water quality challenges.

2.2. Measurement data in dam reservoir

In this research, three measurement sections were established within the Dam reservoir (Fig. 1). In-situ data were collected using a combination of personnel, a boat, and a rapid optical sensor (CTD 310-Idronaut, Fig. S1). The depths recorded at measurement sections of B1, B2, and B3 were 48.30 m, 39.41 m, and 28.02 m, respectively. The CTD 310-Idronaut device with integrated sensors can measure multiple variables in aquatic environments. It enables automatic continuous vertical profiling, saving data within or transmitting it to a connected computer in real-time. Specifically, the CTD 310-Idronaut multiparameter device features a fast-optical DO sensor. It can measure detailed vertical distributions of DO (ppm), depth (m), temperature (°C), electrical conductivity (EC) (mS/cm), salinity (PSU), Chl-a (μg/l), and pH. The study's data were collected at various depths, ranging from the surface to approximately 32 m, on 8 May 2023. Given the significance of DO and Chl-a as water quality indicators, this study focuses on modeling these two parameters within the Dam reservoir.

2.3. Multicollinearity test on the variables

Due to strong interconnections, or collinearity, among predictor variables, the accuracy and reliability of predictions for the target variable can be significantly diminished. Therefore, it is essential to conduct a collinearity test on datasets before selecting variables for modeling. Collinearity arises when independent variables are highly correlated, significantly obscuring the individual contributions of each predictor and leading to misleading results. Multiple methods have been developed to detect and quantify collinearity in order to address this issue. Among these, the Variance Inflation Factor (VIF) has emerged as a widely adopted and robust technique for evaluating the degree of correlation between independent variables and their collective impact on the model's outcomes [35,36]. The VIF is calculated using the following equation:

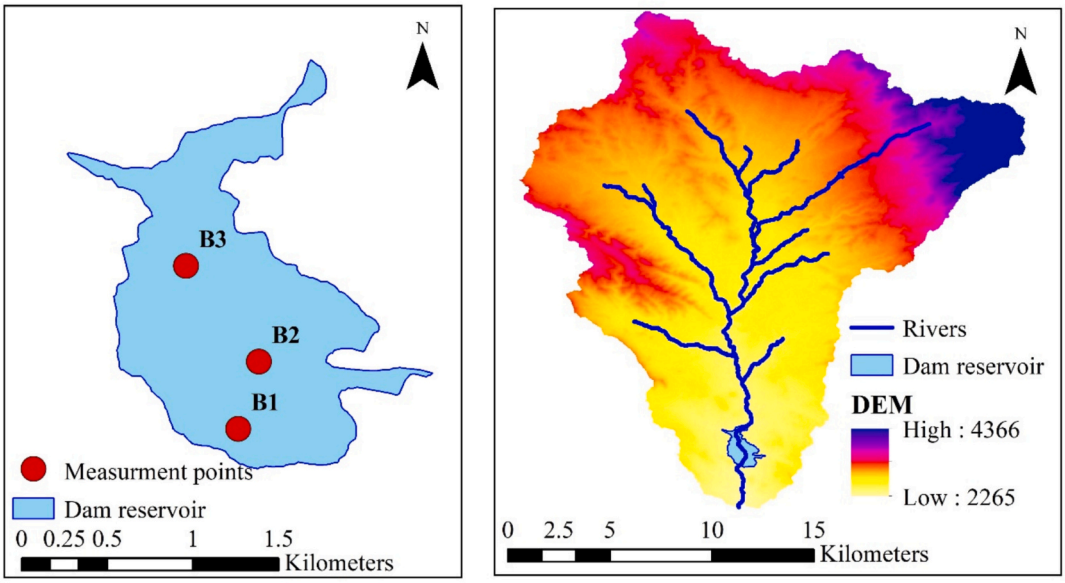

Fig. 1. The location of the study area, including Dam reservoir, and the measurement points.

$$VIF = \left[\frac{1}{1 - R^2J}\right] \quad (1)$$

where R^2J represents the coefficient of determination obtained by regressing the j^{th} independent variable against all other predictors in the dataset.

This metric measures how much the variance of an estimated regression coefficient increases due to multicollinearity. According to Bui et al. [37], a VIF value >10 strongly indicates problematic collinearity, suggesting that the associated variable may be redundant or overly dependent on other predictors. High VIF values require further investigation as they may affect the model's stability and interpretability. Addressing collinearity early in the modeling process can improve the predictive power and robustness of analyses.

2.4. Feature selection

To enhance model interpretability and predictive performance, advanced feature selection techniques namely BFSA, SHAP, and GT were applied in this study.

2.4.1. BFSA method

Feature selection is critical in developing predictive models for water quality assessment, as it enhances model interpretability, reduces overfitting, and improves computational efficiency. The BFSA is a robust, wrapper-based method designed to identify the most relevant features by comparing them with their randomized counterparts, known as shadow features. This method ensures that only the most significant predictors are retained, contributing to more reliable and interpretable models in environmental studies [29].

BFSA operates by iteratively evaluating feature importance using a classification algorithm, typically a Random Forest (RF), and determining whether a given feature provides information beyond what is expected from random noise. Features demonstrating significantly higher importance than shadow features are deemed relevant, while those with lower importance are discarded. This rigorous process makes Boruta particularly effective in handling complex, high-dimensional datasets, such as those encountered in water quality prediction [38].

The algorithm consists of eight essential steps aimed at systematically assessing and improving the significance of variables within the dataset. To begin the process, all variables are duplicated to enhance the information system and create a comprehensive foundation for analysis. Next, the newly added attributes are shuffled to break existing correlations with the target variables (DO and Chl-a). A random forest algorithm is used to calculate Z-scores, which assess the significance of each variable. The algorithm identifies the highest Z-score among shadow variables (artificial attributes) and retains only those variables that exceed this benchmark. For variables whose importance remains ambiguous after this step, an equality analysis is conducted to assess their relevance further. Subsequently, variables with lower importance are discarded, while those with higher Z-scores are preserved. Having served their purpose, shadow variables are removed from the dataset to streamline the evaluation process [39].

2.4.2. SHAP method

ML models are essential for assessing water quality because they can effectively manage complex, non-linear relationships among different environmental parameters. However, the interpretability of these models remains a significant challenge, often limiting their practical application in environmental management. To address this issue, the SHAP method has been employed as a robust method to explain and interpret the contributions of individual features in ML prediction [30].

SHAP, based on cooperative game theory, assigns an importance value to each feature by evaluating all possible feature combinations, thus providing a thorough understanding of each feature's influence on the model's output. This method ensures that the impact of each parameter is accurately assessed, promoting transparent and interpretable ML models in water quality research [40,41].

2.4.3. Gamma Test method

The GT is a non-linear modeling analysis tool used to estimate the noise variance in a dataset, which helps determine how much an output variable can be represented as a smooth function of input variables. This approach is particularly effective in identifying and quantifying non-linear relationships between inputs and outputs, which is essential for constructing accurate predictive models [42].

The key idea behind the GT is that when two input vectors are close together in the input space, their corresponding output values should also be similarly close in the output space, provided that there is a smooth underlying function. Any differences from this expectation are considered to be due to noise. The GT estimates the noise variance in the data by examining the relationship between the distances of nearest neighbors in the input space and the corresponding differences in the output space. This estimation helps in understanding the data's

suitability for modeling and the potential accuracy of predictive models [43].

2.5. Machine learning and deep learning models

Several models, including MLP, XGBoost, LSSVR, GRUs, and CNNs, are considered to predict Chl-a and DO in the dam reservoir. The selected ML/DL models are particularly effective when dealing with complex and nonlinear relationships between input and output variables. The selection of these specific techniques was driven by a comparative analysis of their strengths and suitability for the problem at hand. Each technique brings unique advantages in handling different aspects of the data complexity, such as nonlinearities, generalization, computational efficiency, dependencies, and overfitting. Each model contributes uniquely to addressing the complexities inherent in water quality prediction.

2.5.1. XGBoost model

XGBoost is an advanced version of gradient-boosting algorithms that focuses on speed and performance. It constructs an ensemble of decision trees in sequence, with each new tree correcting the errors made by the previous one. This iterative process aims to minimize a specific loss function, which improves the model's predictive accuracy [44].

A unique aspect of XGBoost is its regularization framework, which imposes penalties on model complexity to avoid overfitting. This is accomplished by incorporating regularization terms into the objective function, which controls the model's complexity and enhances generalization. Additionally, XGBoost introduces a sparsity-aware algorithm capable of handling missing values and optimizing memory usage, making it efficient for large-scale datasets.

The algorithm supports parallel and distributed computing, enabling accelerated training across multiple cores or machines. This scalability is especially beneficial when working with large datasets and complex models. Additionally, XGBoost utilizes a unique tree learning method based on quantile sketching, which allows for the handling of weighted

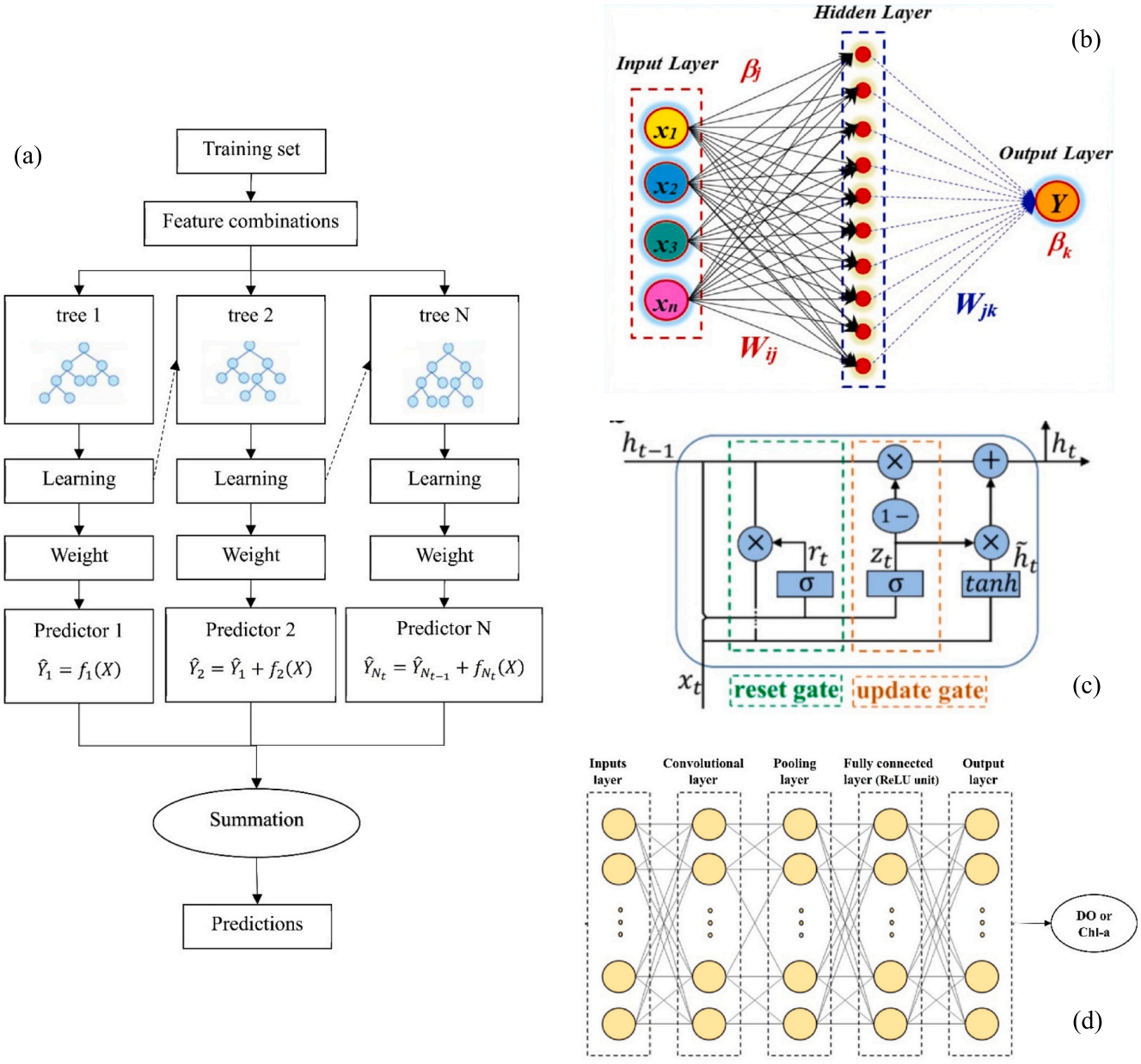

Fig. 2. The schematic diagram of (a) XGBoost, (b) MLP [48], (c) GRUs [53], and (d) CNNs [61] models.

data and enhances the process of finding splits in decision trees [45]. The schematic representation of the XGBoost algorithm is shown in Fig. 2a. Trees are built sequentially, with each tree correcting errors from the previous one. The output ($\widehat{Y}_{N_t}$) can be written as following equation

$$equal\widehat{Y}_{N_t} = \sum_{k=1}^{N_t} f_k(X) \tag{2}$$

where N_t is the number of decision trees and f_k is the k^{th} decision tree.

To gain insight into this model's structure and mathematical foundations, see Chen and Guestrin [21] and Lee et al. [46].

2.5.2. MLP model

The MLP is a type of artificial neural network (ANN) that was chosen due to its ability to learn complex patterns and relationships from data, which is essential when dealing with the multidimensional and nonlinear nature of input and output data. MLP excels at capturing intricate dependencies in data through hidden layers of neurons. In water quality prediction, MLPs are employed to analyze multiple environmental and hydrological parameters, such as temperature, pH, DO, and nutrient concentrations, to accurately forecast water quality indices [47].

An MLP consists of an input layer, one or more hidden layers, and an output layer. Each layer is made up of interconnected neurons, with each neuron applying a weighted sum of its inputs, followed by a nonlinear activation function. Common activation functions include the Rectified Linear Unit (ReLU) and sigmoid functions, which add nonlinearity to the model, allowing it to capture complex patterns within the data. The training process involves optimizing the weights and biases from the input to the hidden layer (W_{ij} and β_j) and from the hidden to the output layer (W_{ik} and β_k). This is done by minimizing the error function that measures the difference between the predicted outputs and the actual outputs [48]. This study applied the Rectified Linear Unit (ReLU) activation function, and the loss function was optimized using the Optuna optimization algorithm. The overall structure of the MLP model is presented in Fig. 2b.

2.5.3. LSSVR model

The LSSVR is an extension of the traditional Support Vector Machine (SVM) framework, tailored for regression tasks. Unlike standard SVMs, which solve a quadratic programming problem with inequality constraints, LSSVR simplifies the optimization process by transforming it into a set of linear equations. These equations utilize equality constraints along with a least squares cost function. This reformulation improves computational efficiency, particularly when working with large datasets [49]. The methodology involves transforming input data into a high-dimensional feature space through the use of kernel functions, allowing for the detection of complex patterns that linear models may overlook. This study employs the Radial Basis Function (RBF) kernel. The LSSVR training process involves minimizing a cost function that balances model complexity and fitting error. This is accomplished by introducing a regularization parameter that balances the smoothness of the regression function with the tolerance for deviations from the actual data points. The solution to this optimization problem is found by solving a system of linear equations, which is computationally simpler than the quadratic programming used in SVMs. This efficiency makes LSSVR a practical choice for real-time water quality monitoring and prediction systems [50].

2.5.4. GRUs model

The GRU is a recurrent neural network (RNN) designed to handle sequential data and capture temporal dependencies. The GRUs tackle issues found in traditional RNNs, like vanishing and exploding gradients, by using gating mechanisms to manage information flow. This architecture allows GRUs to effectively capture long-term dependencies in time-series data, making it especially suitable for applications such as water quality prediction [51].

The training process for a GRUs model involves feeding time-series data into the network, where the gating units adjust the influence of previous states and current inputs. This mechanism enables the model to learn complex temporal patterns without explicitly defined time lags or external feature engineering. The ability of GRUs to manage dependencies over extended time periods allows it to capture the dynamic and non-linear relationships inherent in environmental data. The model's structure includes reset and update gates that control the incorporation of new information and the retention of past data, facilitating accurate predictions based on historical water quality measurements [52]. The GRUs model can be expressed using the following equations [53]:

$$r_t = \sigma(W_r x_t + U_r h_{t-1} + b_r) \tag{3}$$

$$z_t = \sigma(W_z x_t + U_z h_{t-1} + b_z) \tag{4}$$

$$\widetilde{h}_t = tanh[W_h x_t + U_h(r_t \otimes h_{t-1}) + b_z] \tag{5}$$

$$h_t = (1 - z_t) \otimes h_{t-1} + z_t \otimes \widetilde{h}_t \tag{6}$$

where W and U represent the weight matrices, b represents bias vector, z_t shows the update gate, r_t is the reset gate, $\widetilde{h}_t$ represents the candidate state, and h_t shows the hidden state.

The structure of the GRUs model is illustrated in Fig. 2c.

2.5.5. CNNs model

CNNs have become a powerful deep learning architecture for managing spatially structured data, making them particularly effective for various environmental modeling tasks, including water quality assessment. The basic structure of CNNs includes three key layers: convolutional, pooling, and fully connected layers, each with a distinct role in feature extraction and classification [54].

The architecture of CNN models consists of several convolutional layers, each followed by an activation function (typically ReLU) and pooling layers that reduce dimensionality and improve computational efficiency [55,56]. The high-level features extracted are passed through fully connected layers to generate final predictions. Additionally, dropout layers are incorporated to mitigate overfitting, as recommended in previous studies [56].

One of the key advantages of CNNs in water quality prediction is their ability to process raw sensor data with minimal preprocessing, thereby reducing reliance on domain-specific feature engineering [107]. This characteristic makes CNNs adaptable to diverse environmental datasets, including those collected from reservoirs, rivers, and coastal regions. Additionally, CNN models can be integrated with other ML and DL architectures.

In water quality estimation, CNNs have been employed to identify spatial patterns within datasets, utilizing their capacity to automatically detect relevant features [58–60]. The study by Zamani et al. [61] highlighted the effectiveness of CNNs in estimating key water quality indicators (WQIs), such as Chl-a and DO. Their CNNs model was designed to process multidimensional data inputs, including spatial attributes (latitude, longitude, and depth), to extract meaningful features contributing to accurate predictions. The structure of the CNNs model is illustrated in Fig. 2d.

The models and all simulations were developed and implemented in Python, utilizing its comprehensive libraries and tools for computational modeling and analysis. It should also be noted that data from sections B1 and B3 were used for modeling, including training and testing, while data from section B2 were used for assessing best-selected models.

2.6. Uncertainty analysis

Monte Carlo simulation was used to assess the uncertainty of the models. It is considered one of the most effective techniques for uncertainty analysis and has gained considerable popularity in recent years, especially in hydrological modeling studies. The modeling process begins by identifying the optimal probability distribution function for the model's output. A Kolmogorov-Smirnov (KS) test was employed to select the best-fitting distribution from a set of 20 candidate distributions, including normal, gamma, beta, exponential, lognormal, Weibull (minimum and maximum), Pareto, Cauchy, chi-squared, Student's t, uniform, triangular, logistic, exponentially modified normal, skew-normal, Gumbel (right and left), Laplace, and power-law distributions. Next, the selected probability distribution function was used to simulate the model's output. For each prediction, a random value was drawn from the respective distribution and converted into a feasible value using the distribution's mean and variance. This sampling process was repeated 1000 times, and the resulting data were analyzed to assess variations in the model's results [62].

This method uses the 95 % predicted uncertainties (95PPU) factor to establish a range of model outputs, defining the bounds for uncertainty analysis. The 95PPU is calculated by determining the cumulative distribution of prediction values at the 97.5 % and 2.5 % limits based on 1000 predictions. For the uncertainty level to be considered appropriate, the 95PPU range should include "most of the model's prediction". Specifically, an acceptable uncertainty level is achieved when the 95PPU covers 80–100 % of the data. In more complex scenarios with lower estimation accuracy, having 50 % of the measured samples within the 95PPU range is deemed satisfactory. Also, to determine the reliability and predictive capability of the models, the degree of uncertainty ($\overline{d}_x$) is used as an assessment tool. The percentage of the model's prediction data falling within the 95 % confidence interval of Monte Carlo simulations and d-factor are calculated as [63]:

$$95PPU(\%) = \frac{Count(Q \backslash X_L \le Q \le X_U)}{n} \times 100 \tag{7}$$

$$\overline{d}_x = \frac{1}{k}\sum_{i=1}^{k}(X_U - X_L)_i \tag{8}$$

$$d-factor = \frac{\overline{d}_x}{\sigma_x} \tag{9}$$

where k is the total number of samples, σ_x is the standard deviation of the DO or Chl-a variables, and $\overline{d}_x$ is the mean distance between the upper and lower bounds.

2.7. Risk of water blooms

Water blooms, characterized by the rapid proliferation of algae in eutrophic water bodies, are a significant environmental concern. Four-level classification system has been established, aligned with nationally recognized eutrophication standards including no risk (Chl-a < 10 mg/m^3, indicating poor or medium nutrition), low risk (Chl-a 10–20 mg/m^3, slight eutrophication), medium risk (Chl-a 20–40 mg/m^3, moderate eutrophication), and high risk (Chl-a > 40 mg/m^3, severe eutrophication) [23].

Current methods for evaluating bloom risk are mainly data-driven, which makes them vulnerable to variations in data quality and uncertainties in factor relationships. To address this, the risk probability (R) has been calculated using the following equation:

$$R = P \times K \times E \tag{10}$$

where P represents the average probability of water bloom occurrence, K denotes to monitor data accuracy, and E signifies model prediction accuracy. For P, it is assumed that when Chl-a exceeds 40 mg/m^3 (severe eutrophication), $P = 100$ %; for Chl-a concentrations (C) below 40 mg/m^3, P can be calculated as follows:

$$P = \left(\frac{C}{40}\right) \times 100 \tag{11}$$

This probabilistic approach provides a practical framework for water bloom risk assessment, moving beyond traditional methods' rigid classifications and better reflecting real-world scenarios' complexities [23].

2.8. Models evaluation criteria and diagrams

The performance of various models was assessed using multiple statistical evaluation metrics, such as the root mean square error (RMSE), the coefficient of determination (R^2), mean absolute error (MAE), and (PBIAS). The mathematical formulations for these criteria are provided below:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(Y_i^{obs} - Y_i^{pre}\right)^2} \tag{12}$$

$$R^2 = \left(\frac{\sum_{i=1}^{N}\left(Y_i^{obs} - \overline{Y}\right)\left(Y_i^{pre} - \overline{Y}\right)}{\sqrt{\sum_{i=1}^{N}\left(Y_i^{obs} - \overline{Y}\right)^2}\sqrt{\sum_{i=1}^{N}\left(Y_i^{pre} - \overline{Y}\right)^2}}\right)^2 \tag{13}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|\left(Y_i^{pre} - Y_i^{obs}\right)\right| \tag{14}$$

$$PBIAS = \frac{\sum_{i=1}^{N}\left(Y_i^{pre} - Y_i^{obs}\right) \times 100}{\sum_{i=1}^{N} Y_i^{obs}} \tag{15}$$

where N is the total number of observations, Y_i^{pre} is the predicted value, and Y_i^{obs} is the observed value.

In addition, models were evaluated using two plots including Taylor diagram and the error box plot. The Taylor diagram is a powerful graphical method integrating three essential statistical metrics, including standard deviation (SD), centered root mean square difference (RMSD), and the R^2 into a comprehensive visualization. This diagram offers a straightforward method to evaluate the alignment between model predictions and observed data. Plotting these metrics together allows for a straightforward comparison of model performance, highlighting how closely the predicted results match the observed values in terms of variability, error magnitude, and correlation strength. Error box plots provide a visual summary of error distributions, highlighting insights into consistency, spread, and potential outliers in model predictions. These tools work together to create a strong framework for evaluating the accuracy and reliability of models, allowing researchers to identify their predictive strengths and weaknesses.

2.9. Model overview and framework development

The framework of the model consists of a series of essential steps, which are described in detail as follows and shown in Fig. 3.

Step 1. Data preprocessing and evaluation using Pearson correlation coefficient heatmap, density plot, and VIF analysis

The density distribution plots of water quality variables are given in Fig. S2. The x-axis represents the measured values of these parameters, while the y-axis denotes the density of occurrences, providing insights into the frequency distribution of the data. It is necessary to note that the depths recorded at measurement sections of B1, B2, and B3 were 48.30 m, 39.41 m, and 28.02 m, respectively. The density distributions of temperature across the three sections (B1, B2, and B3) indicate relatively

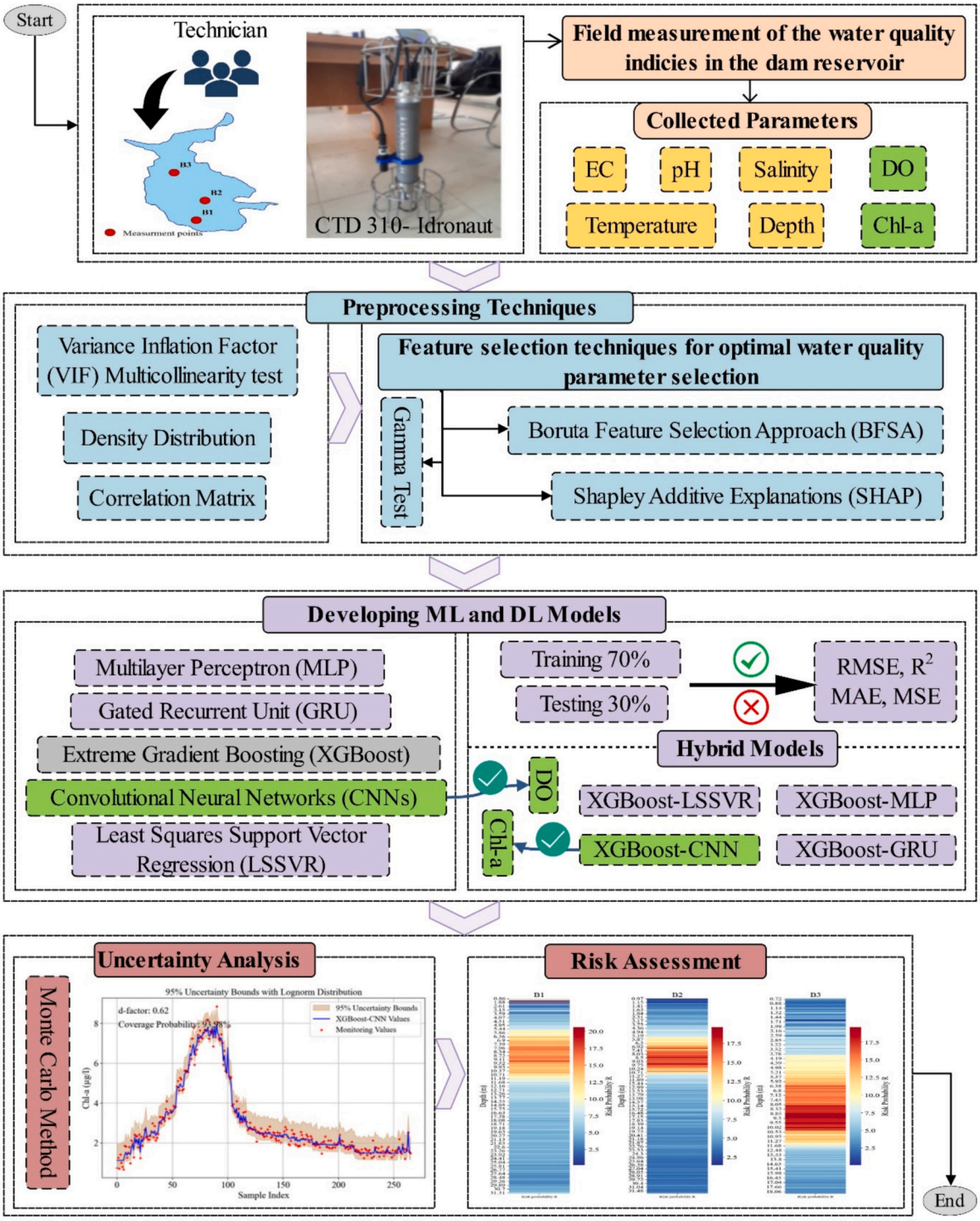

Fig. 3. Flowchart of modeling process for predicting DO and Chl-a.

similar trends, with an unimodal distribution in each case. There are no extreme deviations, indicating stable thermal conditions across sampling sites. The EC distributions exhibit multimodal patterns at all three locations. While the overall trends are similar, B1 shows a slightly broader distribution, suggesting more variability in conductivity. In this section, multiple peaks may indicate influences from inflows or anthropogenic activities. The salinity density distributions in B1 and B2 display a right-skewed pattern, peaking at lower salinity values with a gradual decline toward higher salinity levels. This trend indicates that the reservoir primarily contains freshwater, with only minor saline intrusions. The pH density plots indicate that the reservoir conditions are neutral to slightly alkaline. The broader distribution observed in B3 may suggest more significant fluctuations in pH levels, potentially due to biological activities or localized inputs. The density plots for DO indicate a multimodal distribution, particularly in B1 and B2. The variation suggests possible differences in aeration, photosynthetic activity, and water column stratification across the reservoir. The broader distribution in B3 is affected by the lower depth of measurements. It may reflect localized mixing or external inputs influencing DO concentrations. Chl-a density distributions exhibit variations between sections. The right-skewed nature of the distributions suggests episodic algal blooms or localized eutrophication processes, potentially driven by nutrient

availability and hydrodynamic factors.

The Pearson correlation coefficient heatmaps for B1, B2, and B3 datasets are given in Fig. S3. The Chl-a shows a moderate positive correlation with pH and a weak correlation with other parameters, suggesting pH levels might influence Chl-a concentrations. Ibelings and Maberly [64] found that elevated surface pH facilitates the uptake of atmospheric CO_2 into aquatic systems through a mechanism known as chemical enhancement, which subsequently enhances algal photosynthesis and leads to increased Chl-a concentrations. Han et al. [10] confirmed this result when investigating the correlation between variables in the Namhan River watershed in South Korea.

The correlation between Chl-a and Temperature in the B3 section differs from that observed in B1 and B2, potentially due to variations in hydrological characteristics, the shallower depth of B3, and anthropogenic influences. The correlation of Chl-a with other water quality parameters in the B3 section, located upstream of the dam reservoir, is lower than in the B1 and B2 sections. This emphasizes the need to monitor upstream Chl-a levels to better predict and manage changes in downstream water quality. Different research has confirmed that water temperature is crucial in regulating algal bloom dynamics in aquatic ecosystems [65,66]. Also, a study on rivers by Kim et al. [67] and another on rivers and marginal lakes by Casanova et al. [68] found a strong positive correlation between EC and Chl-a. However, our findings do not fully align with these results, as we observed relatively weak correlations between Chl-a and two variables of EC and Temperature. The current study results are similar to those of Yang et al. [13], who applied correlation analysis to reveal the Chl-a concentration pattern in the Fuchun River, China. They presented that the precise relationship and the extent of the impact remain uncertain, as the correlations between variables are highly intricate due to their strong dependence on spatiotemporal variability, climate dynamics, and anthropogenic influences.

The DO strongly correlates with all input variables of depth, temperature, EC, salinity, and pH. Notably, the strong correlation of dissolved oxygen with depth in datasets B1 and B2, as well as with salinity in section B3, underscores the impact of these factors on oxygen levels. The strong correlations between DO and other variables indicate reliable relationships that can be utilized for predictive modeling and environmental monitoring. However, strong correlations also indicate potential multicollinearity, which should be addressed in multivariate analyses to avoid biased estimates.

The VIF test was conducted on the dataset to identify and eliminate variables exhibiting strong collinearity. Table 1 illustrates the collinearity relationships among the independent variables for predicting Chl-a and DO. The results revealed that nearly all variables demonstrated a low risk and acceptable degree of collinearity ($1 < \text{VIF} \leq 5$) in prediction Chl-a. However, from the results of VIF values for DO predicting, Depth and Temperature variables show strong multicollinearity ($\text{VIF} > 10$), and two variables of Salinity and pH show moderate collinearity ($5 < \text{VIF} \leq 10$) with DO. Given these findings, all variables for Chl-a and two variables of EC and Salinity are selected for further analysis to ensure multicollinearity did not significantly impact the analytical process.

Step 2. Feature selection using BFSA, SHAP, and GT techniques

Table S1 presents the five best feature combinations identified through the GT for predicting Chl-a and three feature combinations for predicting DO in the dam reservoir. The GT was employed to evaluate the predictive relevance of different input variables, minimizing noise and maximizing the accuracy of predictive models. Two key metrics, Gamma and V_{ratio}, were used to assess the performance of selected variables. Lower Gamma and V_{ratio} values indicate better predictive capability with minimal uncertainty.

For Chl-a prediction, the Temperature, EC, and pH combination yielded the lowest Gamma value (1.098) and V_{ratio} (0.221), indicating that these parameters have the highest predictive power for Chl-a concentration in the reservoir. Expanding the selection to include Salinity resulted in a marginal increase in Gamma (1.102) and V_{ratio} (0.222), suggesting a slightly reduced predictive efficiency.

The best-performing variable combination for DO prediction was EC and Salinity, which recorded the lowest Gamma (0.031) and V_{ratio} (0.007). Considering only Salinity as the input vector resulted in a slightly higher Gamma value (0.500), which indicates that the Salinity variable contributes meaningfully to improving prediction accuracy.

The results from the BFSA emphasize Temperature, pH, and EC as the most critical predictors for Chl-a. Based on the BFSA, the importance of EC and Salinity variables were almost similar. The pH (Z-score = 1.29) is the most important variable for Chl-a prediction. These findings underscore the need for continuous monitoring of these key variables to improve the accuracy of machine learning-based water quality models. Since the Z-score values of all variables were lower than the critical threshold, all variables are selected for further modeling of Chl-a.

Fig. S4 illustrates the impact and importance of different water quality parameters in predicting Chl-a concentrations in the dam reservoir, as determined by the SHAP method. SHAP values quantify the contribution of each input variable to the model's predictions and help to identify the most influential features. The distribution of SHAP values and mean absolute SHAP values show that the pH and Temperature variables have the highest variation in SHAP values, indicating their strong influence on model output for predicting Chl-a. Higher pH values tend to have a positive impact on Chl-a predictions. Increased pH levels may be associated with enhanced algal growth. Depth has relatively lower SHAP values and indicates a weaker but still notable influence on Chl-a predictions. Therefore, based on the SHAP analysis, two optimal feature combinations of (1) Temperature and pH, and (2) Temperature, pH, and Depth are selected for further modeling and prediction of Chl-a concentration in the dam reservoir. These findings align with established ecological and hydrological processes. For instance, Saravani et al. [69] emphasized that water temperature and pH regulate biogeochemical processes, including photosynthesis and nutrient cycling. Specifically, temperature directly affects algal growth, with 15–30 °C being optimal for many species [70,71]. Han et al. [10] further highlighted the interaction between pH and temperature: higher temperatures amplify pH's effect on Chl-a, while stratification may create localized depth-specific conditions where cooler temperatures and elevated pH still promote algal growth—possibly due to enhanced CO_2 influx in stratified layers [72]. Water depth and surface area influence sediment resuspension, altering inorganic suspended solids (ISS) and light availability [73,74]. Elevated ISS reduces light penetration, inhibiting algal growth [75,76], while depth-driven variations in the N:P ratio can disrupt nutrient-Chl-a relationships [77]. These mechanisms underscore why reservoir-specific traits (e.g., depth, stratification) must be considered in predictive models [78]. As noted by Zou et al. [79], integrating reservoir morphology (e.g., depth, area) and stratification effects could refine predictions of water bloom outbreaks.

In addition, Salinity has the highest SHAP value distribution, indicating that it is the most significant predictor of DO (Fig. S5). Therefore, based on this analysis, two input vectors of (1) EC and Salinity, and (2) Salinity are considered for further modeling. DO dynamics in aquatic systems are shaped by both biological production (primarily through photosynthesis) [80] and consumption processes (respiration and organic matter decomposition) [81], with salinity and EC serving as critical regulators. Salinity directly reduces oxygen solubility and

Table 1
VIF values to assess the degree of multicollinearity among the independent water quality variables of the dam reservoir.

Variable	R-squared	VIF (Chl-a)	R-squared	VIF (DO)
Depth	0.165	1.197	0.936	15.566
Temperature	0.165	1.198	0.903	10.272
EC	0.053	1.056	0.315	1.460
Salinity	0.168	1.202	0.829	5.839
pH	0.310	1.450	0.887	8.878

promotes stratification that can isolate bottom waters, exacerbating hypoxia [82], while EC reflects ionic nutrient concentrations that drive microbial activity—high EC (e.g., from NH_4^+ or PO_4^{3-}) stimulates oxygen-consuming heterotrophic respiration [83], and low EC may limit algal oxygen production. These features interact with temperature and physical aeration, where elevated EC and salinity collectively signal increased organic loads or reduced mixing potential, aligning with their identified importance in predicting DO declines through oxidative decomposition [84,85] and stratification-linked respiration of algal biomass [82].

Table 2 presents five different input variable combinations for predicting Chl-a and DO, selected using BFSA, SHAP, and GT techniques. Each method identifies the most influential water quality parameters to improve predictive accuracy. The selected feature sets vary, with embedded-based BFSA including all variables, while SHAP focuses on the most impactful ones, such as Temperature, pH, and Depth. GT-selected combinations emphasize EC and Salinity as critical factors for Chl-a and DO. These feature combinations will be further evaluated using XGBoost to determine their effectiveness in predictive modeling. Uddin et al. [86] applied eighteen feature selection techniques, including Boruta, to identify key water quality parameters for predicting the water quality index (WQI) in Ireland. The findings suggest that embedded-based approaches outperform traditional filter methods, such as Principal Component Analysis (PCA) and Pearson Correlation (PCOR), in identifying the most significant water quality indicators.

Step 3. Evaluating various combinations of input variables from step 2 to determine the most effective configuration

The XGBoost model is utilized to assess the various combinations in this stage. The model was fine-tuned using the Optuna optimization algorithm, a hyperparameter optimization framework. The dataset was randomly divided into two subsets, with 70 % allocated for training and 30 % reserved for testing. This partitioning ensured the model was trained on a substantial portion of the data. After testing multiple variable combinations, the optimal configuration was selected based on performance metrics to ensure the best predictive capability.

Step 4. Develop comparing models

The performance of the XGBoost model was evaluated in comparison with several other individual models, including MLP, LSSVR, GRUs, and CNNs. The assessment used multiple evaluation criteria, error box plots, uncertainty analyses, scatter plots, and Taylor diagrams. Based on these assessments, the most effective model was selected for further analysis.

Step 5. Development of hybrid models

In this step, hybrid models are developed by combining the best-performing model from Step 4 with other applied models. The primary objective of this hybridization process is to enhance predictive performance by leveraging the strengths of multiple algorithms. A regularized XGBoost model (optimized via Optuna with L1/L2 regularization) was trained to predict the target variable of Chl-a. Its output predictions were then treated as an additional engineered feature, capturing non-linear relationships and feature importance from tabular data. The XGBoost predictions were used as standalone inputs and fed into MLP, LSSVR, GRUs, and CNNs models to form hybrid frameworks [87]. Therefore, XGBoost-MLP, XGBoost-LSSVR, XGBoost-GRUs, and XGBoost-CNNs models were developed to explore potential improvements in predictive capabilities through hybridization.

Table 2
The best-selected feature combinations of variables based on BFSA, SHAP, and GT techniques.

Method	Selected variables for predicting Chl-a		Selected variables for predicting DO	
BFSA	F1	Depth, temperature, EC, Salinity, pH	I1	EC, Salinity
SHAP	F2	Temperature, pH	I2	EC, Salinity
	F3	Temperature, pH, Depth	I3	Salinity
GT	F4	Temperature, EC, pH	I4	EC, Salinity
	F5	Temperature, EC, Salinity, pH	I5	Salinity

Step 6. Validation of the selected model on the B2 dataset

The datasets obtained from two distinct locations, B1 and B3, were utilized for training and testing the model. In contrast, data from the third location, B2, were exclusively allocated for the validation. This approach ensures that the final model is evaluated on an independent dataset, thereby assessing its generalization capability and performance in predicting water quality parameters under different spatial conditions.

Step 7. Risk assessment of water bloom in dam reservoir

This process analyzes potential risks to water quality, aquatic life, and reservoir usability. The assessment aims to identify early warning indicators to prevent or minimize the adverse effects of water blooms on the ecosystem and human activities.

3. Results and discussion

3.1. Evaluation of the selected input features using the XGBoost model

Fig. 4 illustrates the XGBoost model's evaluation results optimized with the Optuna algorithm for predicting Chl-a and DO concentrations in the dam reservoir. The selected feature combinations, derived from BFSA, SHAP, and GT methods, were tested to determine the best input variables for accurate prediction. In the Chl-a prediction, R^2 values are consistently high (above 0.91) in the training phase and between 0.88 and 0.92 in the testing phase. Based on the R^2 values in the testing phase, feature combinations of F1, F2, and F3 can be considered as the best. RMSE and MAE values vary among the selected feature sets, with F1 and F3 demonstrating the lowest errors in the testing phase and suggesting these feature combinations can enhance model precision. Also, the minimum value of PBIAS and maximum value of 95PPU in the testing phase were observed for F1. The results confirm that using carefully selected features, the XGBoost model can effectively predict Chl-a. The feature combination of F1 demonstrated the highest accuracy and lowest error rates.

In the case of DO prediction, the R^2 values are >0.94, demonstrating exceptional model reliability in capturing DO variations. RMSE and MAE values are notably low, and I1 shows the best performance. PBIAS analysis reveals minor deviations and the lowest value is attained for the I1 (0.043) feature combination in the testing phase. The values of 95PPU are obtained equal to 98.2 % for I1 and 88.8 % for I3 in the testing phase. This result highlights a good level of predictive confidence and uncertainty handling. Therefore, the results confirm the capability of the I1 feature combination, selected by the BFSA method, for predicting DO in dam reservoirs.

3.2. Individual and hybrid models performance

Fig. 5 presents the comparative performance evaluation of models (MLP, LSSVR, GRUs, and CNNs) in predicting Chl-a and DO concentrations. The evaluation uses the best-performing feature sets, F1 and I1, which include Depth, Temperature, EC, Salinity, and pH for F1, and EC and Salinity for I1, as selected by the BFSA method. Among individual models, XGBoost consistently outperformed all others in prediction Chl-a ($R^2 = 0.977$, RMSE = 0.339, MAE = 0.237, and PBIAS = −0.094 in training phase and $R^2 = 0.921$, RMSE = 0.553, MAE = 0.405, and PBIAS = 0.245 in testing phase). Zhu et al. [88] stated that increasing the complexity of a model's structure may enhance its capacity, but it can also lead to instability and may not necessarily improve accuracy. In this study, CNNs and GRUs models were less effective than XGBoost in accurately identifying the patterns of Chl-a and DO concentrations that contributed to the suboptimal performance of these models. In addition, XGBoost's superior performance in predicting Chl-a and DO in dam reservoirs can arise from its ability to handle structured datasets, capture complex interactions, and provide computational efficiency with high interpretability. Deep learning models, including CNNs and GRUs, require large datasets for practical training. However, data availability is

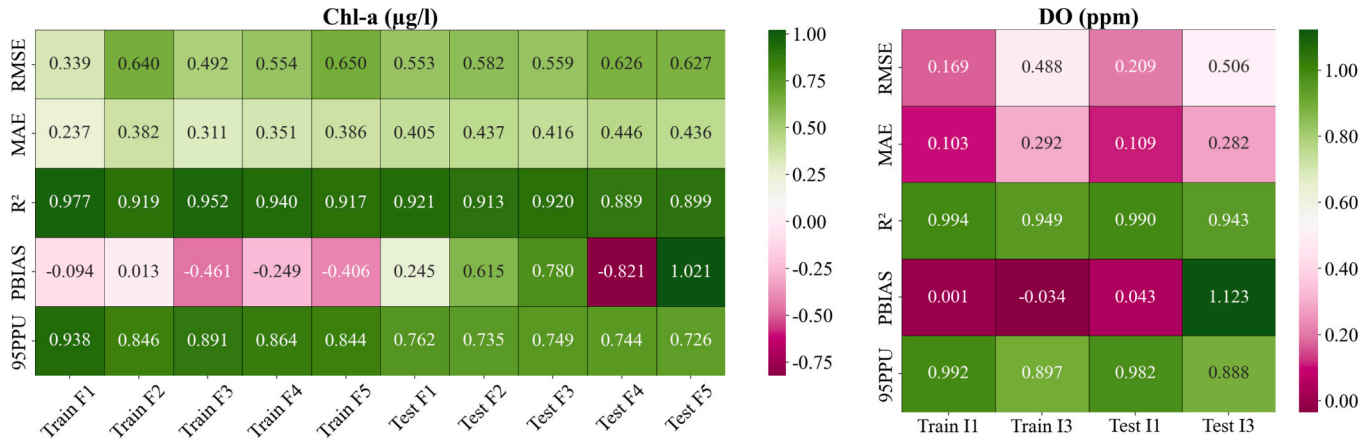

Fig. 4. Performance evaluation of selected Chl-a and DO Prediction features using the XGBoost model.

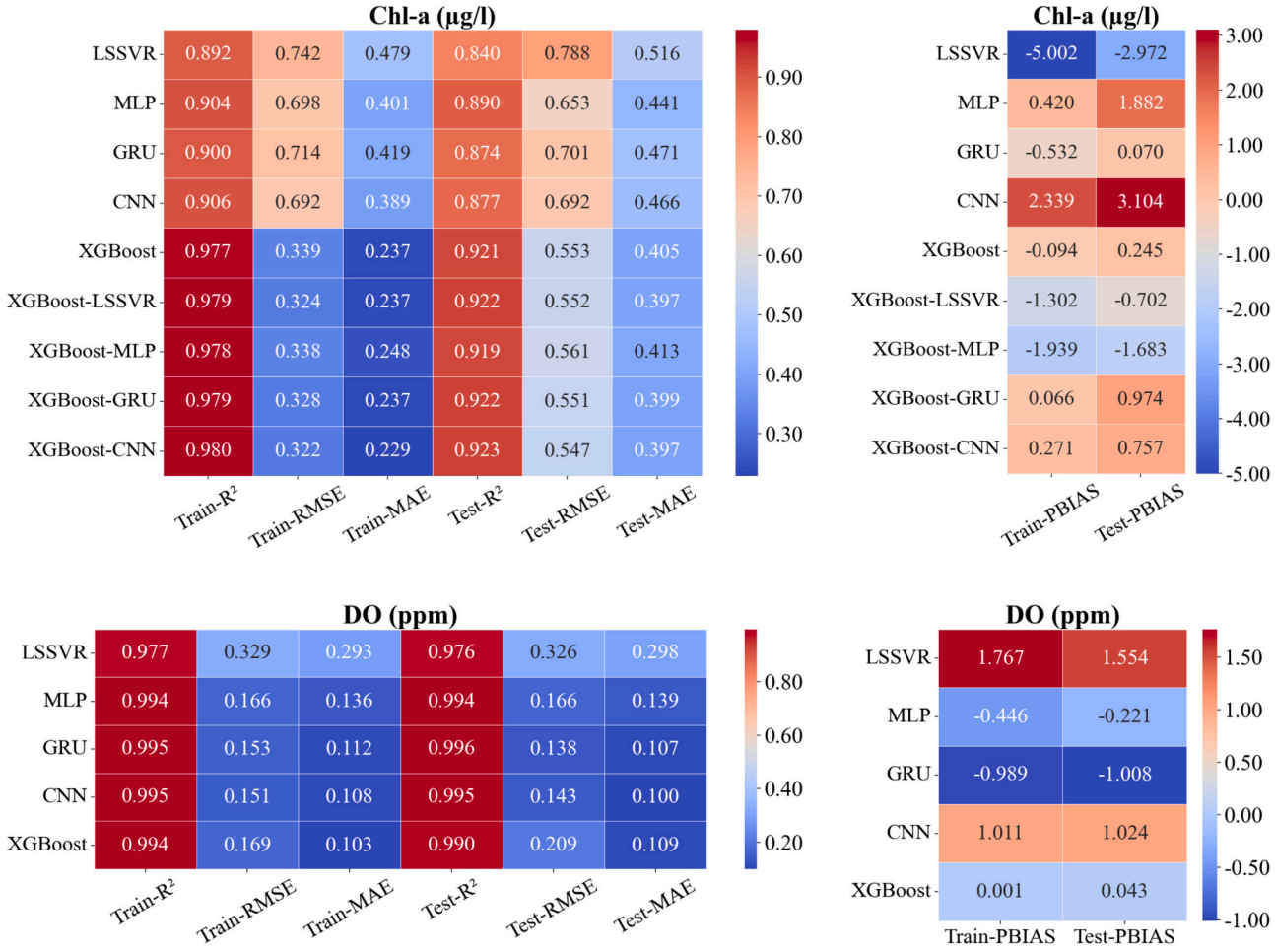

Fig. 5. Performance evaluation of machine/deep learning models for Chl-a and DO Prediction.

often limited in environmental modeling, especially for dam reservoirs, due to continuous monitoring and sampling constraints. XGBoost performs exceptionally well by efficiently capturing relationships between water quality parameters (e.g., Temperature, pH, EC, Salinity, and Depth) without requiring vast training data. Deep learning models, on the other hand, can overfit small datasets, leading to inconsistent predictions in test cases. Additionally, XGBoost employs gradient boosting with decision trees, allowing it to model complex, non-linear relationships between water quality parameters effectively. The lower RMSE values achieved by XGBoost are due to their enhanced ability to mitigate generalization errors, primarily through bootstrap aggregation [21]. Deep learning models rely heavily on neural network architectures and may require extensive tuning to recognize such interactions effectively. Unlike deep learning models, which require high computational power and extensive hyperparameter tuning, XGBoost can achieve optimal performance with fewer computational resources and shorter training times.

For prediction DO, both GRUs (R^2 = 0.995, RMSE = 0.153, MAE =

0.112, and PBIAS = −0.989 in training phase and R^2 = 0.996, RMSE = 0.138, MAE = 0.107, and PBIAS = −1.008 in testing phase) and CNNs (R^2 = 0.995, RMSE = 0.151, MAE = 0.108, and PBIAS = 1.011 in training phase and R^2 = 0.995, RMSE = 0.209, MAE = 0.109, and PBIAS = 1.024 in testing phase) show high performance. Traditional machine learning models like LSSVR and MLP showed higher errors and bias, while the tree-based model of XGBoost performed well but was slightly less stable than deep learning.

Integrating multiple estimation models through a combination approach enhances the robustness of predictions, particularly in cases where single models may not be reliable. To achieve the best performance, exploring various hybrid methodologies and strategies is essential before finalizing the merging model [27]. Given the superior performance of the XGBoost model, it was explored in hybrid models (XGBoost-LSSVR, XGBoost-MLP, XGBoost-GRUs, and XGBoost-CNNs) to enhance the predictive accuracy of Chl-a concentrations. The results demonstrate that hybrid models integrating XGBoost consistently outperform individual models across all performance metrics. XGBoost-GRUs and XGBoost-CNNs outperform standalone GRUs and CNNs models, indicating that integrating tree-based learning with deep learning improves model generalization and prediction accuracy. XGBoost-LSSVR and XGBoost-MLP show enhanced performance over their standalone versions, suggesting that boosting-based feature selection improves prediction robustness. Among all individual and hybrid models, XGBoost-GRUs and XGBoost-CNNs demonstrate the best balance between high accuracy and low error, with R^2 values exceeding 0.97 and minimal MAE/RMSE values. Hybrid models mitigate common weaknesses of individual deep learning models, such as overfitting, slow convergence, and sensitivity to data noise. Fooladi et al. [27] represented that tree-based model of Random Forest showed lower uncertainty in prediction DO and Chl-a in comparison with generalized regression neural network (GRNN), Gaussian process regression (GPR), decision tree (DT), least-squares boosting (LSB), Bayesian ridge (BR), support vector regression (SVR), K-nearest neighbors (KNN), multilayer perceptron (MLP), and group method of data handling (GMDH). Han et al. [10] developed an XGBoost model to predict Chl-a concentrations in a river in South Korea. The results demonstrated the model's strong predictive performance, achieving an R^2 of 0.9487 and RMSE of 3.1661. Shamloo and Sima [25] utilized Landsat-8 and Sentinel-2 satellite data in combination with multiple linear regression (MLR) and artificial neural networks (ANNs) to model Chl-a in Lake Urmia. Their findings revealed that models based on Landsat-8 outperformed those using Sentinel-2, mainly when ANNs were applied to predict Chl-a. The Nash-Sutcliffe Efficiency (NSE) values for ANN models were equal to 0.75, demonstrating superior predictive capabilities compared to MLR models (NSE = 0.74). The findings reinforce that hybrid tree-based and deep learning models provide a highly effective approach for real-time water quality prediction and environmental monitoring in aquatic systems.

Table 3 presents the architecture and parameter configurations of the ML/DL models applied for Chl-a and DO prediction.

The error distribution of different models in predicting Chl-a is given in the error box plot of Fig. S6. The XGBoost and its hybrid models exhibit the lowest median absolute errors with minimal variance, indicating high consistency and stability in training. Based on the Taylor diagram, hybrid models, particularly XGBoost-CNNs and XGBoost-GRUs, demonstrate the highest correlation of 0.99 and the lowest RMSD (between 1.25 and 1.61 lines) in the training phase (Fig. S7). In the testing phase, XGBoost-CNNs remain the closest to the observed values, confirming superior generalization compared to traditional machine learning and deep learning models. Therefore, the XGBoost-CNNs model is chosen as the best model with high performance for predicting Chl-a concentrations in the dam reservoir.

While error reflects a measurable deviation, it represents a deficiency in information and understanding about a system and the complexity and variability inherent in its nature [89]. In this study, the uncertainty analysis is performed on the best-selected model to ensure the reliability and stability of predictive models. Based on Fig. 6, XGBoost-CNNs achieves a high coverage probability (94.16 % in training and 79.37 % in testing phases) and most observed values fall within the 95 % uncertainty bounds. The d-factor in the uncertainty analysis quantifies the width of the uncertainty bounds relative to the predicted values. A lower d-factor indicates narrower uncertainty bounds, suggesting higher model precision, whereas a higher d-factor reflects greater uncertainty in predictions. The relatively low d-factor values (0.69 for training, 0.68 for testing) confirm that XGBoost-CNNs provides stable and precise predictions, with minimal uncertainty in Chl-a predicting. However, the slight variation in the d-factor suggests that the model can be generalized, making it a robust and reliable tool for Chl-a prediction in dam reservoirs.

Table 3
Models structure and parameter configuration for predicting Chl-a and DO.

	Parameters
Model for predicting Chl-a	
LSSVR	c: 24.893, gamma: 14.981
MLP	Learning rate: 0.001, number of hidden layers: 5, hidden size 1: 171, hidden size 2: 165, hidden size 3: 36, hidden size 4: 56, hidden size 5: 100, batch size: 16
GRUs	Learning rate: 0.005, number of hidden layers: 2, hidden size 1: 147, hidden size 1: 68
CNNs	Learning rate: 0.001, filters: 48, kernel size: 4, pool size: 1, dense units: 70, hidden neurons: 30, batch size: 48
XGBoost	Learning rate: 0.014, max depth: 6, subsample: 0.681, colsample bytree: 0.926, L1 regularization: 0.048, L2 regularization: 0.069
XGBoost-LSSVR	c: 99.936, gamma: 0.002
XGBoost-MLP	Learning rate: 0.001, number of hidden layers: 1, hidden size 1: 63, batch size: 32
XGBoost-GRUs	Learning rate: 0.0002, number of hidden layers: 4, hidden size 1: 77, hidden size 1: 19
XGBoost-CNNs	Learning rate: 0.004, filters: 112, kernel size: 1, pool size: 1, dense units: 10, hidden neurons: 10, batch size: 80
Model for predicting DO	
LSSVR	c: 6.720, gamma: 1.449
MLP	Learning rate: 0.0079, number of hidden layers: 1, hidden size: 189, batch size: 16
GRUs	Learning rate: 0.046, number of hidden layers: 2, hidden size 1: 186, hidden size 2: 67
CNNs	Learning rate: 0.003, filters: 96, kernel size: 1, pool size: 1, dense units: 20, hidden neurons: 20, batch size: 16
XGBoost	Learning rate: 0.097, max depth: 3, subsample: 0.969, colsample bytree: 0.169, L1 regularization: 3.248×10^{-7}, L2 regularization: 0.070

The plots to compare observed against predicted Chl-a concentrations at different depths of the reservoir for both training and testing of the XGBoost-CNNs model are given in Fig. S8. XGBoost-CNNs accurately follows the distribution of observed Chl-a values across different depths, though some minor deviations occur. Park et al. [16] evaluated ANN and SVM for early warning prediction of Chl-a, using input variables such as phosphate phosphorus (PO_4-P), ammonium nitrogen (NH_3−N), nitrate nitrogen (NO_3-N), solar radiation, wind speed, and water temperature. Through the Williams–Kloot test and sensitivity analysis, they found that SVM outperformed ANN in prediction accuracy and in capturing the cause-and-effect relationships between Chl-a and environmental variables in both the Juam and Yeongsan Reservoirs. Sensitivity analysis revealed that the most influential input variable differed between models and locations. In the Juam Reservoir, PO_4-P was the most sensitive variable for both ANN and SVM, whereas in the Yeongsan Reservoir, solar radiation was most critical for ANN and NH_3-N for SVM. These findings highlight the context-dependent nature of key drivers in Chl-a modeling and the importance of selecting appropriate algorithms based on system characteristics. Abbas et al. [20] evaluated six DL models including LSTM, CNN, Temporal Convolutional Network (TCN), CNN-LSTM, LSTM-based autoencoder, and input-attention LSTM (IA-

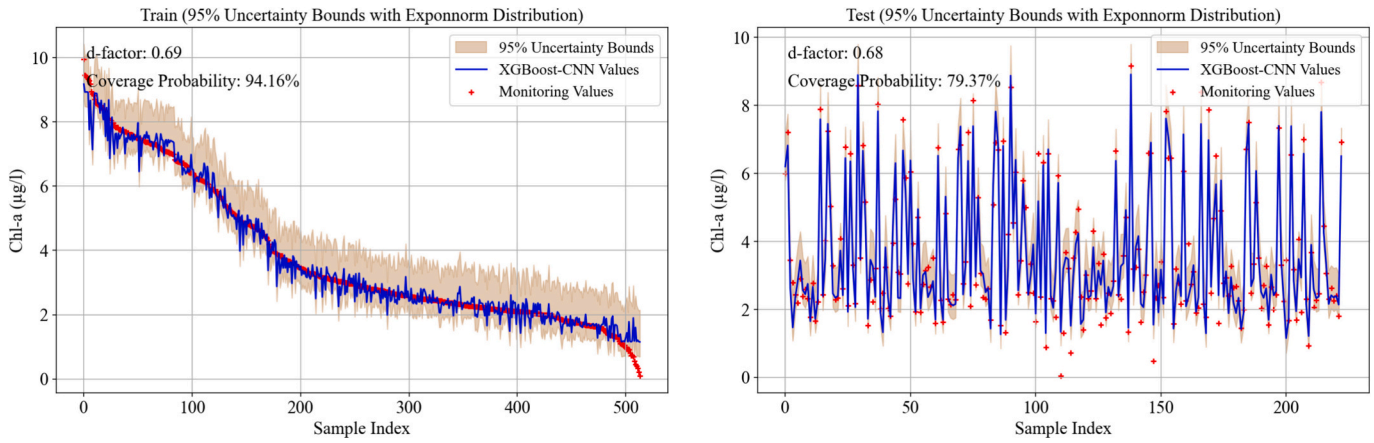

Fig. 6. Uncertainty analysis of hybrid XGBoost-CNNs model in prediction Chl-a concentrations by determining probability distribution function.

LSTM) for hydrological and Chl-a simulations. Their results demonstrated that IA-LSTM outperformed other architectures, achieving an R^2 of 0.85 at the training site and 0.52 at the test site.

The hybrid approach of XGBoost-CNNs leverages CNN's robust feature extraction and XGBoost's predictive power, consistently achieving higher accuracy, lower error rates, and faster processing than standalone models [90]. Premalatha and Bai [90] achieved 98 % accuracy and 100 % recall in healthcare analytics, surpassing decision trees, random forests, and SVMs. Jayakarthik et al. [91] demonstrated that a CNN-XGBoost model optimized with WOA achieved 98 % accuracy in climate change prediction, significantly outperforming individual CNN or XGBoost benchmarks. Prakash and Sangeetha [92] reported a 5 % efficiency gain over traditional models (e.g., SVM, LSTM) in air pollution classification using XGBoost-CNN. These studies collectively underscore that hybridizing CNN and XGBoost capitalizes on their complementary strengths, as noted in the results of the current study.

The error box plot and Taylor diagram for comparing different models in the prediction of DO are given in Figs. S9 and S10. The evaluations confirm that GRUs and CNNs models are the most effective models for DO prediction, outperforming traditional MLP and LSSVR models across all metrics. To find the best model between GRUs and CNNs, Fig. 7 shows the uncertainty analysis of these models in

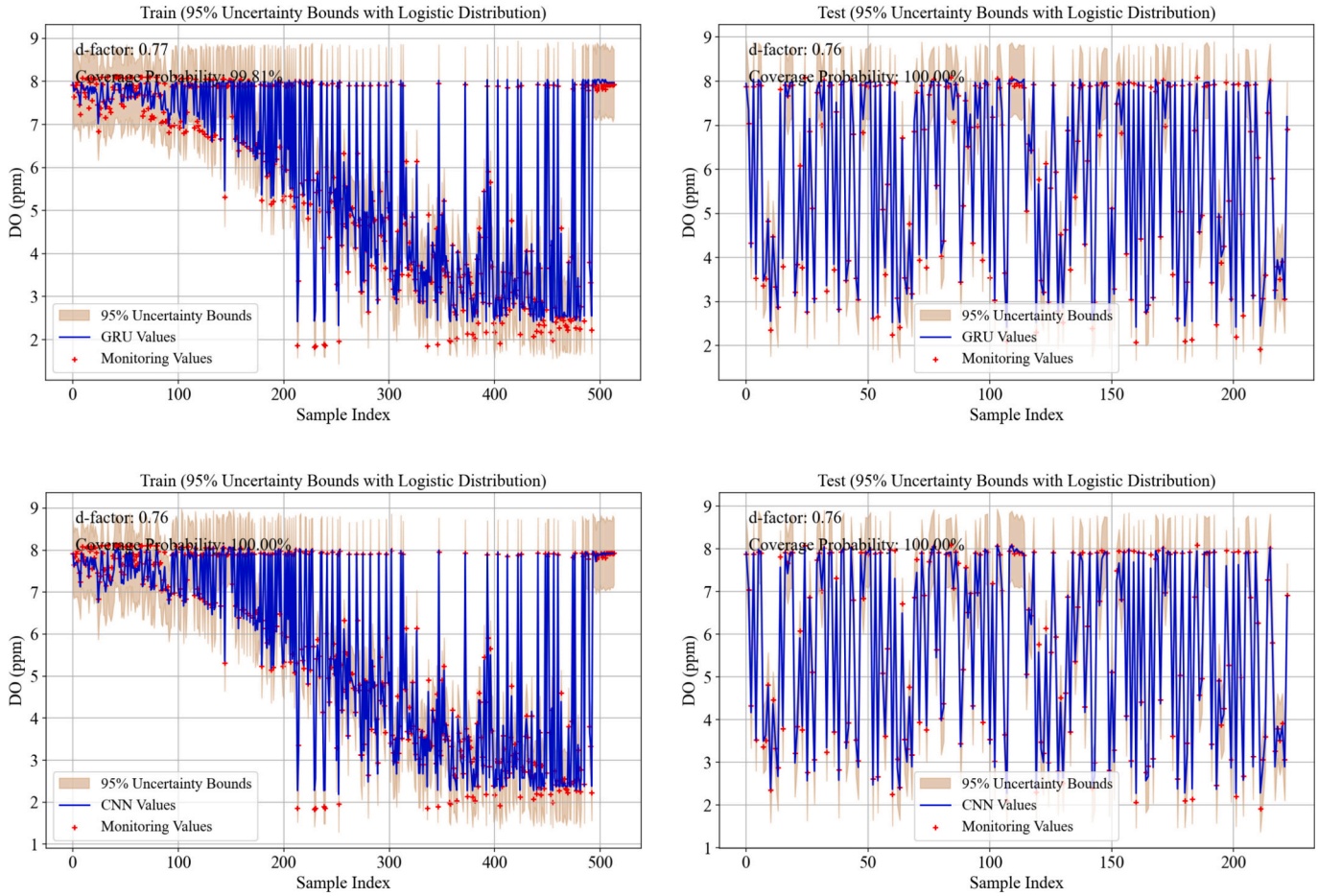

Fig. 7. Uncertainty analysis of individual GRUs and CNNs models in prediction DO concentrations by determining probability distribution function.

predicting DO concentrations. The Logistic function is the most related probability distribution function on the DO dataset for uncertainty analysis. Although the performance of the two models is almost near, 100 % coverage probability is captured within the uncertainty bounds for the CNNs model in both the training and testing phases. Low d-factor values (0.76 in both training and testing) indicate narrow uncertainty bounds and high prediction accuracy of CNNs. The plots to compare observed against predicted DO concentrations at different depths of the reservoir for both training and testing of the CNNs model are given in Fig. S11. Zhang et al. [93] utilized a MLP model enhanced with MI feature selection (MLP-MI) to predict DO concentration in Australia. Their findings demonstrated that MLP-MI provided precise DO predictions, establishing it as a valuable tool for water quality management. Moghadam et al. [94] investigated the effectiveness of various artificial intelligence models, including ANN, RNN, SVM, and DRNN, in predicting DO concentrations in the USA. Their findings demonstrated that DRNN outperformed the other models regarding predictive accuracy. Tiyasha et al. [95] developed and tested four machine learning models: RF, conditional RF, RF generator, and XGBoost, for predicting DO concentrations in Malaysia. Their findings indicated that XGBoost outperformed the other models regarding predictive accuracy. Similarly, Heddam et al. [96] applied various models, including LSTM, SVR, genetic programming (GP), group method of data handling (GMDH), SVR, and GPR, to estimate DO levels in rivers across the USA. Their results revealed that GPR exhibited the highest predictive accuracy among the tested approaches. Alizamir et al. [48] implemented the Bayesian Model Averaging (BMA) approach to integrate deep learning models such as MLPNN, RNN, CNNs, GRUs, LSTM, and SARIMAX for more accurate DO prediction across two USGS stations in the USA. Feature selection techniques, including MI and Recursive Feature Elimination (RFE), were utilized to determine the most influential predictors, with water temperature and previous DO values emerging as key factors. The results indicated that the BMA framework outperformed single-model methods, yielding lower RMSE and MAE values and higher R^2.

The model's performance applied in the current study is inherently constrained by several interconnected uncertainties:

(1) Environmental and measurement uncertainty: Uncontrolled ecological variation (e.g., stochastic weather events, unmeasured biotic interactions) and intrinsic errors from field sampling, sensor calibration drift, and laboratory analytical precision [97] introduce irreducible noise. This limits prediction accuracy and requires careful error handling, as relying on imperfect data can skew model results and reduce confidence in predictions.
(2) Data limitations: Sparse spatiotemporal sampling coverage and data quality issues (e.g., missing values, inconsistencies across sources) restrict the model's ability to capture complex system dynamics fully. This can lead to overfitting on available data, reduced generalizability to unmonitored locations or future conditions, and increased sensitivity to interpolation/extrapolation errors [98].
(3) Model structural and data-driven uncertainty: Aleatoric uncertainty and epistemic uncertainty are two sources of model uncertainty. Aleatoric uncertainty shows inherent stochasticity or noise in the observed data (e.g., natural variability of Chl-a at microscales). Epistemic uncertainty shows limitations arising from model structure, simplifications, incomplete process understanding, or insufficient training data. High epistemic uncertainty indicates low model confidence [99,100].
(4) Predictive distribution accuracy: the model's ability to generate accurate full probability distribution functions (PDFs) for key outputs like Chl-a and DO is paramount [101]. Reliability hinges not just on point predictions but on correctly characterizing the range and likelihood of possible outcomes (e.g., the probability of exceeding a critical Chl-a threshold). Underestimating prediction intervals (e.g., due to unaccounted uncertainties or distribution misspecification) creates false confidence, while overestimation reduces practical utility.
(5) The model must generate precise full probability distribution functions (PDFs) for key outputs (e.g., Chl-a, DO) through rigorous Monte Carlo simulation. The selection of appropriate PDFs directly influences reliability, as it determines the accuracy of uncertainty propagation [101]. Thus, PDF specification in Monte Carlo frameworks is critical to balancing precision and realism.

In conclusion, acknowledging, quantifying, and transparently communicating these sources of uncertainty and their impact on prediction confidence is fundamental to evaluating the true reliability and appropriate application domains of applied models. Future research should focus on mitigating epistemic uncertainty while improving the predictive output distributions.

3.3. Assessment of selected models in the third section of the dam reservoir

This section evaluates the capability of selected CNNs and XGBoost-CNNs models in the third location of the dam reservoir for predicting DO and Chl-a, respectively. Based on Fig. S12, the high R^2 values equal to 0.986 and 0.995 indicate a strong correlation between predicted and observed Chl-a and DO values, respectively. The RMSE equal to 0.224 suggests high model accuracy in predicting Chl-a variations. The value of PBIAS = 1.189 indicates a slight overestimation. Extremely low errors (RMSE = 0.153, MAE = 0.109, and PBIAS = 0.999) in predicting DO in the B2 section show perfect precision of CNNs in DO prediction. These results confirm that both models provide highly accurate predictions, with XGBoost-CNN effectively capturing Chl-a variability and CNNs excelling in DO prediction.

Fig. S13 compares observed values against predicted values at different depths of the B2 section in the dam reservoir. The XGBoost-CNNs model successfully replicates the depth-wise distribution of Chl-a, with minimal deviations across the water column. This model captures surface and subsurface variations and shows strong predictive capability. Also, the CNNs model aligns almost perfectly with observed DO profiles at all depths, demonstrating superior generalization. In addition, a coverage probability of 93.68 % for predicting Chl-a using the XGBoost-CNNs model and 100 % for predicting DO using the CNNs model indicates that most observed values fall within the 95 % confidence interval of the model predictions (Fig. 8). These models exhibit narrow uncertainty bounds, reinforcing their robustness in real-time Chl-a and DO prediction.

3.4. Spatial pattern predictions

The dam reservoir's depth-based distribution of observed and predicted Chl-a and DO concentrations across three sections (B1, B2, and B3) are shown in Fig. 9. The CNNs model was used for DO prediction, while the XGBoost-CNNs model was employed for Chl-a prediction due to their superior performance in earlier evaluations. These visualizations assess the models' effectiveness in capturing spatial and vertical variations in water quality parameters. The XGBoost-CNNs model effectively captures the spatial pattern of Chl-a concentrations, with higher values near the surface and lower concentrations in deeper layers. Localized variations are visible, particularly in B2 and B3, where Chl-a peaks occur at mid-depths, possibly due to light penetration, stratification, and nutrient availability. In the case of DO prediction, the CNNs model effectively captures the expected vertical stratification of DO, with higher concentrations near the surface due to atmospheric oxygen diffusion and gradual depletion with depth due to biological respiration and limited mixing. The transition between oxygen-rich surface layers and oxygen-depleted bottom layers is well-represented across all three sections (B1, B2, and B3). The high accuracy of models in predicting DO

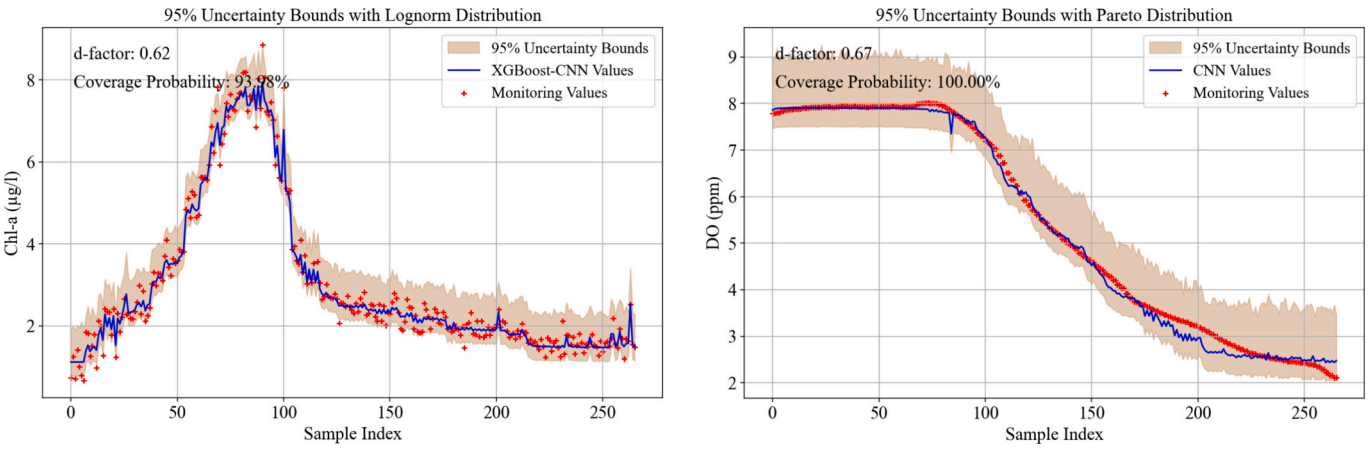

Fig. 8. Uncertainty analysis of CNNs and XGBoost-CNNs models in prediction DO and Chl-a in B2 section, respectively.

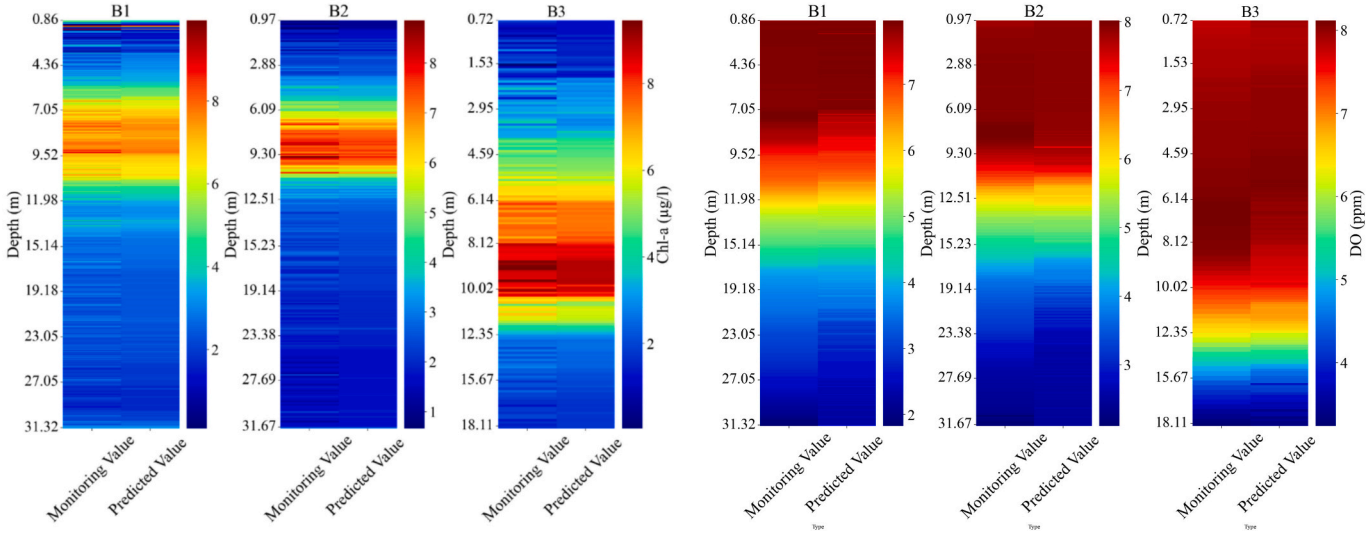

Fig. 9. Depth-based distribution of Chl-a and DO concentrations predicted across three different sections of the dam reservoir (B1, B2, and B3).

and Chl-a levels across varying depths reinforces its capability for real-time water quality monitoring and aquatic ecosystem modeling. Based on Figs. 5 and 9, there are no strong correlations between DO and Chl-a in the dam reservoir. Sutula et al. [102] stated that the relationship between these variables varied significantly across different sub-embayments. The absence of a consistent and significant association between DO and Chl-a in this reservoir suggests that physical factors prevented the formation of low DO conditions, even in areas with high phytoplankton biomass accumulation [108].

3.5. Water blooms risk assessment

There are strong correlations between water bloom abundance and Chl-a concentration [102]. For this reason, risk evaluation is a crucial initial step in determining the effects of Chl-a and offering a framework for assessing the potential development of water blooms. The depth-based distribution of risk probability in three sections (B1, B2, and B3) of the dam reservoir is shown in Fig. 10. This figure highlights areas with potential environmental risks associated with water blooms. The color gradient of the R factor represents the magnitude of risk in producing water blooms, where higher values (red regions) indicate increased risk and lower values (blue regions) suggest stable conditions. B2 and B3 sections show more risk zones at mid-depths (~6–11 m), likely due to fluctuations in Chl-a concentrations and oxygen depletion. Higher risk probability at mid-depths (B2 and B3) suggests regions where oxygen depletion (hypoxia) or algal blooms are more prevalent, potentially due to nutrient enrichment, stratification, or limited mixing. High Chl-a concentrations and temperature-driven phytoplankton growth are linked to near-surface risk zones, which may impact light penetration and DO availability. Deeper regions exhibit lower risk levels, likely due to decreased biological activity and stable physical-chemical conditions. The results highlight the importance of continuous monitoring and adaptive management strategies to mitigate water quality risks. Mid-depth regions require targeted interventions such as aeration or controlled nutrient loading to minimize the impact of algal blooms and oxygen depletion. The risk categories (no risk, low risk, medium risk, and severe risk) are determined based on predefined thresholds and shown in Fig. S14. The findings indicate that mid-depth zones are critical areas for phytoplankton growth. The spatial variations in risk probability emphasize the need for location-specific management approaches in dam reservoirs to ensure long-term ecological sustainability and water quality preservation.

3.6. Research limitations and prospects

Despite advancements in predicting Chl-a and DO concentrations using ML and DL models, several limitations must be acknowledged. The dataset utilized in this study was restricted to three specific sections (B1,

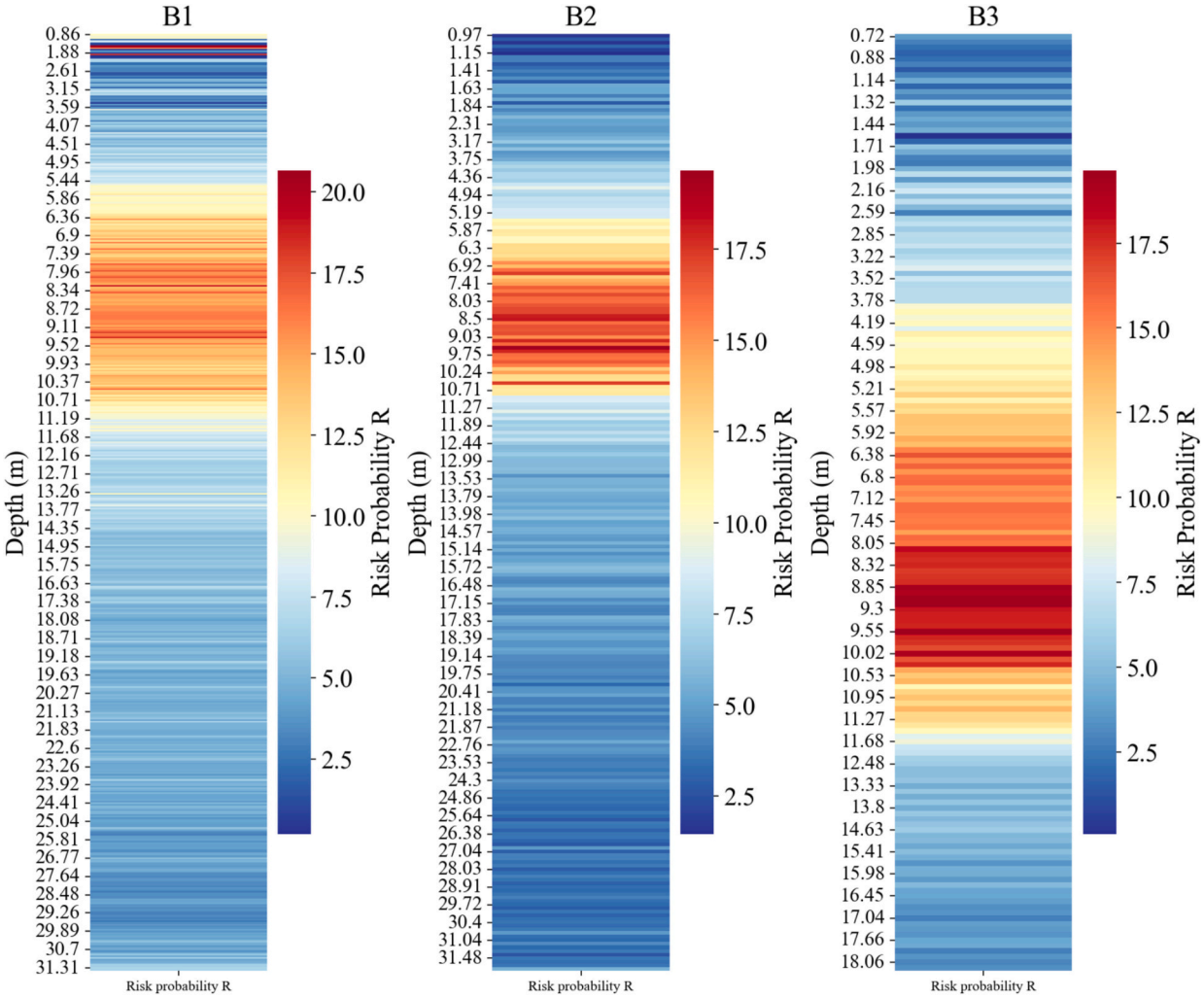

Fig. 10. Depth-based distribution of risk probability for water blooms across three different sections (B1, B2, and B3) of the dam reservoir

B2, and B3) of the dam reservoir, potentially limiting the generalizability of the findings across different spatial and temporal scales. While the selected sections capture variations in water quality parameters, expanding monitoring locations and increasing temporal coverage could provide a more comprehensive understanding of water quality dynamics. Future research should incorporate additional sampling sites across various depths and geographical locations within the reservoir and extended observation periods to account for seasonal and interannual variations. Also, the study relies on a single day of measurements (8 May 2023) across three sections. Since water quality parameters can change quickly, it is necessary to monitor data over a long time period for accounting seasonal variations.

Another significant limitation stems from the complex nature of water bloom formation in reservoirs and lakes. While this study successfully identified key predictors for Chl-a and DO concentrations, water blooms are influenced by many factors beyond the selected water quality parameters. The formation of water blooms in lakes and reservoirs is a highly intricate phenomenon, influenced not only by water quality variables such as water temperature, pH, and EC, but also by meteorological factors such as precipitation, wind velocity, and temperature [1,103]. Moreover, hydrodynamic processes, including nutrient transport, sediment resuspension, and stratification, are crucial in determining bloom outbreaks [78]. Meng et al. [104] developed a Reservoir Water Bloom Risk Index (RWBRI) to assess large-scale eutrophication in reservoirs across Fujian Province, China. Their approach integrated multiple variables, including the Trophic State Index (TSI), Floating Algae Index (FAI), temperature, precipitation, and wind speed, using Sentinel-2 MSI data at the pixel scale. Similarly, Xue et al. [105] evaluated algal bloom risk in the Xiashan and Jihongtan reservoirs by analyzing total dissolved phosphorus (TDP) concentrations and N/P ratios. Future work could focus on developing a multi-parameter risk index for water bloom prediction. This index would incorporate key variables such as nutrient levels (e.g., phosphorus and nitrogen), climatic factors, and hydrodynamic conditions to improve the accuracy and comprehensiveness of risk assessments.

The Monte-Carlo-based uncertainty analysis provided probabilistic insights into the reliability of the predictions. However, uncertainties remain due to potential measurement errors in water quality parameters and inherent model assumptions.

From an application standpoint, integrating real-time monitoring systems with AI-driven models remains a promising yet underexplored area. While this study employed historical datasets for training and validation, deploying sensor networks capable of real-time data acquisition could significantly enhance the practical utility of predictive frameworks. Implementing Internet of Things (IoT)-based water quality monitoring and cloud-based AI analytics would enable automated early warning systems for bloom risk assessment and water resource management.

4. Conclusions

This study utilized the different ML/DL models, including MLP,

XGBoost, LSSVR, GRUs, and CNNs, to predict Chl-a and DO in the dam reservoir in Iran. Three feature selection approaches, including BFSA, GT, and SHAP, were used to determine the most important features in predicting Chl-a and DO after removing variables with high multicollinearity. The study applied an advanced technique to quantitatively evaluate risk assessment of water bloom occurrence using Chl-a concentrations. The findings of this study responded to objectives as follows:

- Feature selection and optimal input variables: Advanced feature selection techniques identified the most relevant predictors for Chl-a and DO. For Chl-a prediction, pH, temperature, EC, and Depth variables emerged as the most significant variables. At the same time, for DO, salinity and EC were determined as the most influential predictors based on the BFSA method. Considering both BFSA and SHAP methods ensured that only the most statistically significant and relevant features were utilized, leading to improved model interpretability and accuracy.
- Model performance and comparative analysis: XGBoost demonstrated the highest accuracy in predicting Chl-a concentrations among individual models, outperforming traditional ML and DL approaches. For DO prediction, CNNs and GRUs models showed superior predictive capabilities, effectively capturing spatial variations in the reservoir.
- Hybrid model enhancement: hybrid ML-DL frameworks, particularly XGBoost-CNNs and XGBoost-GRUs, significantly improved prediction accuracy of Chl-a compared to standalone models. The hybrid XGBoost-CNNs model achieved the highest performance in predicting Chl-a, leveraging the strength of tree-based learning and CNNs' capability to capture spatial patterns.
- Uncertainty analysis and risk assessment: Monte-Carlo-based uncertainty analysis demonstrated that the predictive models provided reliable estimates, with the XGBoost-CNNs model achieving a high coverage probability (94.16 % in training and 79.37 % in testing) for Chl-a predictions and the CNNs model attaining 100 % coverage probability for DO predictions. The risk assessment framework classified water bloom risks into two categories: no risk and low risk based on Chl-a concentrations. The results revealed that mid-depth regions (5–13 m) exhibited the highest risk.
- Application to Dam reservoir: The study successfully applied the predictive framework to Dam, an under-researched reservoir in Iran's semi-arid region. The findings provided critical insights into the spatial and vertical distributions of Chl-a and DO, demonstrating the applicability of ML and DL models for real-time water quality assessment in deep reservoirs. The validation of models using independent data from the B2 section of the reservoir confirmed their strong generalization ability and potential for future deployment in similar water bodies.

CRediT authorship contribution statement

Akram Seifi: Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Methodology, Formal analysis, Conceptualization. **Hossien Riahi Madvar:** Writing – original draft, Visualization, Validation, Resources, Investigation, Formal analysis, Data curation. **Rouhollah Davarpanah:** Writing – original draft, Software, Methodology, Investigation, Formal analysis. **Mumtaz Ali:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Methodology, Formal analysis. **Abdul-Wahab Mashat:** Writing – original draft, Visualization, Formal analysis.

Declaration of competing interest

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

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jwpe.2025.108341.

Data availability

Data will be made available on request.

References

[1] S.B. Watson, C. Miller, G. Arhonditsis, G.L. Boyer, W. Carmichael, M.N. Charlton, S.W. Wilhelm, The re-eutrophication of Lake Erie: harmful algal blooms and hypoxia, Harmful Algae 56 (2016) 44–66.
[2] Z. Wu, X. Wang, Y. Chen, Y. Cai, J. Deng, Assessing river water quality using water quality index in Lake Taihu Basin, China, Sci. Total Environ. 612 (2018) 914–922.
[3] P.M. Glibert, Harmful algae at the complex nexus of eutrophication and climate change, Harmful Algae 91 (2020) 101583.
[4] J. Huisman, G.A. Codd, H.W. Paerl, B.W. Ibelings, J.M. Verspagen, P.M. Visser, Cyanobacterial blooms, Nat. Rev. Microbiol. 16 (8) (2018) 471–483.
[5] R. Noori, E. Ansari, R. Bhattarai, Q. Tang, S. Aradpour, M. Maghrebi, B. Kløve, Complex dynamics of water quality mixing in a warm mono-mictic reservoir, Sci. Total Environ. 777 (2021) 146097.
[6] R. Noori, M. Noury, M.K. Poshtegal, M. Sadrinasab, M. Mahdian, R. Bhattarai, S. Abolfathi, Thermal stratification and mixing of dam reservoirs in Iran, Watershed Ecology and the Environment 6 (2024) 138–145.
[7] K.B. Kim, M.K. Jung, Y.F. Tsang, H.H. Kwon, Stochastic modeling of chlorophyll-a for probabilistic assessment and monitoring of algae blooms in the Lower Nakdong River, South Korea, J. Hazard. Mater. 400 (2020) 123066.
[8] W. Luo, S. Zhu, S. Wu, J. Dai, Comparing artificial intelligence techniques for chlorophyll-a prediction in US lakes, Environ. Sci. Pollut. Res. 26 (2019) 30524–30532.
[9] Y. Song, Hydrodynamic impacts on algal blooms in reservoirs and bloom mitigation using reservoir operation strategies: a review, J. Hydrol. 620 (2023) 129375.
[10] J.W. Han, T. Kim, S. Lee, T. Kang, J.K. Im, Machine learning and explainable AI for chlorophyll-a prediction in Namhan River watershed, South Korea, Ecol. Indic. 166 (2024) 112361.
[11] X. Zhu, H. Guo, J.J. Huang, S. Tian, W. Xu, Y. Mai, An ensemble machine learning model for water quality estimation in coastal area based on remote sensing imagery, J. Environ. Manag. 323 (2022) 116187.
[12] L. Gatz, Freshwater harmful algal blooms: causes, challenges, and policy considerations (updated), in: Key Congressional Reports for September 2019: Part VII, Geetharamani, 2020, pp. 131–171.
[13] J. Yang, Y. Zheng, W. Zhang, Y. Zhou, Y. Zhang, Comparative analysis of machine learning methods for prediction of chlorophyll-a in a river with different hydrology characteristics: a case study in Fuchun River, China, J. Environ. Manag. 364 (2024) 121386.
[14] B.D. Shoener, S.M. Schramm, F. Béline, O. Bernard, C. Martínez, B.G. Plósz, J. S. Guest, Microalgae and cyanobacteria modeling in water resource recovery facilities: a critical review, Water Research X 2 (2019) 100024.
[15] K.M. Kim, J.H. Ahn, Machine learning predictions of chlorophyll-a in the Han river basin, Korea, J. Environ. Manag. 318 (2022) 115636.
[16] Y. Park, K.H. Cho, J. Park, S.M. Cha, J.H. Kim, Development of early-warning protocol for predicting chlorophyll-a concentration using machine learning models in freshwater and estuarine reservoirs, Korea, Sci. Total Environ. 502 (2015) 31–41.
[17] Z. Du, M. Qin, F. Zhang, R. Liu, Multistep-ahead forecasting of chlorophyll a using a wavelet nonlinear autoregressive network, Knowl.-Based Syst. 160 (2018) 61–70.
[18] H.I. Kim, D. Kim, M. Mahdian, M.M. Salamattalab, S.M. Bateni, R. Noori, Incorporation of water quality index models with machine learning-based techniques for real-time assessment of aquatic ecosystems, Environ. Pollut. 355 (2024) 124242.
[19] M.S. Zare, M.R. Nikoo, G. Al-Rawas, M. Al-Wardy, R. Nazari, M. Karimi, A. H. Gandomi, Integrating machine learning and data envelopment analysis for reliable reservoir water quality index assessment considering uncertainty, Hydrol. Sci. J. 70 (7) (2025) 1109–1124.
[20] A. Abbas, M. Park, S.S. Baek, K.H. Cho, Deep learning-based algorithms for long-term prediction of chlorophyll-a in catchment streams, J. Hydrol. 626 (2023) 130240.
[21] T. Chen, C. Guestrin, Xgboost: a scalable tree-boosting system, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785–794 (August).
[22] W. Tian, Z. Liao, X. Wang, Transfer learning for neural network model in chlorophyll-a dynamics prediction, Environ. Sci. Pollut. Res. 26 (2019) 29857–29871.
[23] H. Huang, J. Zhang, Prediction of chlorophyll a and risk assessment of water blooms in Poyang Lake based on a machine learning method, Environ. Pollut. 347 (2024) 123501.

[24] Z. Mozafari, R. Noori, S.M. Siadatmousavi, H. Afzalimehr, J. Azizpour, Satellite-based monitoring of eutrophication in the earth's largest transboundary lake, GeoHealth 7 (5) (2023) e2022GH000770.
[25] A. Shamloo, S. Sima, Investigating the potential of remote sensing-based machine-learning algorithms to model Secchi-disk depth, total phosphorus, and chlorophyll-a in Lake Urmia, J. Great Lakes Res. 50 (4) (2024) 102370.
[26] Z. Mozafari, R. Noori, S.M. Bateni, C. Jun, D. Kim, M.J. Saravani, S. Abolfathi, Impact of climatic factors on eutrophication in the world's largest lake, Ecol. Indic. 175 (2025) 113497.
[27] M. Fooladi, M.R. Nikoo, R. Mirghafari, C.A. Madramootoo, G. Al-Rawas, R. Nazari, Robust clustering-based hybrid technique enabling reliable reservoir water quality prediction with uncertainty quantification and spatial analysis, J. Environ. Manag. 362 (2024) 121259.
[28] A. Najah, A. El-Shafie, O.A. Karim, A.H. El-Shafie, Application of artificial neural networks for water quality prediction, Neural Comput. & Applic. 22 (1) (2013) 187–201.
[29] M.B. Kursa, W.R. Rudnicki, Feature selection with the Boruta package, J. Stat. Softw. 36 (11) (2010) 1–13.
[30] S.M. Lundberg, S.I. Lee, A unified approach to interpreting model predictions, Adv. Neural Inf. Proces. Syst. 30 (2017).
[31] H. Zhou, J. Zhang, Y. Zhou, X. Guo, Y. Ma, A feature selection algorithm of decision tree based on feature weight, Expert Syst. Appl. 164 (2021) 113842.
[32] D. Goldstein, C. Aldrich, Q. Shao, L. O'Connor, A field-scale framework for assessing the influence of measure-while-drilling variables on geotechnical characterization using a Boruta-SHAP approach, Mining 5 (1) (2025) 20.
[33] J. Park, W.H. Lee, K.T. Kim, C.Y. Park, S. Lee, T.Y. Heo, Interpretation of ensemble learning to predict water quality using explainable artificial intelligence, Sci. Total Environ. 832 (2022) 155070.
[34] G.U. Mahesswari, P.U. Maheswari, SmartScanPCOS: a feature-driven approach to cutting-edge prediction of polycystic ovary syndrome using machine learning and explainable artificial intelligence, Heliyon 11 (2024) e39205.
[35] Z. Ebrahimi-Khusfi, A.R. Nafarzadegan, F. Dargahian, Predicting the number of dusty days around the desert wetlands in southeastern Iran using feature selection and machine learning techniques, Ecol. Indic. 125 (2021) 107499.
[36] H. Gholami, A. Mohamadifar, A.L. Collins, Spatial mapping of the provenance of storm dust: application of data mining and ensemble modelling, Atmos. Res. 233 (2020) 104716.
[37] D.T. Bui, B. Pradhan, O. Lofman, I. Revhaug, O.B. Dick, Spatial prediction of landslide hazards in Hoa Binh province (Vietnam): a comparative assessment of the efficacy of evidential belief functions and fuzzy logic models, Catena 96 (2012) 28–40.
[38] M. Jamei, M. Ali, B. Karimi, M. Karbasi, A.A. Farooque, Z.M. Yaseen, Surface water electrical conductivity and bicarbonate ion determination using a smart hybridization of optimal Boruta package with Elman recurrent neural network, Process. Saf. Environ. Prot. 174 (2023) 115–134.
[39] H. Gholami, A. Mohammadifar, S. Golzari, D.G. Kaskaoutis, A.L. Collins, Using the Boruta algorithm and deep learning models for mapping land susceptibility to atmospheric dust emissions in Iran, Aeolian Res. 50 (2021) 100682.
[40] R.K. Makumbura, L. Mampitiya, N. Rathnayake, D. Meddage, S. Henna, T. L. Dang, Y. Hoshino, U. Rathnayake, Advancing water quality assessment and prediction using machine learning models, coupled with explainable artificial intelligence (XAI) techniques like Shapley additive explanations (SHAP) for interpreting the black-box nature, Res. Eng. Des. 23 (2024) 102831.
[41] S. Wang, H. Peng, S. Liang, Prediction of estuarine water quality using interpretable machine learning approach, J. Hydrol. 605 (2021) 127320, https://doi.org/10.1016/j.jhydrol.2021.127320.
[42] R. Noori, A. Karbassi, M.S. Sabahi, Evaluation of PCA and gamma test techniques on ANN operation for weekly solid waste prediction, J. Environ. Manag. 91 (3) (2009) 767–771.
[43] S. Azadi, H. Amiri, M.G. Mooselu, H. Liltved, R. Castro-Muñoz, X. Sun, G. Boczkaj, Network design for surface water quality monitoring in a road construction project using Gamma Test theory, Water Resources and Industry 26 (2021) 100162.
[44] H. Lu, X. Ma, Hybrid decision tree-based machine learning models for short-term water quality prediction, Chemosphere 249 (2020) 126169.
[45] M. Niazkar, A. Menapace, B. Brentan, R. Piraei, D. Jimenez, P. Dhawan, M. Righetti, Applications of XGBoost in water resources engineering: a systematic literature review (Dec 2018–may 2023), Environ. Model. Softw. 174 (2024) 105971.
[46] H.W. Lee, M. Kim, H.W. Son, B. Min, J.H. Choi, Machine-learning-based water quality management of river with serial impoundments in the Republic of Korea, J. Hydrol. Reg. Stud. 41 (2022) 101069.
[47] H. Isik, T. Akkan, Water quality assessment with artificial neural network models: performance comparison between SMN, MLP and PS-ANN methodologies, Arab. J. Sci. Eng. 50 (1) (2025) 369–387.
[48] M. Alizamir, K. Moradveisi, K.O. Ahmed, J. Bahrami, S. Kim, S. Heddam, An efficient data fusion model based on Bayesian model averaging for robust water quality prediction using deep learning strategies, Expert Syst. Appl. 261 (2025) 125499.
[49] Z. Guo, G. Bai, Application of least squares support vector machine for regression to reliability analysis, Chin. J. Aeronaut. 22 (2) (2009) 160–166.
[50] J.C.Y. Ngu, W.S. Yeo, Prediction of dissolved oxygen using least square support vector regression model, in: 2022 International Conference on Green Energy, Computing and Sustainable Technology (GECOST), 2022, pp. 70–74.
[51] J. Chung, C. Gulcehre, K. Cho, Y. Bengio, Empirical evaluation of gated recurrent neural networks on sequence modeling, arXiv preprint, 2014. arXiv:1412.3555.
[52] X. Cao, Y. Liu, J. Wang, C. Liu, Q. Duan, Prediction of dissolved oxygen in pond culture water based on K-means clustering and gated recurrent unit neural network, Aquac. Eng. 91 (2020) 102122.
[53] S. Huang, Y. Wang, J. Xia, Which riverine water quality parameters can be predicted by meteorologically-driven deep learning? Sci. Total Environ. 946 (2024) 174357.
[54] Z. Li, F. Liu, W. Yang, S. Peng, J. Zhou, A survey of convolutional neural networks: analysis, applications, and prospects, IEEE Transactions on Neural Networks and Learning Systems 33 (12) (2021) 6999–7019.
[55] N. Kalchbrenner, E. Grefenstette, P. Blunsom, A Convolutional Neural Network for Modelling Sentences, arXiv (Cornell University), 2014.
[56] Q. Zhang, J. Xiao, C. Tian, J.C. Lin, S. Zhang, A robust deformed convolutional neural network (CNN) for image denoising, CAAI Trans. Intell. Technol. 8 (2) (2022) 331–342.
[58] S. Majnooni, M. Fooladi, M.R. Nikoo, G. Al-Rawas, A.T. Haghighi, R. Nazari, M. Al-Wardy, A.H. Gandomi, Smarter water quality monitoring in reservoirs using interpretable deep learning models and feature importance analysis, J Water Process Eng 60 (2024) 105187.
[59] J. Sha, X. Li, M. Zhang, Z. Wang, Comparison of forecasting models for real-time monitoring of water quality parameters based on hybrid deep learning neural networks, Water 13 (11) (2021) 1547.
[60] X. Wang, X. Tang, M. Zhu, Z. Liu, G. Wang, Predicting abrupt depletion of dissolved oxygen in Chaohu lake using CNN-BiLSTM with improved attention mechanism, Water Res. 261 (2024) 122027.
[61] M.G. Zamani, M.R. Nikoo, G. Al-Rawas, R. Nazari, D. Rastad, A.H. Gandomi, Hybrid WT–CNN–GRU-based model for the estimation of reservoir water quality variables considering spatio-temporal features, J. Environ. Manag. 358 (2024) 120756.
[62] S.H. Masteali, M. Bayat, P. Bettinger, M. Ghorbanpour, Uncertainty analysis of linear and non-linear regression models in the modeling of water quality in the Caspian Sea basin: application of Monte-Carlo method, Ecol. Indic. 170 (2025) 112979.
[63] A. Seifi, M. Dehghani, V.P. Singh, Uncertainty analysis of water quality index (WQI) for groundwater quality evaluation: application of Monte-Carlo method for weight allocation, Ecol. Indic. 117 (2020) 106653.
[64] B.W. Ibelings, S.C. Maberly, Photoinhibition and the availability of inorganic carbon restrict photosynthesis by surface blooms of cyanobacteria, Limnol. Oceanogr. 43 (3) (1998) 408–419.
[65] Y.J. Jiang, W. He, W.X. Liu, N. Qin, H.L. Ouyang, Q.M. Wang, F.L. Xu, The seasonal and spatial variations of phytoplankton community and their correlation with environmental factors in a large eutrophic Chinese lake (Lake Chaohu), Ecol. Indic. 40 (2014) 58–67.
[66] Z. Yu, K. Yang, Y. Luo, C. Shang, Spatial-temporal process simulation and prediction of chlorophyll-a concentration in Dianchi Lake based on wavelet analysis and long-short term memory network, J. Hydrol. 582 (2020) 124488.
[67] H.G. Kim, S. Hong, K.S. Jeong, D.K. Kim, G.J. Joo, Determination of sensitive variables regardless of hydrological alteration in artificial neural network model of chlorophyll a: case study of Nakdong River, Ecol. Model. 398 (2019) 67–76.
[68] S.M.C. Casanova, E.A. Panarelli, R. Henry, Rotifer abundance, biomass, and secondary production after the recovery of hydrologic connectivity between a river and two marginal lakes (São Paulo, Brazil), Limnologica 39 (4) (2009) 292–301.
[69] M.J. Saravani, R. Noori, C. Jun, D. Kim, S.M. Bateni, P. Kianmehr, R.I. Woolway, Predicting chlorophyll-a concentrations in the world's largest lakes using Kolmogorov-Arnold networks, Environ. Sci. Technol. 59 (3) (2025) 1801–1810.
[70] Q. Shen, J. Zhu, L. Cheng, J. Zhang, Z. Zhang, X. Xu, Enhanced algae removal by drinking water treatment of chlorination coupled with coagulation, Desalination 271 (1–3) (2011) 236–240.
[71] S.P. Singh, P. Singh, Effect of temperature and light on the growth of algae species: a review, Renew. Sust. Energ. Rev. 50 (2015) 431–444.
[72] B. Boehrer, M. Schultze, Stratification of lakes, Rev. Geophys. 46 (2) (2008).
[73] L. Håkanson, M. Jansson, Principles of Lake Sedimentology 316, Springer-verlag, Berlin, 1983.
[74] D.P. Hamilton, S.F. Mitchell, An empirical model for sediment resuspension in shallow lakes, Hydrobiologia 317 (1996) 209–220.
[75] P.L. Brezonik, R.W. Bouchard Jr., J.C. Finlay, C.G. Griffin, L.G. Olmanson, J. P. Anderson, R. Hozalski, Color, chlorophyll a, and suspended solids effects on Secchi depth in lakes: implications for trophic state assessment, Ecol. Appl. 29 (3) (2019) e01871.
[76] O.T. Lind, The effect of non-algal turbidity on the relationship of Secchi depth to chlorophyll a, Hydrobiologia 140 (1986) 27–35.
[77] D. Hamilton, S. Mitchell, Wave-induced shear stresses, plant nutrients and chlorophyll in seven shallow lakes, Freshw. Biol. 38 (1) (1997) 159–168.
[78] R. Quinlan, A. Filazzola, O. Mahdiyan, A. Shuvo, K. Blagrave, C. Ewins, S. Sharma, Relationships of total phosphorus and chlorophyll in lakes worldwide, Limnol. Oceanogr. 66 (2) (2021) 392–404.
[79] W. Zou, G. Zhu, Y. Cai, A. Vilmi, H. Xu, M. Zhu, B. Qin, Relationships between nutrient, chlorophyll a and Secchi depth in lakes of the Chinese Eastern Plains ecoregion: implications for eutrophication management, J. Environ. Manag. 260 (2020) 109923.
[80] M. Radwan, P. Willems, A. El-Sadek, J. Berlamont, Modelling of dissolved oxygen and biochemical oxygen demand in river water using a detailed and a simplified model, Int. J. River Basin Manag. 1 (2) (2003) 97–103.
[81] D.A. Gardner-Dale, I.M. Bradley, J.S. Guest, Influence of solids residence time and carbon storage on nitrogen and phosphorus recovery by microalgae across diel cycles, Water Res. 121 (2017) 231–239.

[82] N.N. Rabalais, W.J. Cai, J. Carstensen, D.J. Conley, B. Fry, X. Hu, J. Zhang, Eutrophication-driven deoxygenation in the coastal ocean, Oceanography 27 (1) (2014) 172–183.
[83] K. Kunlasak, C. Chitmanat, N. Whangchai, J. Promya, L. Lebel, Relationships of dissolved oxygen with chlorophyll-a and phytoplankton composition in tilapia ponds, Int. J. Geosci. 4 (05) (2013) 46.
[84] J. Crossman, G. Bussi, P.G. Whitehead, D. Butterfield, E. Lannergård, M.N. Futter, A new, catchment-scale integrated water quality model of phosphorus, dissolved oxygen, biochemical oxygen demand and phytoplankton: INCA-Phosphorus Ecology (PEco), Water 13 (5) (2021) 723.
[85] A.M. Morgan, T.V. Royer, M.B. David, L.E. Gentry, Relationships among nutrients, chlorophyll-a, and dissolved oxygen in agricultural streams in Illinois, J. Environ. Qual. 35 (4) (2006) 1110–1117.
[86] M.G. Uddin, S. Nash, A. Rahman, A.I. Olbert, Assessing optimization techniques for improving water quality model, J. Clean. Prod. 385 (2023) 135671.
[87] S. Luo, B. Wang, Q. Gao, Y. Wang, X. Pang, Stacking integration algorithm based on CNN-BiLSTM-Attention with XGBoost for short-term electricity load forecasting, Energy Rep. 12 (2024) 2676–2689.
[88] S. Zhu, B. Hrnjica, M. Ptak, A. Choiński, B. Sivakumar, Forecasting of water level in multiple temperate lakes using machine learning models, J. Hydrol. 585 (2020) 124819.
[89] S.L. Ellison, A. Williams, Quantifying uncertainty in analytical measurement, in: CITAC Guide Number, Eurachem/CITAC, Teddington, UK, 2012, p. 141.
[90] G. Premalatha, V.T. Bai, Design and implementation of intelligent patient in-house monitoring system based on efficient XGBoost-CNN approach, Cogn. Neurodyn. 16 (5) (2022) 1135–1149.
[91] R. Jayakarthik, K.T. Chandrashekhara, O. Sampath, D. Kumar, L. Biban, J. P. Maroor, S.L. Malluvalasa, Climate impact prediction: whale-optimized conv-XGBoost with remote sensing and sociological data, Remote Sensing in Earth Systems Sciences 7 (4) (2024) 443–456.
[92] S. Prakash, K. Sangeetha, Systems classification of air pollutants using Adam optimized CNN with XGBoost feature selection, Analog Integr. Circ. Sig. Process 122 (3) (2025) 35.
[93] Y. Zhang, P. Fitch, M.P. Vilas, P.J. Thorburn, Applying multi-layer artificial neural network and mutual information to the prediction of trends in dissolved oxygen, Front. Environ. Sci. 7 (2019) 46.
[94] S.V. Moghadam, A. Sharafati, H. Feizi, S.M.S. Marjaie, S.B.H.S. Asadollah, D. Motta, An efficient strategy for predicting river dissolved oxygen concentration: application of deep recurrent neural network model, Environ. Monit. Assess. 193 (2021) 1–18.
[95] Tung Tiyasha, T.M. Tung, Z.M. Yaseen, Deep learning for prediction of water quality index classification: tropical catchment environmental assessment, Nat. Resour. Res. 30 (6) (2021) 4235–4254.
[96] S. Heddam, S. Kim, A.D. Mehr, M. Zounemat-Kermani, A. Malik, A. Elbeltagi, O. Kisi, Predicting dissolved oxygen concentration in river using new advanced machines learning: long-short term memory (LSTM) deep learning, in: Computers in Earth and Environmental Sciences, Elsevier, 2022, pp. 1–20.
[97] E. Bertone, A. Ajmar, F.G. Tonolo, R.J. Dunn, N.J. Doriean, W.W. Bennett, J. Purandare, Satellite-based estimation of total suspended solids and chlorophyll-a concentrations for the Gold Coast Broadwater, Australia, Mar. Pollut. Bull. 201 (2024) 116217.
[98] A.M. Saranathan, B. Smith, N. Pahlevan, Per-pixel uncertainty quantification and reporting for satellite-derived chlorophyll-a estimates via mixture density networks, IEEE Trans. Geosci. Remote Sens. 61 (2023) 1–18.
[99] M. Abdar, F. Pourpanah, S. Hussain, D. Rezazadegan, L. Liu, M. Ghavamzadeh, S. Nahavandi, A review of uncertainty quantification in deep learning: techniques, applications and challenges, Information fusion 76 (2021) 243–297.
[100] E. Hüllermeier, W. Waegeman, Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods, Mach. Learn. 110 (3) (2021) 457–506.
[101] H. Wang, M. Convertino, Algal bloom ties: systemic biogeochemical stress and Chlorophyll-a shift forecasting, Ecol. Indic. 154 (2023) 110760.
[102] M. Sutula, R. Kudela, J.D. Hagy III, L.W. Harding Jr., D. Senn, J.E. Cloern, M. Beck, Novel analyses of long-term data provide a scientific basis for chlorophyll-a thresholds in San Francisco Bay, Estuar. Coast. Shelf Sci. 197 (2017) 107–118.
[103] P. Shi, M. Zhu, R. You, H. Li, W. Zou, H. Xu, G. Zhu, Rainstorm events trigger algal blooms in a large oligotrophic reservoir, J. Hydrol. 622 (2023) 129711.
[104] H. Meng, J. Zhang, Z. Zheng, Y. Lai, H. Geng, Risk assessment and spatio-temporal characteristics analysis of water bloom in three large-scale eutrophic reservoirs in Fujian Province, China, Ecol. Indic. 158 (2024) 111539.
[105] L. Xue, J. Hu, Z. Wang, G. Pei, L. Chen, Assessing risks of algal blooms in water transfer based on algal growth potential, Environ. Monit. Assess. 195 (7) (2023) 871.
[106] I. Busari, D. Sahoo, R.B. Jana, Prediction of chlorophyll-a as an indicator of harmful algal blooms using deep learning with Bayesian approximation for uncertainty assessment, J. Hydrol. 630 (2024) 130627.
[107] W. Huo, W. Li, Z. Zhang, C. Sun, F. Zhou, G. Gong, Performance prediction of proton-exchange membrane fuel cell based on convolutional neural network and random forest feature selection, Energy Convers. Manag. 243 (2021) 114367.
[108] S.V. Smith, J.T. Hollibaugh, Water, salt, and nutrient exchanges in San Francisco Bay, Limnol. Oceanogr. 51 (1part2) (2006) 504–517.