Algae Development in Rivers with Artificially Constructed Weirs: Dominant Influence of Discharge over Temperature

Hyunju Kim [a], Gyesik Lee [b,*], Chang-Gu Lee [c], Seong-Jik Park [d,*]

[a] Faculty of Liberal Education, Seoul National University, Seoul, 08826, Republic of Korea

[b] School of Computer Engineering and Applied Mathematics, Hankyong National University, Anseong, 17579, Republic of Korea

[c] Department of Environmental and Safety Engineering, Ajou University, Suwon 16499, Republic of Korea

[d] Department of Bioresources and Rural System Engineering, Hankyong National University, Anseong, 17579, Republic of Korea

* ***Corresponding author***: S. J. Park

E-mail address: parkseongjik@hknu.ac.kr

ORCID: 0000-0003-2122-5498

G. Lee

E-mail address: gslee@hknu.ac.kr

Abstract

Algal blooms contribute to water quality degradation, unpleasant odors, taste issues, and the presence of harmful substances in artificially constructed weirs. Mitigating these adverse effects through effective algal bloom management requires identifying the contributing factors and predicting algal concentrations. This study focused on the upstream region of the Seungchon Weir in Korea, which is characterized by elevated levels of total nitrogen and phosphorus due to a significant influx of water from a sewage treatment plant. We employed four distinct machine learning models to predict chlorophyll-a (Chl-a) concentrations and identified the influential variables linked to local algal bloom events. The gradient boosting model enabled an in-depth exploration of the intricate relationships between algal occurrence and water quality parameters, enabling accurate identification of the causal factors. The models identified the discharge flow rate (D-Flow) and water temperature as the primary determinants of Chl-a levels, with feature importance values of 0.236 and 0.212, respectively. Enhanced model precision was achieved by utilizing daily average D-Flow values, with model accuracy and significance of the D-Flow amplifying as the temporal span of daily averaging increased. Elevated Chl-a concentrations correlated with diminished D-Flow and temperature, highlighting the pivotal role of D-Flow in regulating Chl-a concentration. This trend can be attributed to the constrained discharge of the Seungchon Weir during winter. Calculating the requisite D-Flow to maintain a desirable Chl-a concentration of up to 20 mg/m^3 across varying temperatures revealed an escalating demand for D-Flow with rising temperatures. Specific D-Flow ranges, corresponding to each season and temperature condition, were identified as particularly influential on Chl-a concentration. Thus, optimizing Chl-a reduction can be achieved by strategically increasing D-Flow within these specified ranges for each season and temperature variation. This study highlights the importance of maintaining sufficient D-Flow levels to mitigate algal proliferation within river systems featuring weirs.

Keywords: algae bloom; machine learning; temperature; discharge flow rate; constructed weirs; chlorophyll-a

1. Introduction

Algae are single-celled or multicellular organisms that provide food and oxygen to aquatic animals and play vital roles in aquatic ecosystems. However, excessive algal growth, known as algal blooms, can lead to severe environmental problems such as fish mortality, contamination of drinking water, destruction of aquatic habitats, and human diseases caused by algal toxins (Zhou et al., 2014; Hong et al., 2020; Kim et al., 2021a). Algal blooms are a global phenomenon and have caused environmental and social problems in many countries. For example, in Korea, the construction of weirs in four major rivers significantly increased the frequency and severity of algal blooms, mainly due to the reduced flow rate caused by the increased water depth and storage capacity (Lee et al., 2017). Thus, predicting and controlling algal blooms has become a critical issue, and effective control measures are needed.

Chlorophyll is a group of pigments that play a critical role in photosynthesis in all phototrophic organisms, including algae and some bacterial species (Suggett et al., 2010). Monitoring chlorophyll concentrations can provide valuable insights into phytoplankton biomass and nutritional status (Furnas, 1990), making it an indirect indicator of nutrient levels, such as phosphorus and nitrogen, in surface water (Keller et al., 2018). As a widely used water quality parameter, the concentration of chlorophyll-a (Chl-a) is commonly utilized to assess algal levels (Kwak, 2021; Shin et al., 2017). In Korea, the concentrations of Chl-a in rivers, together with cyanobacterial cell densities, are used as criteria for the algal warning system (Park et al., 2015).

Monitoring and predicting surface water quality is crucial for mitigating the potential

damage caused by harmful algae and improving water quality. Traditional methods for predicting algal blooms using conventional water quality indicators are costly, labor-intensive, and time-consuming, posing significant challenges in providing timely monitoring and management interventions (Ly et al., 2023). Model development can help predict algal growth and evolution, enabling proactive measures to be taken (Deng et al., 2021; Ly et al., 2023). Identifying the predictive potential of environmental factors on algal bloom occurrence can facilitate the development of effective management strategies (Yu et al., 2021).

Process-based models have been used since the 1980s to predict algal blooms and elucidate the relationships between algal growth and environmental factors (Deng et al., 2021; Yang et al., 2019). Although these models can provide relatively accurate predictions of water quality parameters, they face challenges when dealing with large quantities of input data (Ahmed et al., 2019). Statistical methods, such as principal component analysis with multiple linear regression, have previously been used to predict Chl-a concentrations too (Çamdevýren et al., 2005).

Recently, machine learning-based prediction has become popular in many scientific fields, especially for the environmental modeling of complex nonlinear phenomena (Kang et al., 2022; Lee et al., 2024; Yu et al., 2021). Machine learning models map the relationships between the inputs and outputs of a system rather than completing complex process mechanisms, enabling accurate predictions of highly nonlinear relationships without prior knowledge of the system (Deng et al., 2021). For example, CEEMDAN-RF and CEEMDAN-XGBoost, two hybrid decision tree (DT)-based models, have been applied to predict water quality based on parameters such as temperature, dissolved oxygen, pH, specific conductance, turbidity, and fluorescent dissolved organic matter (Lu and Ma, 2020). Park et al. (2021) used artificial neural networks and support vector machines to predict algal bloom alert levels by incorporating environmental variables, such as nutrients and meteorological factors, in freshwater reservoirs.

Mozo et al. (2022) developed machine learning models for algal bloom predictions using three years of Chl-a data but did not identify the primary factors controlling algal blooms, as their analysis only focused on a single parameter. Ly et al. (2023) compared five different machine learning models for predicting algal growth, revealing that XGBoost and stacking methods demonstrated superior performance over other models. Additionally, Hong et al. (2023) employed reinforcement learning to autonomously calibrate sequential water quality parameters in the Environmental Fluid Dynamics Code model for Chl-a prediction.

This study aimed to address the significant challenges arising from algal blooms in artificially constructed weirs. The investigation involved identifying the contributing factors and predicting algal concentrations using four different machine learning models: elastic net (EN), decision tree (DT), random forest (RF), and gradient boosting (GB). These models are widely employed in environmental studies. EN is a representative regularized linear regression model that combines the advantages of both ridge and lasso regression and performed well in predicting surface sediment concentrations (Aires et al. 2022). DT, RF, and GB are tree-based regression models commonly used for predicting unknown variables. Yaqub et al. (2022) used DT, RF, and extreme GB models to remove pharmaceutical products from water in a managed aquifer recharge system, with XGBoost showing notably better results. In contrast, Kim et al. (2022) found RF to be suitable for predicting Chl-a levels, and Lian et al. (2021) indicated that RF can effectively predict variations in algal blooms.

This study not only compared model performance in predicting Chl-a concentrations but also analyzed the factors influencing Chl-a concentration in a river with an artificially constructed weir. Among the key determinants influencing Chl-a concentration, the discharge flow rate (D-Flow) stands out. We considered utilizing different temporal average values of D-Flow in our analyses, and we found that extended temporal averaging bolstered model accuracy

and the significance of D-Flow. Additionally, a D-Flow analysis was conducted to determine the necessary discharge for maintaining the desired Chl-a concentration across varying temperatures and seasons, achieved by segmenting the D-Flow distribution according to the Chl-a concentration range. This study offers fundamental insights for practical weir operation, elucidating the factors that impact algal blooms in river systems with weir structures, and proposing optimal D-Flow to sustain favorable Chl-a concentrations.

2. Materials and Methods

2.1. Study area and data acquisition

The Seungchon Weir in the Yeongsan River was constructed as part of the Four Major Rivers Project in Korea, which started in October 2009 and was completed in May 2012. The concentration of Chl-a in the Yeongsan River increased significantly after the construction of the weir compared to the other three major rivers (Lee et al., 2017), and has the most severe water pollution among the four major rivers in the project. Reports indicate eutrophication due to high phosphorus and nitrogen concentrations, low oxygen depletion, and high phytoplankton concentrations downstream throughout the year (Son et al., 2013). The Yeongsan River, characterized by a short length and a small basin area, consistently experiences insufficient water flow. Approximately 70% of the mainstream's flow is derived from discharges from sewage treatment plants. Furthermore, the Yeongsan River basin exhibits a low percentage of forested areas and a high proportion of agricultural land, contributing to the deterioration of water quality due to non-point pollution sources. The Seungchon Weir is 512 m long and 12 m wide, with a management elevation level of 7.5 m upstream, providing 9 million m^3 of water (Fig. 1). The weir has a multifunctional structure consisting of four truss-type movable liftgate weirs (total length 180 m, elevation above mean sea level [ELm] 2.5 m) and three fixed weirs (total length 304.5 m, ELm 7.5 m). It includes two fishways, one on the left side of the weir

and the other along the original river path. Additionally, an operational hydroelectric power plant with a maximum power generation capacity of 800 kW and a water usage rate of 28 m^3/s has been installed and is operational (Chong et al., 2015). Water quality data for this study were collected at a latitude of 35° 4' 14" and longitude of 126° 46' 35", 1.1 km upstream of the Seungchon Weir.

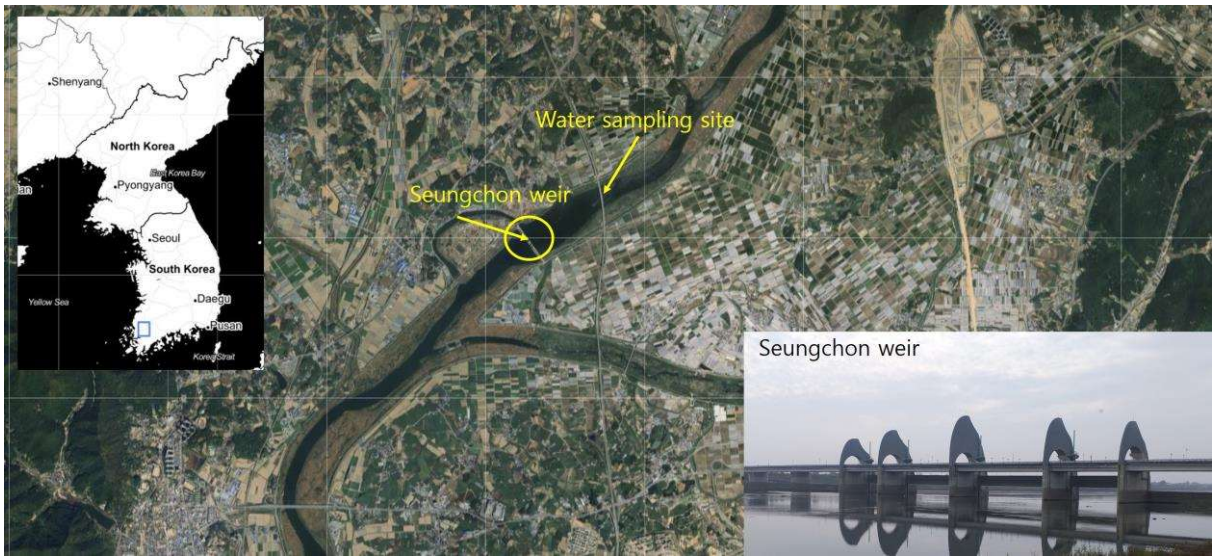


Fig. 1. Map and digital image of the study area encompassing the Seungchon Weir and the water sampling site.

This study utilized hourly water quality data obtained from a real-time water quality information system (https://water.nier.go.kr) at the National Institute of Environmental Sciences of the Korean Ministry of Environment (KME). A total of 82,257 observations from January 1, 2013, to May 23, 2022, were analyzed, with parameters including water temperature (Temp), hydrogen ion concentration (pH), electrical conductivity (EC), dissolved oxygen (DO), total organic carbon (TOC), total nitrogen (TN), total phosphorus (TP), and Chl-a concentrations. Data on the D-Flow and water level (WL) at the upstream location were obtained from the Han River Flood Control Office (https://www.hrfco.go.kr) and converted

into 1 h intervals to correspond to the hourly water quality data.

2.2. Data sets

Six of the eight water quality parameters used in this study (listed above had missing values: Temp (21.5%), EC (21.6%), TOC (28.7%), TN (27.1%), TP (28.3%), and Chl-a (23.6%). To fill in the missing values, we used linear interpolation with a 12-hour window in each direction. If data were missing for more than 24 hours, they were excluded from the input. In addition, the upstream WL was maintained at ELm 7.5 m, and data with a WL greater than 10 m were considered outliers and excluded.

2.3. Machine learning models

The models used to predict the Chl-a concentration were EN, DT, RF, and GB.

2.3.1. EN

ENs are linear regression models that combine L1- and L2-regularization to regularize model parameters called weights, using the L1-norm $\| \cdot \|_1$ and L2-norm $\| \cdot \|_2$, respectively (Zou and Hastie, 2005). During training, the EN model attempts to learn a weight matrix W that minimizes the following formula (Eq. 1) (Friedman et al., 2010):

$$(\|X\,W - y\|_2^2 + r\,\alpha\,\|W\|_1 + \frac{1}{2}(1 - r)\,\alpha\,\|W\|_2^2) \quad (1)$$

where X is the matrix representing the training set, y is the target vector, r is the ratio of L1-regularization, and α the intensity of regularization. The L1-regularization performs variable selection by eliminating the weights of the least important variables, while the L2-regularization keeps the weights as small as possible to minimize the model's variance (Ogutu et al., 2012).

2.3.2. DT

DTs are nonlinear models with a binary tree structure. Each node in a DT holds information from a dataset, and the process of splitting a node corresponds to the partitioning of the dataset into two subsets.

Node t of the binary tree is split into two child nodes t^L and t^R to maximize the reduction in misclassification cost (Breiman et al., 1984). An algorithm called the classification and regression tree (CART) performs this task, as demonstrated in Fig. 2a.

In node t, for instance, the CART algorithm chooses a variable j and a value c, and decides for each data x whether x should belong to node t^L or t^R depending on the result of the comparison $x_j \leq c$, where x_j is the value contained in x for the variable j. The performance of DTs depends on the extent to which the partitioning is executed. Moreover, there is always a risk of overfitting.

2.3.3. RF

RFs combine a series of DTs to prevent model overfitting and improve prediction accuracy (Liu et al., 2012). As shown in Fig. 2b, an RF model simultaneously trains a multitude of DT using a subset of the training set selected by sampling with replacement for each tree. Its final prediction is made by aggregating the predictions of all of the individual trees (Breiman, 2001). The classification tasks were decided by the majority. For regression tasks, the average of individual trees was used.

2.3.4. GB

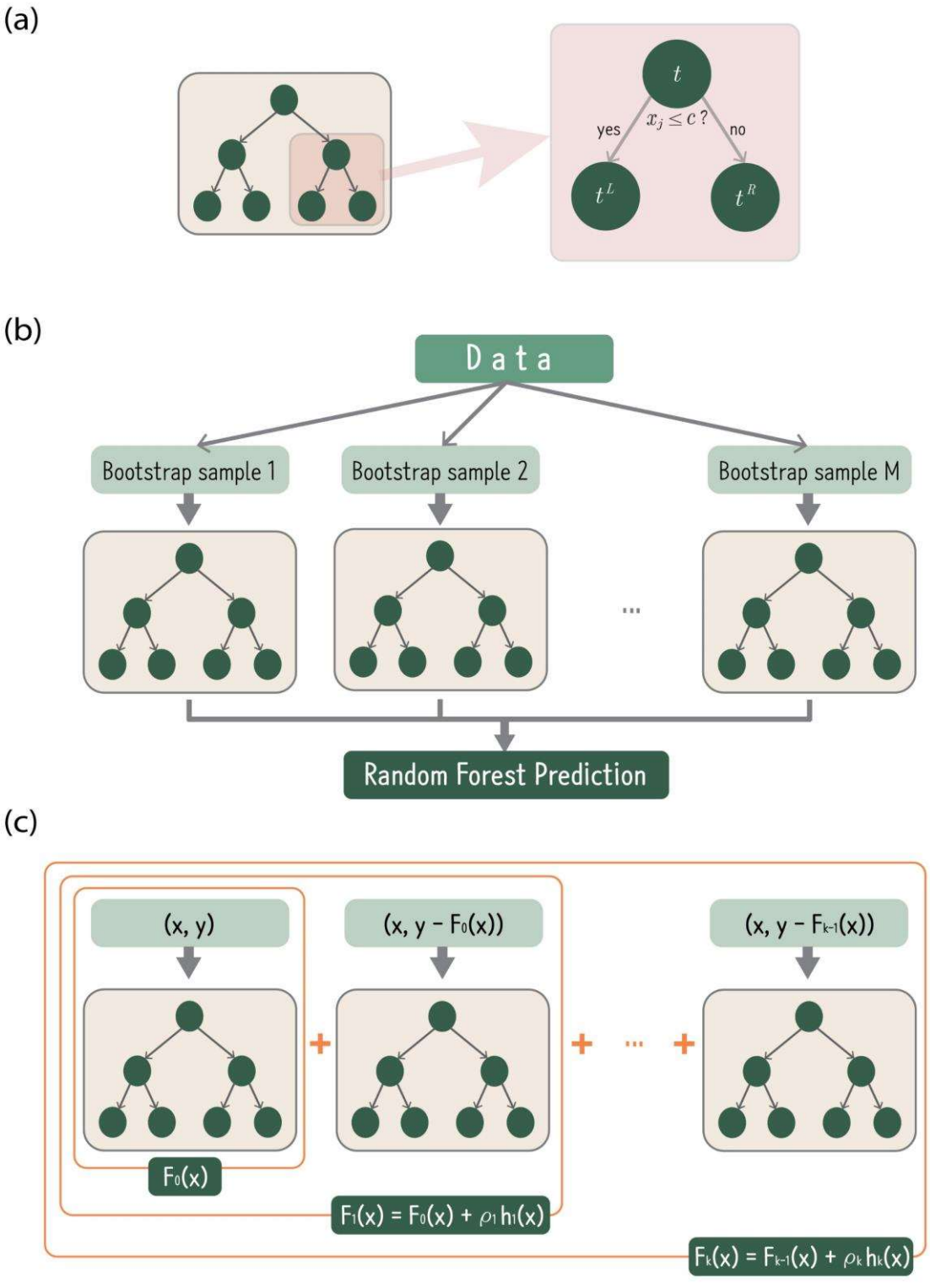
GB is a method for gradually improving prediction performance by sequentially training weak estimators (Makhotin et al., 2019). Friedman (2001, 2002) proposed a GB algorithm that

starts with a DT as the base estimator and cumulatively trains a new model that predicts the residual error of the previous model. Fig. 2c illustrates the proposed algorithm.

Adding the weak estimator h_k trained with the (k-1)-th residual error results in a stronger estimator (Eq. 2):

$$F_k(x) = F_{k-1}(x) + \rho_k \, h_k(x) \quad (2)$$

where F_0 is the initial weak estimator, and ρ_k is the k-th learning rate.



214

215 **Fig. 2.** Schematic diagram of (a) decision tree, (b) random forest, and (c) gradient boosting

models.

2.4. Model performance evaluation

The performance of each model was assessed using two metrics: the root mean squared error (RMSE) and coefficient of determination (R^2). RMSE (Eq. 3) is a widely used regression evaluation metric that measures the average error between the estimated value $\widehat{y_k}$ and true value y_k over all N data samples.

$$RMSE = \sqrt{\frac{1}{N}\sum_{k=1}^{N}(\widehat{y_k} - y_k)^2} \tag{3}$$

R^2 (Eq. 4) is displayed as a value less than or equal to 1 and is used to evaluate the fitting performance of the model. The closer R^2 is to 1, the better the fitting performance of the model

$$R^2 = 1 - \frac{\sum_{k=1}^{N}(\widehat{y_k} - y)^2}{\sum_{k=1}^{N}\left(y_k - \underline{y}\right)^2} \tag{4}$$

where $\underline{y}$ is the average of the original data and $\widehat{y_k}$ and y_k are the predicted and true values, respectively.

2.5. Application of models

The EN, DT, RF, and GB models were implemented using Python's Scikit-learn library (https://scikit-learn.org/stable/), a Python library that provides almost all the tools needed for data analysis and training machine learning models. The dataset was divided into two sets for performance evaluation after training: a training set containing 75% of the dataset and a test set comprising the remaining 25%. However, the regression model performed better when the response variable had a normal distribution. Fig. S1a shows that the distribution of Chl-a was skewed to the left; therefore, we made it approach a normal distribution by log transformation

(Fig. S1b). Hyperparameter tuning for each model was performed through a grid search to optimize model performance. The sklearn library's GridSearchCV was employed to systematically explore various hyperparameter combinations and identify the best-performing set. The optimal hyperparameters for the models utilized in this study are summarized in Table S1.

To assess the effectiveness of four distinct models and identify the feature importance of water quality parameters in predicting Chl-a, the input variables included pH, DO, Temp, EC, TOC, TN, TP, WL, and D-Flow, all obtained concurrently with Chl-a measurements. Additionally, an in-depth investigation was conducted using the GB model, which demonstrated superior accuracy. The model was enhanced by incorporating the average daily values of D-Flow and Temp from the preceding 1 to 14 days.

3. Results and Discussion

3.1. Statistical analysis of water quality and hydraulic parameters

Statistical analysis of the water quality and hydraulic conditions of the Seungchon Weir are provided in Table 1. The TOC concentration fell within a moderate range (level 3), according to the water quality standards of the KME. However, the mean concentrations of TN and TP corresponded to extremely poor water quality (> 1.5 mg/L, level 6) and poor water quality levels (0.10–0.15 mg/L, level 5), respectively, according to KME standards. An excess of phosphorus (> 0.1 mg/L) has been shown to have severely detrimental effects on freshwater systems, contributing to eutrophication (Mng'ong'o et al., 2022). The mean Chl-a concentration was high (35–70 mg/m^3, level 5). These findings indicate a nutrient-rich environment in the Seungchon Weir, indicating high eutrophication and favorable conditions for algal growth (Mng'ong'o et al., 2022).

Table 1.

Statistics of water quality and hydraulic parameters of the Seungchon Weir including water temperature (Temp), electrical conductivity (EC), total organic carbon (TOC), total nitrogen (TN), total phosphorus (TP), water level (WL), discharge flow rate (D-Flow), and chlorophyll-a (Chl-a).

	Temp (℃)	EC (μS/cm)	TOC (mg/L)	TN (mg/L)	TP (mg/L)	WL (ELm)	D-Flow (m^3/s)	Chl-a (mg/ m^3)
Mean	17.962	339.594	4.364	5.397	0.119	6.847	26.508	50.334
SD	8.028	88.196	1.026	2.001	0.064	0.864	62.494	37.231
Min	1.800	92.000	1.600	0.706	0.003	4.038	0.000	1.000
25%	10.000	277.000	3.600	3.817	0.074	5.980	11.400	20.813
50%	18.200	339.000	4.264	5.157	0.106	7.490	15.237	42.500
75%	25.100	411.000	5.000	6.726	0.150	7.530	25.962	72.200
Max	36.100	730.000	19.300	12.357	0.745	8.072	2710.73	453.500

Pearson's correlation analysis was performed using a dataset comprising 49,111 samples, which included all water quality and hydraulic parameters out of a total of 82,257 observations. The correlation analysis is presented as a heatmap plot (Fig. S2). Temperature was significantly and inversely correlated with EC and TN. TN and EC showed strong positive correlations. DO and pH were highly correlated with Chl-a, indicating that both these parameters influenced algal outbreaks. This correlation can be attributed to algae producing oxygen through

photosynthesis and hydroxyl ions by utilizing inorganic carbon (bicarbonate) in the water (Liu et al., 2016).

3.2. Comparison of machine learning models

We employed four distinct machine learning models (EN, DT, RF, and GB) to forecast the Chl-a concentration using pH, DO, Temp, EC, TOC, TN, TP, WL, and D-Flow as input variables (Fig. S3). In the DT, RF, and GB models, pH emerged as the most influential variable, with a feature importance value (FIV) of approximately 0.3. DO ranked fifth in importance across all three models, with a FIV of approximately 0.08. Notably, photosynthesis by algae can lead to elevated DO and pH values (Raven et al., 2020). Consequently, as pH and DO are outcomes of algal occurrence rather than factors influencing its growth, these two variables were excluded from further analysis. The GB and RF models exhibited remarkable performances based on their high R^2 values and low RMSE values, whereas the performance of EN was too low to be useful (Table 2). Since the EN is a linear regression model that assumes a linear relationship between the independent and dependent variables (Zou and Hastie, 2005), it may not be suitable for capturing the complex interactions between Chl-a and water quality parameters. Figure 3 shows the Chl-a concentration observed and predicted by each of the four models for a randomly selected subset of 200 data samples from the test set. Notably, for the GB and RF models, the observed and predicted line plots coincided approximately, whereas the EN model showed less alignment between the two lines. Notably, when the Chl-a concentration was high, disparities between the observed and predicted values emerged between the GB and RF models. In these cases, the predicted values were significantly lower than the observed values.

Table 2.

Root mean squared error (RMSE) and determination coefficient (R^2) values of four machine learning models: elastic net (EN), decision tree (DT), random forest (RF), and gradient boosting (GB)

	EN	DT	RF	GB
RMSE	0.724	0.298	0.192	0.187
R^2	0.338	0.888	0.954	0.956

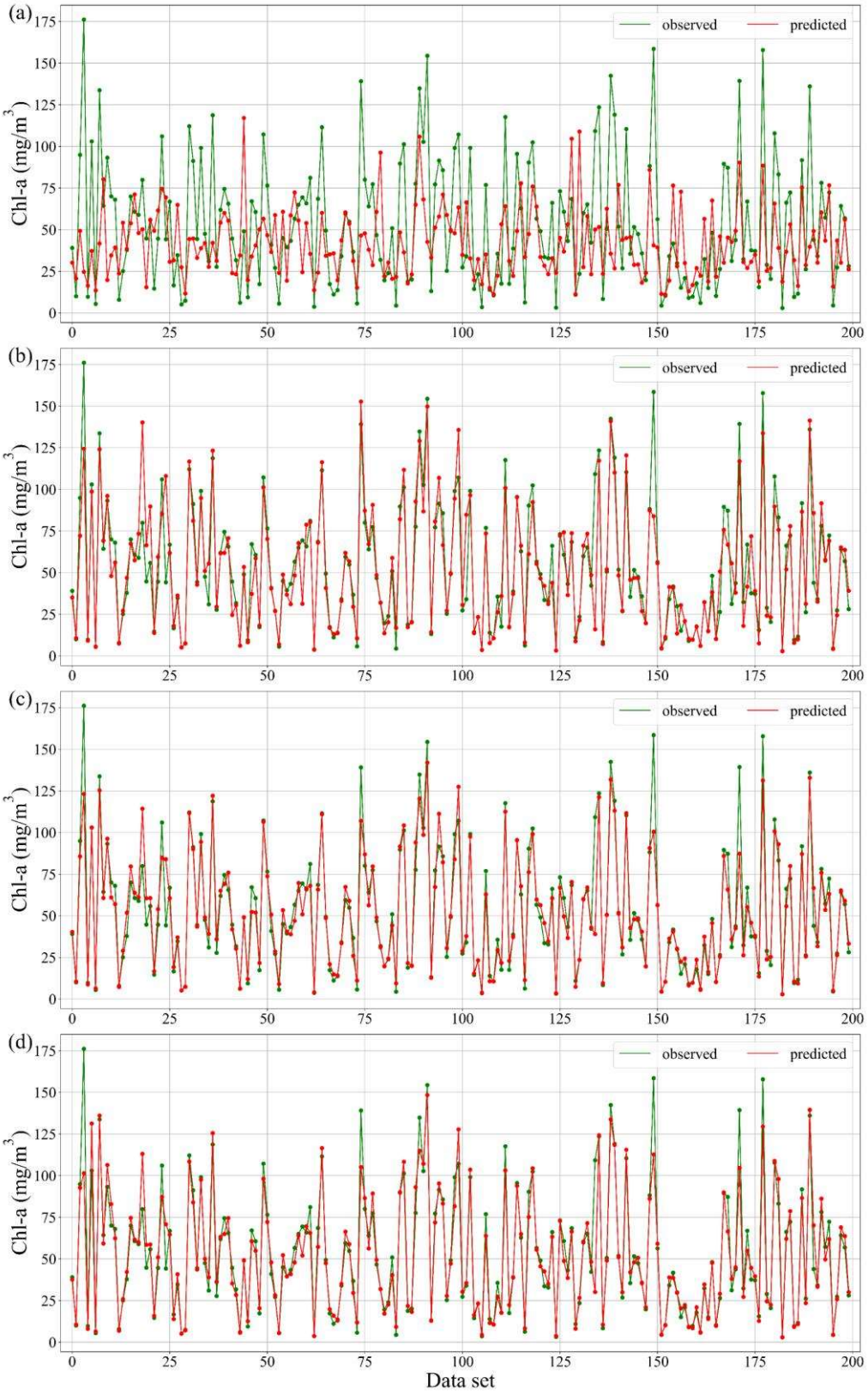


Fig. 3. Line graphs of observed (green line) and predicted (red line) values using the (a) elastic net (EN), (b), decision tree (DT), (c) random forest (RF), and (d) gradient boosting (GB)

models for 200 randomly selected data samples from the test set.

3.3. Feature importance analysis

The investigation into the factors influencing the Chl-a concentration involved assessing the FIV using the GB, RF, and DT models. The EN model was excluded because of its inadequate prediction accuracy. This exploration revealed consistent patterns in the order and FIV across the three models (Fig. 4). D-Flow was the most influential parameter affecting Chl-a concentration, followed by temperature. EC and WL contributed substantially, both with FIV exceeding 0.1.

The optimal conditions for algal growth include low flow rates, elevated temperatures in the presence of sunlight, and nutrient availability. Rivers with shorter water residence times and higher flow rates exhibit lower susceptibility to algal blooms than lakes and reservoirs because the reduced water residence time curbs nutrient uptake by algal cells and hampers algal metabolism (Kim et al., 2021b; Li et al., 2021; Xu et al., 2017; Zamparas and Zacharias, 2014). Water level and flow velocity are intricately connected, with rising water levels in the mainstream leading to a deceleration in flow velocity within the backwater sections of tributaries (Long et al., 2011). The potential for algal blooms may escalate owing to the extended duration of warm water temperatures. Among algae, cyanobacteria exhibit a competitive edge in higher temperature conditions (exceeding 25 °C). Elevated temperatures in surface waters lead to enhanced vertical stratification of the water column, which in turn fosters the development of cyanobacterial blooms (Coffey et al., 2019).

TN, TOC, and TP were found to have a relatively small influence on algal development than the other parameters. The impact of N and P on Chl-a levels has been a topic of controversy in various studies, as they were identified as primary factors significantly correlated with Chl-a in reservoirs in some studies (Gamez et al. 2019; Gupta et al., 2023;

Jargal et al., 2023), but their influence on Chl-a in 15 shallow lakes in China was inconsistent (Lv et al., 2011). The 25th percentile concentrations of TN and TP were 3.817 mg/L and 0.074 mg/L, respectively, exceeding the threshold levels for Microcystis-dominated blooms (TN: 0.80 mg/L, TP: 0.05 mg/L) (Xu et al., 2015). The relatively stable water quality throughout the year, attributed to 70–80% of the water entering the Yeongsan River originating from an upstream sewage treatment plant, may contribute to the diminished importance of nitrogen and phosphorus in this study. Consequently, the role of nutrients in fostering algal growth was less important than that of D-Flow and Temp in this study.

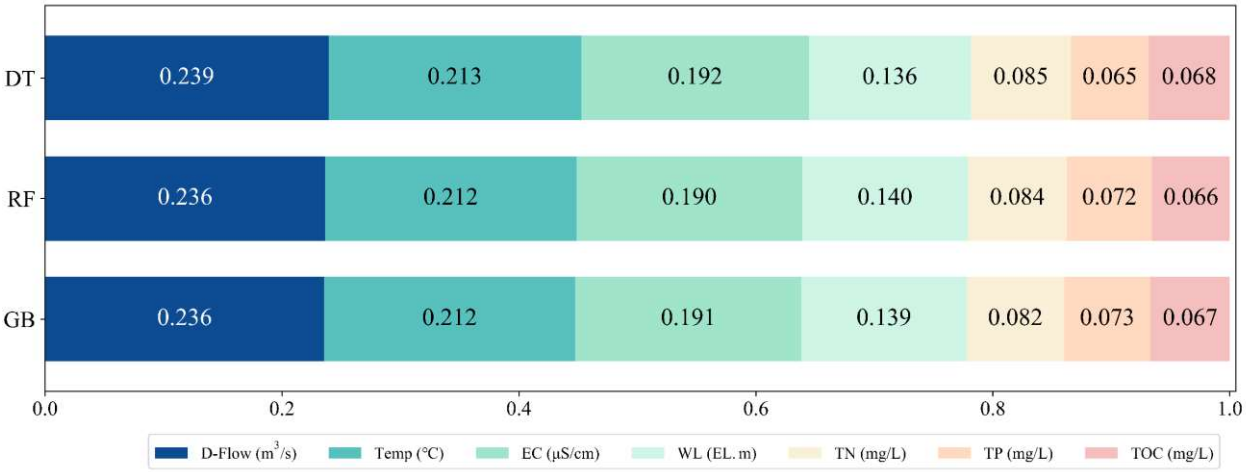


Fig. 4. Feature importance in the prediction of chlorophyll-a (Chl-a) concentrations obtained from the decision tree (DT), random forest (RF), and gradient boosting (GB) models.

3.4. Impact of D-Flow on Chl-a

3.4.1 Feature importance depending on the time span of averaged past D-Flow and Temp values

To delve deeper into the impact and quantify the effects of D-Flow and Temp, recognized as the pivotal factors influencing the Chl-a concentration (Section 3.3), a comprehensive study was conducted employing the GB model, which showed superior accuracy compared to the

other models (Section 3.2). Fourteen GB models were trained, each incorporating the daily mean values of D-Flow and Temp of the previous 1 to 14 days, respectively, as new variables. This analysis focused on assessing how historical D-Flow and Temp averages impacted the current algal concentration, rather than their concurrent influence on algal development. In Fig. 5a, it can be observed that as the historical averaging time span increases, there is a noticeable upward trend in the R^2 value and negated RMSE of the trained model, indicating an enhanced overall performance of the model. Negated RMSE is used instead of the RMSE to ensure that both line graphs exhibit a consistent upward trend. On the other hand, Fig. 5b summarizes the changes in feature importance depending on the time span and highlights the interplay between the averages of the D-Flow and Temp. Notably, the 3-day mean D-Flow shows significant importance, with D-Flow from days 1–5 retaining profound importance. The hydraulic retention time of 3.93 days, corresponding to the average D-Flow (26.508 m^3/sec, as indicated in Table 1), falls within these ranges. Temp was consistently one of the two most influential factors throughout the simulations. This analysis provides valuable insights into the hierarchy of variables that exert the greatest influence on Chl-a concentrations, highlighting the enduring importance of D-Flow and the consistent prominence of Temp.

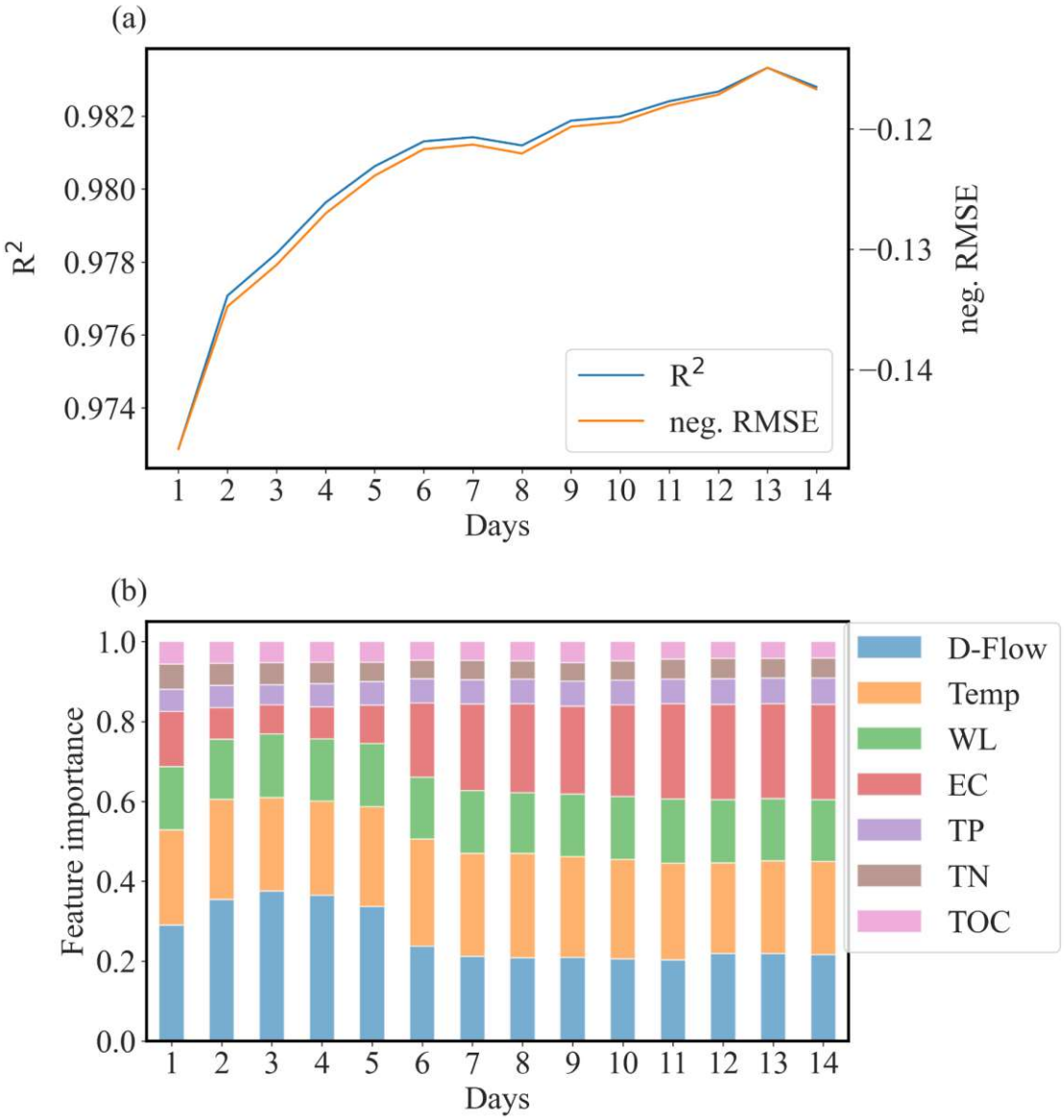


Fig. 5. Gradient boosting (GB) model performance with the time span of past discharge flow rate (D-Flow) and water temperature (Temp): (a) the values of determination coefficient (R^2) and negated root mean squared error (neg. RMSE) and (b) feature importance with averaging of past D-Flow and Temp values, ranging from 1 to 14 days.

3.4.2. Chl-a concentration dependent on the distribution of D-Flow and Temp

The distributions of the D-Flow and Temp were investigated. Figs. 6 and S4 show the respective distributions of D-Flow and Temp under different Chl-a concentrations (Figs 6a–d

and S4a–d for ≤ 20, > 20, > 70, and > 100 mg/m^3, respectively) for two weeks. A Chl-a concentration of 20 mg/m^3 corresponds to level 3 of Korea's Lake Environmental Water Quality Standards, representing a normal level. A concentration of 70 mg/m^3 indicates level 5, indicating poor water quality. At 100 mg/m^3, the water enters the algal bloom stage, which was the most critical phase of Korea's algal warning system before 2015. During this stage, proactive measures have to be implemented, including enhanced water purification, toxicity analyses, public awareness campaigns, algal removal, and discharge adjustments.

As expected, an increase in Chl-a concentration was associated with a reduction in D-Flow values. Significantly, when the algal concentration remained at 20 mg/m^3 or less (indicating good water quality conditions) for more than two weeks, the D-Flow distributions mostly ranged from 12–28 m^3/sec, corresponding to a hydraulic retention time of 3.72 to 8.68 days. In the Chl-a range above 20 mg/m^3, the D-Flow exhibited a broad range, from 6–22 m^3/sec. When Chl-a exceeded 70 mg/m^3, the D-Flow demonstrated a more focused distribution ranging from 6–16 m^3/sec (equivalent to 6.51–17.36 days of hydraulic retention time). Despite a limited dataset for Chl-a concentrations above 100 mg/m^3, D-Flow showed an intensive distribution within the 6–12 m^3/sec range.

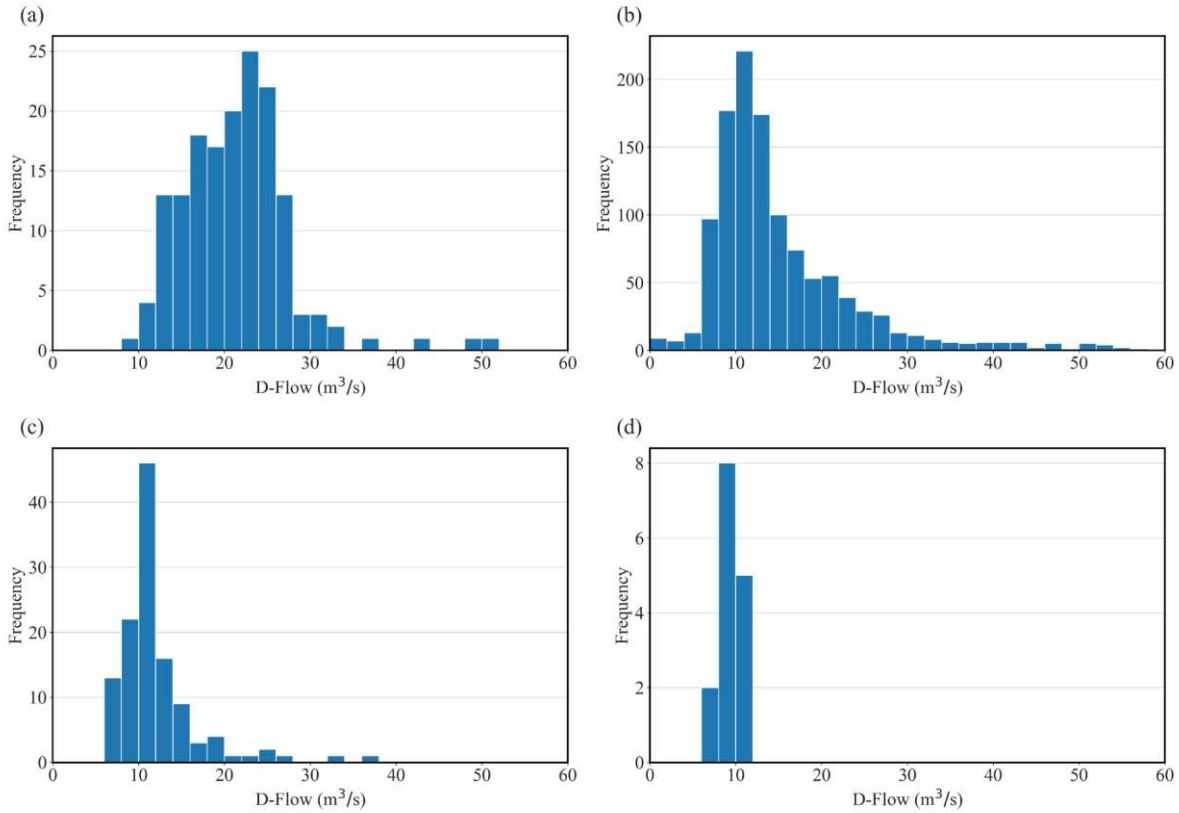


Fig. 6. Distribution of discharge flow rate (D-Flow) under different chlorophyll-a (Chl-a) concentrations for 14 days: (a) ≤ 20 mg/m^3, (b) > 20 mg/m^3, (c) > 70 mg/m^3, and (d) > 100 mg/m^3.

Figure S4 illustrates the Temp distribution corresponding to varying Chl-a concentrations over 14 days. The distribution of Temp when the Chl-a concentrations surpassed 70 mg/m^3 over the 14 days was relatively consistent across seasons. However, there was a notable increase in frequencies at Temp below 10 °C. Instances where Chl-a exceeded 100 mg/m^3 for 14 days were relatively rare, comprising only 15 samples. Nevertheless, the temperature distribution in these cases was more extreme than in the cases above 70 mg/m^3. Remarkably, Temp consistently ranged from 6–10 °C, indicating that occurrences of Chl-a exceeding 100 mg/m^3 primarily transpired during winter. This outcome contradicts the conventional understanding that warmer temperatures foster algal growth more effectively than colder

conditions (Hage et al., 2018; Silva et al., 2021). This result likely stems from the distinctive management practices of the Seungchon Weir and the winter climate of Korea. During winter, the Seungchon Weir experiences minimal inflow, including rainfall. The floodgate of the Seungchon Weir is closed in the winter to store water, which, combined with the dry winter season, contributes to reduced D-Flow.

Figure 7 illustrates the associations between the D-Flow/Temp and Chl-a concentrations over 14 days. As shown in Figure S4, the positive association between D-Flow and Temp, attributed to the management practices of the Seungchon Weir, is also evident in Figure 7. Lower D-Flow and Temp were consistently associated with higher Chl-a concentrations. As mentioned earlier, the observed D-Flow outcomes align with the expectations and findings of existing studies. However, the Temp results contradicted those in the prevailing literature (Chong et al., 2015; Li et al., 2021; Long et al., 2011; Wehr and Descy, 1998), which is due to the low D-Flow during winter. Therefore, it can be deduced that D-Flow had a greater influence on determining the Chl-a concentration than Temp in this study.

Flushing is a proposed algal removal technique that involves a temporary surge in discharge from a weir (Chong et al., 2015; Wehr and Descy, 1998), and its efficacy has been substantiated. Flow control strategies, including flushing, have been explored in the Barwon-Darling River system in Australia, and shown to mitigate algal growth (Mitrovic et al., 2006). Similarly, in the United States, an in-lake mesocosm experiment conducted in a small bay during a harmful algal outbreak demonstrated successful algal reduction through flushing (Hayden et al., 2012).

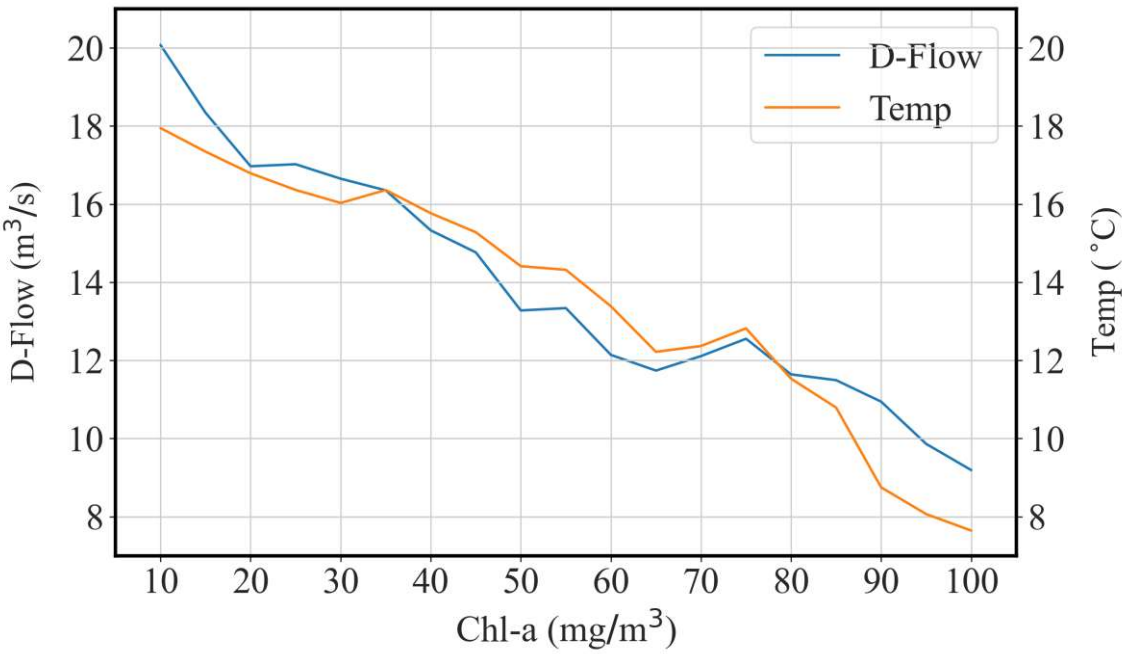


Fig. 7. Discharge flow rate (D-Flow) and water temperature (Temp) under different chlorophyll-a (Chl-a) concentrations for two weeks

3.4.3. Concentration of Chl-a according to the correlation between Temp and D-Flow

D-Flow was a pivotal determinant of the Chl-a concentration in the studied weir-constructed region (Section 3.4.2). Subsequent investigations were conducted on the Chl-a concentration to determine the minimum D-Flow necessary to constrain the Chl-a concentration below critical levels. Alterations based on D-Flow and Chl-a concentrations were assessed across seasons (spring, March–May; summer, June–August; autumn, September–November; and winter, December–February), as shown in Fig. 8.

Considering a Chl-a concentration of 20 mg/m^3 as the benchmark for good water quality, the daily mean flow rates required to achieve a concentration below this level were analyzed. In spring, summer, autumn, and winter, the daily mean flow rates necessary to reduce Chl-a concentrations to below 20 mg/m^3 were approximately 95, 85, 40, and 20 m^3/sec, respectively.

When averaged over 2 days, spring, summer, autumn, and winter required flow rates of approximately 30, 60, 35, and 15 m^3/sec, respectively. This implies that although elevated algal concentrations were observed due to limited D-Flow, Chl-a levels could be maintained within normal ranges despite modest D-Flow. Conversely, the elevated temperatures in summer required greater D-Flow to maintain optimal Chl-a concentrations.

Figure 8 also shows the sensitivity of Chl-a concentration to specific D-Flow ranges. In winter, Chl-a concentrations sharply declined with increasing D-Flow from 11.5 m^3/sec to 16.5 m^3/sec. In spring and fall, similar sharp decreases occurred with D-Flow increases from 11.5 m^3/sec to 26 m^3/sec. Winter exhibited a broader D-Flow range impacting Chl-a concentrations, and a distinct decrease was observed with D-Flow increasing from 11.5 m^3/sec to 50 m^3/sec. Each season exhibited specific D-Flow ranges where Chl-a concentration was particularly sensitive. Therefore, effectively reducing Chl-a concentration can be achieved by increasing D-Flow within the corresponding range for each season.

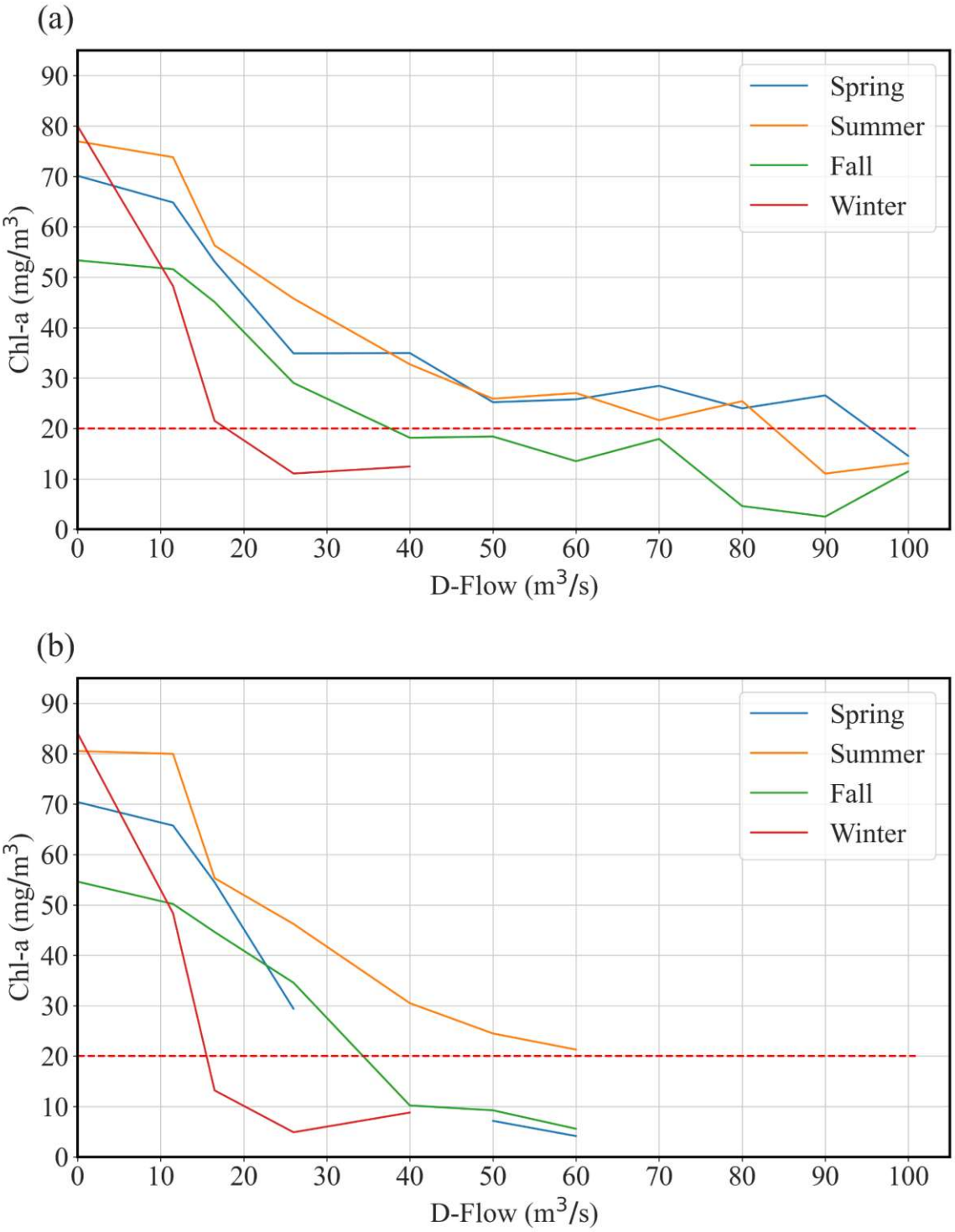


Fig. 8. Chlorophyll-a (Chl-a) concentration in relation to (a) 1-day and (b) 2-day average discharge flow rate (D-Flow) across the four seasons.

To examine the effect of Temp on the Chl-a concentration in more detail, the Chl-a concentration at changing D-Flow was examined by dividing the water temperature into increments of 5 °C. Figure 9 illustrates the relationship between Chl-a concentrations and daily average D-Flow (Fig. 9a for the 1-day and Fig. 9b for the 2-day average discharge flow rate). In Fig. 9a, it can be observed that as the temperature increased, the required D-Flow to achieve a Chl-a concentration of 20 mg/m^3 tended to increase, albeit with notable fluctuations at 15 and 20 °C. In Fig. 9b, a significant increase in the necessary D-Flow was evident at higher temperatures. Specifically, when Temp ranged from 0 to 10 °C, the D-Flow of 15 m^3/sec was necessary to attain the target Chl-a concentration of 20 mg/m^3. In contrast, at temperatures between 25 and 30 °C, the required D-Flow was much higher, at 60 m^3/sec. This finding highlights the substantial impact of D-Flow and Temp on the Chl-a concentration. The elevated Temp demand increased D-Flow to achieve the desired Chl-a level of 20 mg/m^3, signifying good water quality.

In Fig. 9, the sensitivity of Chl-a concentration to specific D-Flow ranges was observed, and these D-Flow ranges were dependent on water temperature. The sharp decrease in Chl-a concentration within specific D-Flow ranges resembled the results illustrated in Fig. 8. At 0–10 °C, a notable drop in Chl-a concentration was observed with the increase of D-Flow from 0 to 16.5 m^3/sec. Between 10–15 °C, the D-Flow ranges leading to the drop in Chl-a concentration extended to 0–26 m^3/sec. In the 25–30 °C range, the D-Flow ranges resulting in the drop of Chl-a concentration further extended to 0–40 m^3/sec. Above 30 °C, such a sharp drop in Chl-a concentration with the increase in D-Flow was not observed. In summary, the increase in water temperature led to the extension of D-Flow ranges affecting the sharp drop in Chl-a concentration, necessitating higher D-Flow for decreasing Chl-a concentration.

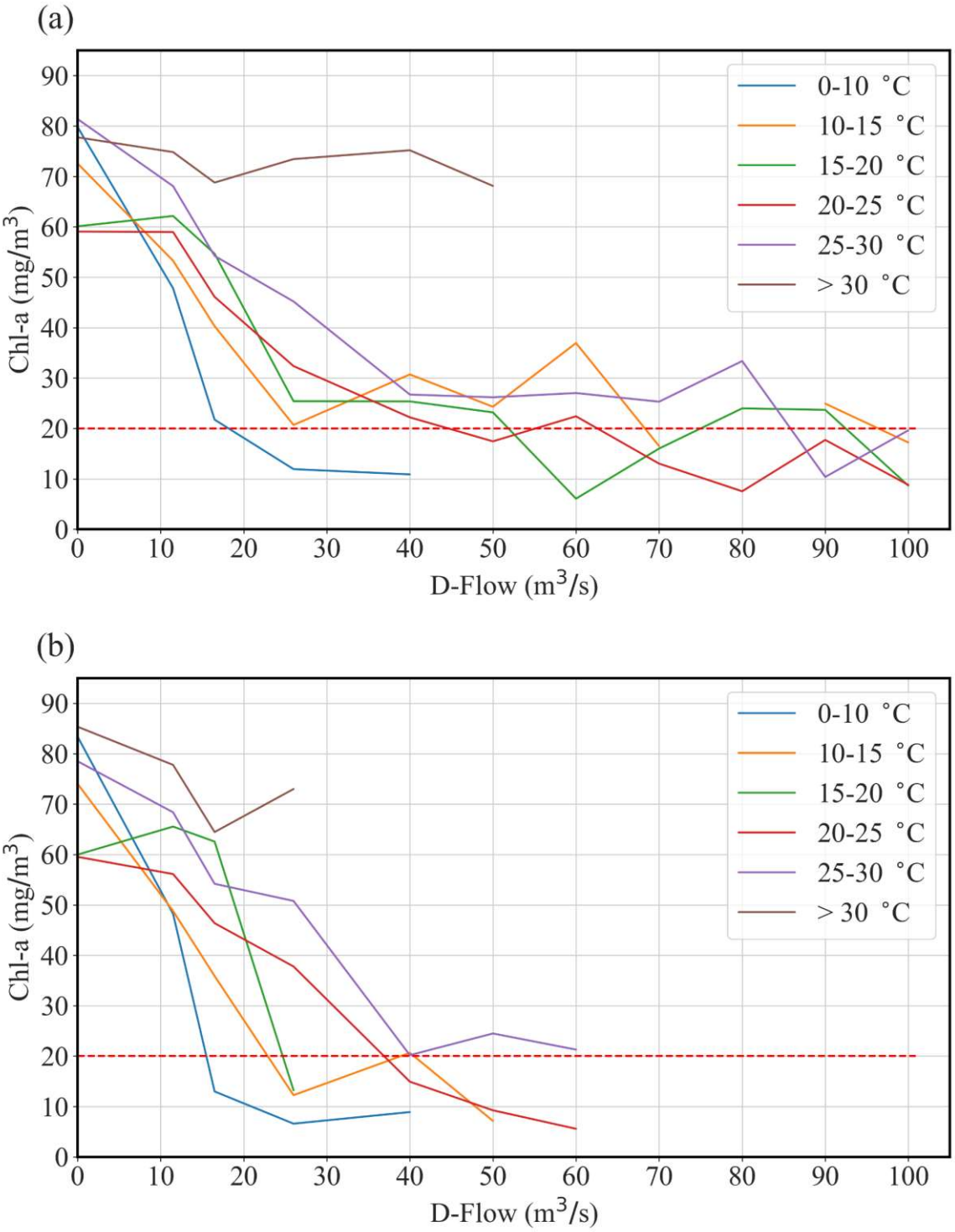


Fig. 9. Chlorophyll-a (Chl-a) concentration in relation to (a) 1-day and (b) 2-day average discharge flow rate (D-Flow) under varying water temperatures.

4. Conclusions

This study harnessed the potential of machine learning algorithms to forecast Chl-a concentrations while delving into the contributors to algal presence in a river ecosystem containing artificially constructed weirs. Four distinct machine learning models (EN, DT, RF, and GB) were deployed to predict Chl-a concentrations and identify the water quality and hydraulic factors influencing these levels. Among the model variants, GB exhibited the highest R^2 and the lowest RMSE values, indicating its superior suitability. The discharge flow rate and water temperature emerged as pivotal factors determining Chl-a concentration. To enhance the model accuracy, we tested the impact of past D-Flow and Temp on Chl-a concentration. Notably, the importance of D-Flow escalated as the averaging timeframe increased. The investigation involved plotting the Chl-a concentration over two weeks alongside the D-Flow and temperature data. Interestingly, higher Chl-a concentrations were observed at lower levels of both D-Flow and temperature, contrary to previous studies that suggested a direct increase in Chl-a with rising temperatures. In the Seungchon Weir, D-Flow emerged as the primary factor influencing Chl-a concentration, outweighing the impact of temperature. While the effect of temperature on Chl-a was relatively small, the required D-Flow for maintaining optimal Chl-a levels (20 mg/m^3) displayed variability in response to temperature fluctuations. Specifically, as the temperature increased, the D-Flow necessary to sustain a favorable Chl-a concentration also increased. The concentration of Chl-a demonstrated distinct sensitivity to varying D-Flow ranges within each season and temperature category. Therefore, the targeted reduction of Chl-a levels can be effectively achieved by strategically adjusting D-Flow within the designated range corresponding to each specific season and temperature condition. This study identifies D-Flow as the primary driver of algal development in artificially constructed river systems, and maintaining an appropriate D-Flow level is essential for sustaining ideal Chl-a concentrations. This study introduced a novel approach for identifying the causes of river algal

blooms and proposed a weir management strategy, along with a minimum D-Flow value, to prevent algal outbreaks. However, additional validation in diverse water bodies is necessary to confirm our hypothesis.

Funding: This research did not receive any specific grants from funding agencies in the public, commercial, or not-for-profit sectors.

Authors' contributions: Hyunju Kim: Writing—original draft, data analysis, and visualization; **Gyesik Lee:** Data analysis, writing—original draft, writing—review and editing; **Chang-Gu Lee:** Writing—review and editing; **Seong-Jik Park:** Conceptualization, writing—original draft, writing—review and editing, and supervision. All authors have read and agreed to the published version of the manuscript.

Data availability: All data generated or analyzed during this study are included in this published article. The datasets used and/or analyzed in the current study are available from the corresponding author upon reasonable request.

References

Ahmed, A.N., Othman, F.B., Afan, H.A., Ibrahim, R.K., Fai, C.M., Hossain, M.S., Ehteram, M., Elshafie, A., 2019. Machine learning methods for better water quality prediction. Journal of Hydrology 578, 124084. https://doi.org/10.1016/j.jhydrol.2019.124084.

Aires, U.R.V., Silva, D.D., Filho, E.I.F., Rodrigues, L.N., Uliana, E.M., Amorim, R.S.S., Ribeiro, C.B.M., Campos, J.A., 2022. Modeling of surface sediment concentration in the Doce River basin using satellite remote sensing. Journal of Environmental Management. 323(1), 116207. https://doi.org/10.1016/j.jenvman.2022.116207.

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 1984. Classification and regression

trees. Wadsworth International Group, Belmont CA.

Breiman, L., 2001. Random forests. Machine Learning 45, 5-32. https://doi.org/10.1023/A:1010933404324.

Çamdevýren, H., Demýr, N., Kanik, A., Keskýn, S., 2005. Use of principal component scores in multiple linear regression models for prediction of Chlorophyll-a in reservoirs. Ecological Modelling 181(4), 581-589. https://doi.org/10.1016/j.ecolmodel.2004.06.043.

Chong, S., Yi, H.S., Hwang, H.S., Kim, H.J., 2015. Modeling the flushing effect of multi-purpose weir operation on algae removal in Yeongsan River. Journal of the Korean Society of Environmental Engineering 37(10), 563-572. https://doi.org/10.4491/KSEE.2015.37.10.563.

Coffey, R., Paul, M.J., Stamp, J., Hamilton, A., Johnson, T., 2019. A review of water quality responses to air temperature and precipitation changes 2: nutrients, algal blooms, sediment, pathogens. JAWRA. 55(4), 844-868. https://doi.org/10.1111/1752-1688.12711.

Deng, T., Chau, K.W., Duan, H.F., 2021. Machine learning based marine water quality prediction for coastal hydro-environment management. Journal of Environmental Management 284, 112051. https://doi.org/10.1016/j.jenvman.2021.112051.

Friedman, J.H., 2001. Greedy function approximation: A gradient boosting machine. Annals of Statistics 29(5), 1189-1232. https://doi.org/10.1214/aos/1013203451.

Friedman, J.H. 2002. Stochastic gradient boosting. CSDA. 38(4), 367-378. https://doi.org/10.1016/S0167-9473(01)00065-2.

Friedman, J.H., Hastie, T., Tibshirani, R., 2010. Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software 33(1), 1-22. https://doi.org/10.18637/jss.v033.i01.

Furnas, M.J., 1990. In situ growth rates of marine phytoplankton: approaches to measurement,

community and species growth rates. Journal of Plankton Research 12(6), 1117-1151. https://doi.org/10.1093/plankt/12.6.1117.

Gamez, T.E., Benton, L., Manning, S.R., 2019. Observations of two reservoirs during a drought in Central Texas, USA: strategies for detecting harmful algal blooms. Ecological Indicators 104, 588–593. https://doi.org/10.1016/j.ecolind.2019.05.022.

Gupta, A., Hantush, M.M., Govindaraju, R.S., 2023. Sub-monthly time scale forecasting of harmful algal blooms intensity in Lake Erie using remote sensing and machine learning. Science of the Total Environment, 900, 165781. https://doi.org/10.1016/j.scitotenv.2023.165781.

Hage, A., Luckett, N., Holbrook, G.P., 2018. Phycoremediation of Municipal Wastewater by the Cold- Adapted Microalga Monoraphidium sp. Dek19. Water Environmental Research 90(11), 1938-1946. https://doi.org/10.2175/106143017X15131012188060.

Hayden, N.J., Roelke, D.L., Brooks, B.W., Grover, J.P., Neisch, M.T., Valenti Jr, T.W., Prosser, K.N., Gable, G.M., Umphres, G.D., Hewitt, N.C., 2012. Beyond hydraulic flushing: Deep water mixing takes the harm out of a haptophyte algal bloom. Harmful Algae, 20, 42-57. https://doi.org/10.1016/j.hal.2012.07.006.

Hong, S.H., Ndingwan, A.M., Yoo, S.C., Lee, C.G., Park, S.J., 2020. Use of calcined sepiolite in removing phosphate from water and returning phosphate to soil as phosphorus fertilizer. J. Environ. Manage. 270, 110817. https://doi.org/10.1016/j.jenvman.2020.110817.

Hong, S.M., Abbas, A., Kim, S., Kwon, D.H., Yoon, N., Yun, D., Lee, S., Pachepsky, Y., Pyo, J.C., Cho, K.H., 2023. Autonomous calibration of EFDC for predicting chlorophyll-a using reinforcement learning and a real-time monitoring system. Environmental Modelling & Software, 168, 105805. https://doi.org/10.1016/j.envsoft.2023.105805.

Jargal, N., Lee, E.H., An, K.G., 2023. Monsoon-induced response of algal chlorophyll to trophic state, light availability, and morphometry in 293 temperate reservoirs. Journal of Environmental Management 337, 117737. https://doi.org/10.1016/j.jenvman.2023.117737.

Kang, J.K., Seo, E.J., Lee, C.G., Jeong, S., Park, S.J., 2022. Application of response surface methodology and artificial neural network for the preparation of Fe-loaded biochar for enhanced Cr (VI) adsorption and its physicochemical properties and Cr (VI) adsorption characteristics. Environmental Science and Pollution Research 29(40), 60852-60866. https://doi.org/10.1007/s11356-022-20009-3

Keller, S., Maier, P.M., Riese, F.M., Norra, S., Holbach, A., Börsig, N., Wilhelms, A., Moldaenke, C., Zaake A., Hinz, S., 2018. Hyperspectral data and machine learning for estimating CDOM, chlorophyll a, diatoms, green algae and turbidity. International Journal of Environmental Research and Public Health 15(9), 1881. https://doi.org/10.3390/ijerph15091881.

Kim, J., Jones, J.R., Seo, D., 2021a. Factors affecting harmful algal bloom occurrence in a river with regulated hydrology. Journal of Hydrology: Regional Studies. 33, 100769. https://doi.org/10.1016/j.ejrh.2020.100769.

Kim, J.H., Shin, J.K., Lee, H., Lee, D.H., Kang, J.H., Cho, K.H., Lee, Y.G., Chon, K., Park, Y., 2021b. Improving the performance of machine learning models for early warning of harmful algal blooms using an adaptive synthetic sampling method. Water Research 207, 117821. https://doi.org/10.1016/j.watres.2021.117821.

Kwak, J., 2021. A study on the 3-month prior prediction of Chl-a concentration in the Daechong Lake using hydrometeorological forecasting data. Journal of Wetlands Research 23(2), 144-153. https://doi.org/10.17663/JWR.2021.23.2.144.

Lee, H.J., Kim, H.J., Choi, K.S., 2017. Investigation and monitoring of causes of algal blooms

in the four major rivers. Water for Future. 50(6), 20-25.

Lee, Y.J., Im, E.C., Lee, G., Hong, S.C., Lee, C.G., Park, S.J. 2024. Comparison of ammonia volatilization in paddy and field soils fertilized with urea and ammonium sulfate during rice, potato, and Chinese cabbage cultivation. Atmospheric Pollution Research, 15(4), 102049. https://doi.org/10.1016/j.apr.2024.102049.

Li, J., Yin, W., Jia, H., Xin, X., 2021. Hydrological management strategies for the control of algal blooms in regulated lowland rivers. Hydrological Processes 35(6), e14171. https://doi.org/10.1002/hyp.14171.

Lian, A., Han, D., Song, X., Yang, S., 2021. Impacts of storm events on Chlorophyll-a variations and controlling factors for algal bloom in a river receiving reclaimed water. Journal of Environmental Management 297(1), 113376. https://doi.org/10.1016/j.jenvman.2021.113376.

Liu, Y., Wang, Y., Zhang, J., 2012. New machine learning algorithm: Random forest. In:Liu, B., Ma, M., Chang, J.(eds) Information Computing and Applications. ICICA 2012. Lecture Notes in Computer Science. Springer, Berlin, Heidelberg.

Liu, N., Yang, Y., Li, F., Ge, F., Kuang, Y., 2016. Importance of controlling pH-depended dissolved inorganic carbon to prevent algal bloom outbreaks. Bioresources And Technology 220, 246-252. https://doi.org/10.1016/j.biortech.2016.08.059.

Long, T.Y., Wu, L., Meng, G.H., Guo, W.H., 2011. Numerical simulation for impacts of hydrodynamic conditions on algae growth in Chongqing Section of Jialing River, China. Ecological Modelling 222(1), 112-119.

Lu, H., Ma, X., 2020. Hybrid decision tree-based machine learning models for short-term water quality prediction. Chemosphere. 249, 126169. https://doi.org/10.1016/j.chemosphere.2020.126169.

Lv, J., Wu, H., Chen, M., 2011. Effects of nitrogen and phosphorus on phytoplankton

composition and biomass in 15 subtropical, urban shallow lakes in Wuhan, China. Limnologica 41, 48–56. https://doi.org/10.1016/j.limno.2010.03.003.

Ly, Q.V., Tong, N.A., Lee, B.M., Nguyen, M.H., Trung, H.T., Le Nguyen, P., Hoang, T.H., Hwang, Y., Hur, J., 2023. Improving algal bloom detection using spectroscopic analysis and machine learning: A case study in a large artificial reservoir, South Korea. Science of The Total Environment. 901, 166467. https://doi.org/10.1016/j.scitotenv.2023.166467

Makhotin, I., Koroteev, D., Burnaev, E., 2019. Gradient boosting to boost the efficiency of hydraulic fracturing. Journal of Petroleum Exploration and Production Technology 9, 1919-1925. https://doi.org/10.1007/s13202-019-0636-7.

Mitrovic, S.M., Chessman, B.C., Bowling, L.C., Cooke, R.H., 2006. Modelling suppression of cyanobacterial blooms by flow management in a lowland river. River Research Applications. 22(1), 109-114. https://doi.org/10.1002/rra.875.

Mng'ong'o, M., Munishi, L.K., Blake, W., Comber, S., Hutchinson, T.H., Ndakidemi, P.A., 2022. Towards sustainability: Threat of water quality degradation and eutrophication in Usangu agro-ecosystem Tanzania. Marine Pollution Bulletin, 181, 113909. https://doi.org/10.1016/j.marpolbul.2022.113909.

Ogutu, J.O., Schulz-Streeck, T., Piepho, H.P., 2012. Genomic selection using regularized linear regression models: ridge regression, lasso, elastic net and their extensions. BMC proceedings. 6, S10. https://doi.org/10.1186/1753-6561-6-S2-S10.

Park, Y., Cho, K.H., Park, J., Cha, S.M., Kim, J.H., 2015. Development of early-warning protocol for predicting chlorophyll-a concentration using machine learning models in freshwater and estuarine reservoirs, Korea. Science of the Total Environment 502, 31-41. https://doi.org/10.1016/j.scitotenv.2014.09.005.

Park, Y., Lee, H.K., Shin, J.K., Chon, K., Kim, S., Cho, K.H., Kim, J.H., Baek, S.S., 2021. A machine learning approach for early warning of cyanobacterial bloom outbreaks in a

freshwater reservoir. Journal of Environmental Management 288, 112415. https://doi.org/10.1016/j.jenvman.2021.112415.

Raven, J.A., Gobler, C.J., Hansen, P.J., 2020. Dynamic CO2 and pH levels in coastal, estuarine, and inland waters: Theoretical and observed effects on harmful algal blooms. Harmful Algae, 91, 101594. https://doi.org/10.1016/j.hal.2019.03.012.

Shin, C.M., Min, J.H., Park, S.Y., Choi, J., Park, J.H., Song, Y.S., Kim, K., 2017. Operational water quality forecast for the Yeongsan River using EFDC model. Journal of the Korean Society of Water Environment 33(2), 219-229. https://doi.org/10.15681/KSWE.2017.33.2.219.

Silva, L., Calleja, M.L., Ivetic, S., Huete-Stauffer, T., Roth, F., Carvalho, S., Morán, X.A.G., 2021. Heterotrophic bacterioplankton responses in coral-and algae-dominated Red Sea reefs show they might benefit from future regime shift. Science of the Total Environment 751, 141628. https://doi.org/10.1016/j.scitotenv.2020.141628.

Suggett, D.J., Prášil, O., Borowitzka, M.A., 2010. Chlorophyll a fluorescence in aquatic sciences: methods and applications, Springer.

Water Resources Management Information System, 2018. Water quality standards. http://www.wamis.go.kr/wke/wke_wqbase_lst.do (accessed 26 August 2023)

Wehr, J.D., Descy, J.P., 1998. Use of phytoplankton in large river management, Journal of Phycology 34(5), 741–749. https://doi.org/10.1046/j.1529-8817.1998.340741.x.

Xu, H., Paerl, H.W., Qin, B., Zhu, G., Hall, N.S., Wu, Y., 2015. Determining critical nutrient thresholds needed to control harmful cyanobacterial blooms in eutrophic Lake Taihu, China. Environmental Science and Technology 49(2), 1051-1059. https://doi.org/10.1021/es503744q

Xu, H., Paerl, H.W., Zhu, G., Qin, B., Hall, N.S., Zhu, M., 2017. Long-term nutrient trends and

harmful cyanobacterial bloom potential in hypertrophic Lake Taihu, China. Hydrobiologia, 787, 229-242. https://doi.org/10.1007/s10750-016-2967-4.

Yang, Q., Liu, G., Hao, Y., Zhang, L., Giannetti, B.F., Wang, J., Casazza, M., 2019. Donor-side evaluation of coastal and marine ecosystem services. Water Research 166, 115028. https://doi.org/10.1016/j.watres.2019.115028.

Yaqub, M., Ngoc, N.M., Park, S., Lee, W., 2022. Predictive modeling of pharmaceutical product removal by a managed aquifer recharge system: Comparison and optimization of models using ensemble learners. Journal of Environmental Management. 324(15), 116345. https://doi.org/10.1016/j.jenvman.2022.116345.

Yu, P., Gao, R., Zhang, D., Liu, Z.P., 2021. Predicting coastal algal blooms with environmental factors by machine learning methods. Ecological Indicators 123, 107334. https://doi.org/10.1016/j.ecolind.2020.107334.

Zamparas, M., Zacharias, I., 2014. Restoration of eutrophic freshwater by managing internal nutrient loads. A review. Science of the Total Environment 496, 551-562. https://doi.org/10.1016/j.scitotenv.2014.07.076.

Zhou, S., Shao, Y., Gao, N., Deng, Y., Li, L., Deng, J., Tan, C., 2014. Characterization of algal organic matters of Microcystis aeruginosa: biodegradability, DBP formation and membrane fouling potential. Water Research 52, 199-207. https://doi.org/10.1016/j.watres.2014.01.002.

Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. Journal of the royal statistical society series B: Statistical Methodology 67(2), 301-320. https://doi.org/10.1111/j.1467-9868.2005.00503.x.

Algae Development in Rivers with Artificially Constructed Weirs: Dominant Influence of Discharge over Temperature

Hyunju Kim [a], Gyesik Lee [b,*], Chang-Gu Lee [c], Seong-Jik Park [d,*]

[a] Faculty of Liberal Education, Seoul National University, Seoul, 08826, Republic of Korea

[b] School of Computer Engineering and Applied Mathematics, Hankyong National University, Anseong, 17579, Republic of Korea

[c] Department of Environmental and Safety Engineering, Ajou University, Suwon 16499, Republic of Korea

[d] Department of Bioresources and Rural System Engineering, Hankyong National University, Anseong, 17579, Republic of Korea

* ***Corresponding author***: S. J. Park

E-mail address: parkseongjik@hknu.ac.kr

ORCID: 0000-0003-2122-5498

G. Lee

E-mail address: gslee@hknu.ac.kr

Abstract

Algal blooms contribute to water quality degradation, unpleasant odors, taste issues, and the presence of harmful substances in artificially constructed weirs. Mitigating these adverse effects through effective algal bloom management requires identifying the contributing factors and predicting algal concentrations. This study focused on the upstream region of the Seungchon Weir in Korea, which is characterized by elevated levels of total nitrogen and phosphorus due to a significant influx of water from a sewage treatment plant. We employed four distinct machine learning models to predict chlorophyll-a (Chl-a) concentrations and identified the influential variables linked to local algal bloom events. The gradient boosting model enabled an in-depth exploration of the intricate relationships between algal occurrence and water quality parameters, enabling accurate identification of the causal factors. The models identified the discharge flow rate (D-Flow) and water temperature as the primary determinants of Chl-a levels, with feature importance values of 0.236 and 0.212, respectively. Enhanced model precision was achieved by utilizing daily average D-Flow values, with model accuracy and significance of the D-Flow amplifying as the temporal span of daily averaging increased. Elevated Chl-a concentrations correlated with diminished D-Flow and temperature, highlighting the pivotal role of D-Flow in regulating Chl-a concentration. This trend can be attributed to the constrained discharge of the Seungchon Weir during winter. Calculating the requisite D-Flow to maintain a desirable Chl-a concentration of up to 20 mg/m^3 across varying temperatures revealed an escalating demand for D-Flow with rising temperatures. Specific D-Flow ranges, corresponding to each season and temperature condition, were identified as particularly influential on Chl-a concentration. Thus, optimizing Chl-a reduction can be achieved by strategically increasing D-Flow within these specified ranges for each season and temperature variation. This study highlights the importance of maintaining sufficient D-Flow levels to mitigate algal proliferation within river systems featuring weirs.

Keywords: algae bloom; machine learning; temperature; discharge flow rate; constructed weirs; chlorophyll-a

1. Introduction

Algae are single-celled or multicellular organisms that provide food and oxygen to aquatic animals and play vital roles in aquatic ecosystems. However, excessive algal growth, known as algal blooms, can lead to severe environmental problems such as fish mortality, contamination of drinking water, destruction of aquatic habitats, and human diseases caused by algal toxins (Zhou et al., 2014; Hong et al., 2020; Kim et al., 2021a). Algal blooms are a global phenomenon and have caused environmental and social problems in many countries. For example, in Korea, the construction of weirs in four major rivers significantly increased the frequency and severity of algal blooms, mainly due to the reduced flow rate caused by the increased water depth and storage capacity (Lee et al., 2017). Thus, predicting and controlling algal blooms has become a critical issue, and effective control measures are needed.

Chlorophyll is a group of pigments that play a critical role in photosynthesis in all phototrophic organisms, including algae and some bacterial species (Suggett et al., 2010). Monitoring chlorophyll concentrations can provide valuable insights into phytoplankton biomass and nutritional status (Furnas, 1990), making it an indirect indicator of nutrient levels, such as phosphorus and nitrogen, in surface water (Keller et al., 2018). As a widely used water quality parameter, the concentration of chlorophyll-a (Chl-a) is commonly utilized to assess algal levels (Kwak, 2021; Shin et al., 2017). In Korea, the concentrations of Chl-a in rivers, together with cyanobacterial cell densities, are used as criteria for the algal warning system (Park et al., 2015).

Monitoring and predicting surface water quality is crucial for mitigating the potential

damage caused by harmful algae and improving water quality. Traditional methods for predicting algal blooms using conventional water quality indicators are costly, labor-intensive, and time-consuming, posing significant challenges in providing timely monitoring and management interventions (Ly et al., 2023). Model development can help predict algal growth and evolution, enabling proactive measures to be taken (Deng et al., 2021; Ly et al., 2023). Identifying the predictive potential of environmental factors on algal bloom occurrence can facilitate the development of effective management strategies (Yu et al., 2021).

Process-based models have been used since the 1980s to predict algal blooms and elucidate the relationships between algal growth and environmental factors (Deng et al., 2021; Yang et al., 2019). Although these models can provide relatively accurate predictions of water quality parameters, they face challenges when dealing with large quantities of input data (Ahmed et al., 2019). Statistical methods, such as principal component analysis with multiple linear regression, have previously been used to predict Chl-a concentrations too (Çamdevýren et al., 2005).

Recently, machine learning-based prediction has become popular in many scientific fields, especially for the environmental modeling of complex nonlinear phenomena (Kang et al., 2022; Lee et al., 2024; Yu et al., 2021). Machine learning models map the relationships between the inputs and outputs of a system rather than completing complex process mechanisms, enabling accurate predictions of highly nonlinear relationships without prior knowledge of the system (Deng et al., 2021). For example, CEEMDAN-RF and CEEMDAN-XGBoost, two hybrid decision tree (DT)-based models, have been applied to predict water quality based on parameters such as temperature, dissolved oxygen, pH, specific conductance, turbidity, and fluorescent dissolved organic matter (Lu and Ma, 2020). Park et al. (2021) used artificial neural networks and support vector machines to predict algal bloom alert levels by incorporating environmental variables, such as nutrients and meteorological factors, in freshwater reservoirs.

Mozo et al. (2022) developed machine learning models for algal bloom predictions using three years of Chl-a data but did not identify the primary factors controlling algal blooms, as their analysis only focused on a single parameter. Ly et al. (2023) compared five different machine learning models for predicting algal growth, revealing that XGBoost and stacking methods demonstrated superior performance over other models. Additionally, Hong et al. (2023) employed reinforcement learning to autonomously calibrate sequential water quality parameters in the Environmental Fluid Dynamics Code model for Chl-a prediction.

This study aimed to address the significant challenges arising from algal blooms in artificially constructed weirs. The investigation involved identifying the contributing factors and predicting algal concentrations using four different machine learning models: elastic net (EN), decision tree (DT), random forest (RF), and gradient boosting (GB). These models are widely employed in environmental studies. EN is a representative regularized linear regression model that combines the advantages of both ridge and lasso regression and performed well in predicting surface sediment concentrations (Aires et al. 2022). DT, RF, and GB are tree-based regression models commonly used for predicting unknown variables. Yaqub et al. (2022) used DT, RF, and extreme GB models to remove pharmaceutical products from water in a managed aquifer recharge system, with XGBoost showing notably better results. In contrast, Kim et al. (2022) found RF to be suitable for predicting Chl-a levels, and Lian et al. (2021) indicated that RF can effectively predict variations in algal blooms.

This study not only compared model performance in predicting Chl-a concentrations but also analyzed the factors influencing Chl-a concentration in a river with an artificially constructed weir. Among the key determinants influencing Chl-a concentration, the discharge flow rate (D-Flow) stands out. We considered utilizing different temporal average values of D-Flow in our analyses, and we found that extended temporal averaging bolstered model accuracy

and the significance of D-Flow. Additionally, a D-Flow analysis was conducted to determine the necessary discharge for maintaining the desired Chl-a concentration across varying temperatures and seasons, achieved by segmenting the D-Flow distribution according to the Chl-a concentration range. This study offers fundamental insights for practical weir operation, elucidating the factors that impact algal blooms in river systems with weir structures, and proposing optimal D-Flow to sustain favorable Chl-a concentrations.

2. Materials and Methods

2.1. Study area and data acquisition

The Seungchon Weir in the Yeongsan River was constructed as part of the Four Major Rivers Project in Korea, which started in October 2009 and was completed in May 2012. The concentration of Chl-a in the Yeongsan River increased significantly after the construction of the weir compared to the other three major rivers (Lee et al., 2017), and has the most severe water pollution among the four major rivers in the project. Reports indicate eutrophication due to high phosphorus and nitrogen concentrations, low oxygen depletion, and high phytoplankton concentrations downstream throughout the year (Son et al., 2013). The Yeongsan River, characterized by a short length and a small basin area, consistently experiences insufficient water flow. Approximately 70% of the mainstream's flow is derived from discharges from sewage treatment plants. Furthermore, the Yeongsan River basin exhibits a low percentage of forested areas and a high proportion of agricultural land, contributing to the deterioration of water quality due to non-point pollution sources. The Seungchon Weir is 512 m long and 12 m wide, with a management elevation level of 7.5 m upstream, providing 9 million m^3 of water (Fig. 1). The weir has a multifunctional structure consisting of four truss-type movable liftgate weirs (total length 180 m, elevation above mean sea level [ELm] 2.5 m) and three fixed weirs (total length 304.5 m, ELm 7.5 m). It includes two fishways, one on the left side of the weir

and the other along the original river path. Additionally, an operational hydroelectric power plant with a maximum power generation capacity of 800 kW and a water usage rate of 28 m^3/s has been installed and is operational (Chong et al., 2015). Water quality data for this study were collected at a latitude of 35° 4' 14" and longitude of 126° 46' 35", 1.1 km upstream of the Seungchon Weir.

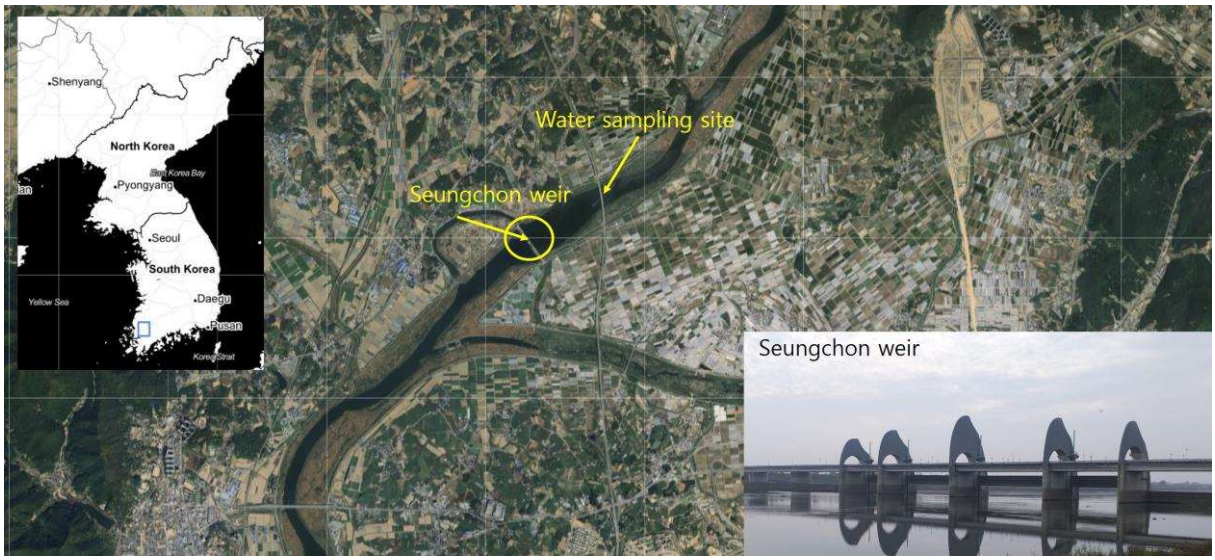


Fig. 1. Map and digital image of the study area encompassing the Seungchon Weir and the water sampling site.

This study utilized hourly water quality data obtained from a real-time water quality information system (https://water.nier.go.kr) at the National Institute of Environmental Sciences of the Korean Ministry of Environment (KME). A total of 82,257 observations from January 1, 2013, to May 23, 2022, were analyzed, with parameters including water temperature (Temp), hydrogen ion concentration (pH), electrical conductivity (EC), dissolved oxygen (DO), total organic carbon (TOC), total nitrogen (TN), total phosphorus (TP), and Chl-a concentrations. Data on the D-Flow and water level (WL) at the upstream location were obtained from the Han River Flood Control Office (https://www.hrfco.go.kr) and converted

into 1 h intervals to correspond to the hourly water quality data.

2.2. Data sets

Six of the eight water quality parameters used in this study (listed above had missing values: Temp (21.5%), EC (21.6%), TOC (28.7%), TN (27.1%), TP (28.3%), and Chl-a (23.6%). To fill in the missing values, we used linear interpolation with a 12-hour window in each direction. If data were missing for more than 24 hours, they were excluded from the input. In addition, the upstream WL was maintained at ELm 7.5 m, and data with a WL greater than 10 m were considered outliers and excluded.

2.3. Machine learning models

The models used to predict the Chl-a concentration were EN, DT, RF, and GB.

2.3.1. EN

ENs are linear regression models that combine L1- and L2-regularization to regularize model parameters called weights, using the L1-norm $\| \cdot \|_1$ and L2-norm $\| \cdot \|_2$, respectively (Zou and Hastie, 2005). During training, the EN model attempts to learn a weight matrix W that minimizes the following formula (Eq. 1) (Friedman et al., 2010):

$$(\|X\,W - y\|_2^2 + r\,\alpha\,\|W\|_1 + \frac{1}{2}(1 - r)\,\alpha\,\|W\|_2^2) \quad (1)$$

where X is the matrix representing the training set, y is the target vector, r is the ratio of L1-regularization, and α the intensity of regularization. The L1-regularization performs variable selection by eliminating the weights of the least important variables, while the L2-regularization keeps the weights as small as possible to minimize the model's variance (Ogutu et al., 2012).

2.3.2. DT

DTs are nonlinear models with a binary tree structure. Each node in a DT holds information from a dataset, and the process of splitting a node corresponds to the partitioning of the dataset into two subsets.

Node t of the binary tree is split into two child nodes t^L and t^R to maximize the reduction in misclassification cost (Breiman et al., 1984). An algorithm called the classification and regression tree (CART) performs this task, as demonstrated in Fig. 2a.

In node t, for instance, the CART algorithm chooses a variable j and a value c, and decides for each data x whether x should belong to node t^L or t^R depending on the result of the comparison $x_j \leq c$, where x_j is the value contained in x for the variable j. The performance of DTs depends on the extent to which the partitioning is executed. Moreover, there is always a risk of overfitting.

2.3.3. RF

RFs combine a series of DTs to prevent model overfitting and improve prediction accuracy (Liu et al., 2012). As shown in Fig. 2b, an RF model simultaneously trains a multitude of DT using a subset of the training set selected by sampling with replacement for each tree. Its final prediction is made by aggregating the predictions of all of the individual trees (Breiman, 2001). The classification tasks were decided by the majority. For regression tasks, the average of individual trees was used.

2.3.4. GB

GB is a method for gradually improving prediction performance by sequentially training weak estimators (Makhotin et al., 2019). Friedman (2001, 2002) proposed a GB algorithm that

starts with a DT as the base estimator and cumulatively trains a new model that predicts the residual error of the previous model. Fig. 2c illustrates the proposed algorithm.

Adding the weak estimator h_k trained with the (k-1)-th residual error results in a stronger estimator (Eq. 2):

$$F_k(x) = F_{k-1}(x) + \rho_k \, h_k(x) \quad (2)$$

where F_0 is the initial weak estimator, and ρ_k is the k-th learning rate.

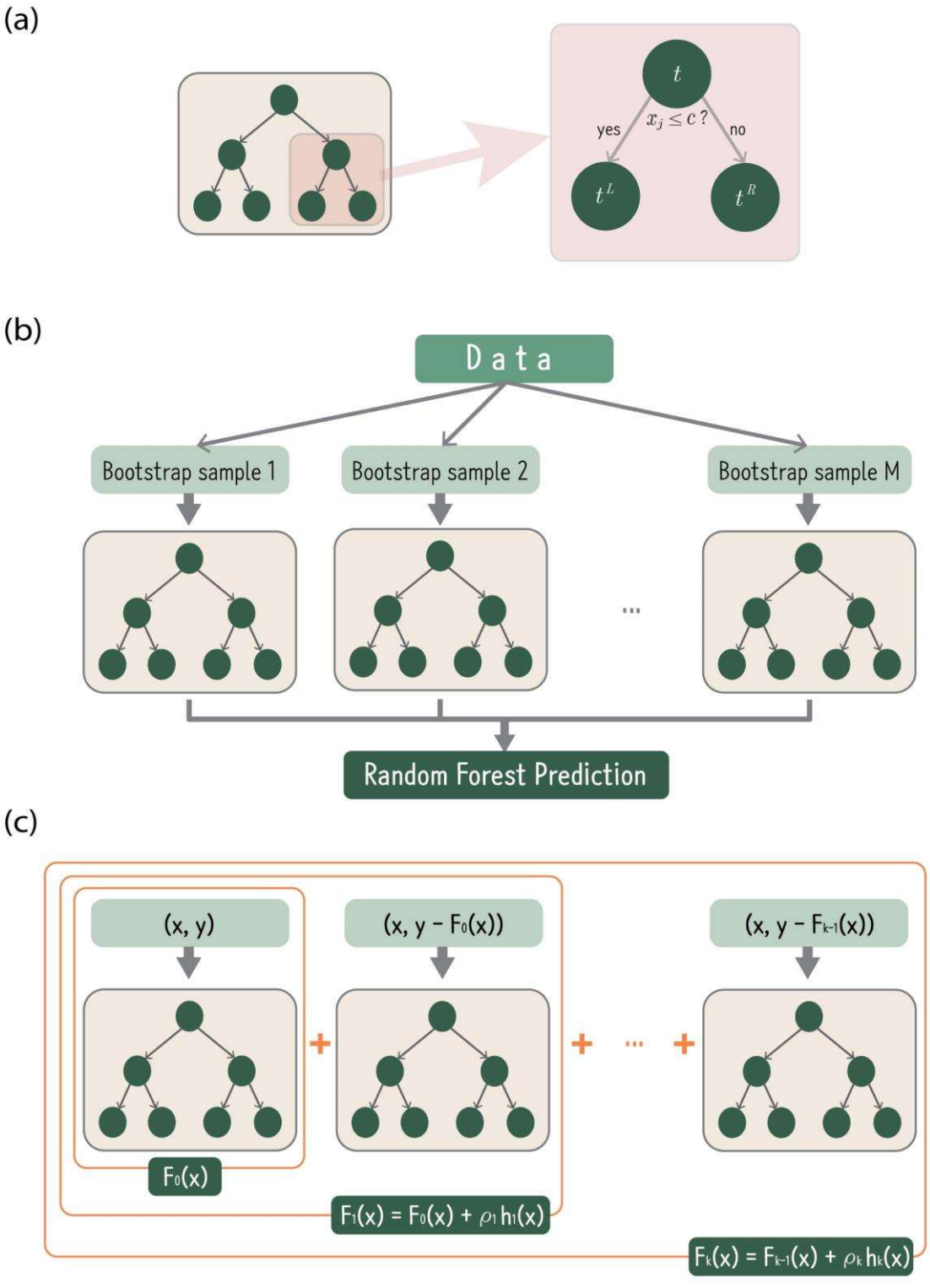


Fig. 2. Schematic diagram of (a) decision tree, (b) random forest, and (c) gradient boosting

models.

2.4. Model performance evaluation

The performance of each model was assessed using two metrics: the root mean squared error (RMSE) and coefficient of determination (R^2). RMSE (Eq. 3) is a widely used regression evaluation metric that measures the average error between the estimated value $\widehat{y_k}$ and true value y_k over all N data samples.

$$RMSE = \sqrt{\frac{1}{N}\sum_{k=1}^{N}(\widehat{y_k} - y_k)^2} \tag{3}$$

R^2 (Eq. 4) is displayed as a value less than or equal to 1 and is used to evaluate the fitting performance of the model. The closer R^2 is to 1, the better the fitting performance of the model

$$R^2 = 1 - \frac{\sum_{k=1}^{N}(\widehat{y_k} - y)^2}{\sum_{k=1}^{N}\left(y_k - \underline{y}\right)^2} \tag{4}$$

where $\underline{y}$ is the average of the original data and $\widehat{y_k}$ and y_k are the predicted and true values, respectively.

2.5. Application of models

The EN, DT, RF, and GB models were implemented using Python's Scikit-learn library (https://scikit-learn.org/stable/), a Python library that provides almost all the tools needed for data analysis and training machine learning models. The dataset was divided into two sets for performance evaluation after training: a training set containing 75% of the dataset and a test set comprising the remaining 25%. However, the regression model performed better when the response variable had a normal distribution. Fig. S1a shows that the distribution of Chl-a was skewed to the left; therefore, we made it approach a normal distribution by log transformation

(Fig. S1b). Hyperparameter tuning for each model was performed through a grid search to optimize model performance. The sklearn library's GridSearchCV was employed to systematically explore various hyperparameter combinations and identify the best-performing set. The optimal hyperparameters for the models utilized in this study are summarized in Table S1.

To assess the effectiveness of four distinct models and identify the feature importance of water quality parameters in predicting Chl-a, the input variables included pH, DO, Temp, EC, TOC, TN, TP, WL, and D-Flow, all obtained concurrently with Chl-a measurements. Additionally, an in-depth investigation was conducted using the GB model, which demonstrated superior accuracy. The model was enhanced by incorporating the average daily values of D-Flow and Temp from the preceding 1 to 14 days.

3. Results and Discussion

3.1. Statistical analysis of water quality and hydraulic parameters

Statistical analysis of the water quality and hydraulic conditions of the Seungchon Weir are provided in Table 1. The TOC concentration fell within a moderate range (level 3), according to the water quality standards of the KME. However, the mean concentrations of TN and TP corresponded to extremely poor water quality (> 1.5 mg/L, level 6) and poor water quality levels (0.10–0.15 mg/L, level 5), respectively, according to KME standards. An excess of phosphorus (> 0.1 mg/L) has been shown to have severely detrimental effects on freshwater systems, contributing to eutrophication (Mng'ong'o et al., 2022). The mean Chl-a concentration was high (35–70 mg/m^3, level 5). These findings indicate a nutrient-rich environment in the Seungchon Weir, indicating high eutrophication and favorable conditions for algal growth (Mng'ong'o et al., 2022).

Table 1.

Statistics of water quality and hydraulic parameters of the Seungchon Weir including water temperature (Temp), electrical conductivity (EC), total organic carbon (TOC), total nitrogen (TN), total phosphorus (TP), water level (WL), discharge flow rate (D-Flow), and chlorophyll-a (Chl-a).

	Temp (℃)	EC (μS/cm)	TOC (mg/L)	TN (mg/L)	TP (mg/L)	WL (ELm)	D-Flow (m^3/s)	Chl-a (mg/ m^3)
Mean	17.962	339.594	4.364	5.397	0.119	6.847	26.508	50.334
SD	8.028	88.196	1.026	2.001	0.064	0.864	62.494	37.231
Min	1.800	92.000	1.600	0.706	0.003	4.038	0.000	1.000
25%	10.000	277.000	3.600	3.817	0.074	5.980	11.400	20.813
50%	18.200	339.000	4.264	5.157	0.106	7.490	15.237	42.500
75%	25.100	411.000	5.000	6.726	0.150	7.530	25.962	72.200
Max	36.100	730.000	19.300	12.357	0.745	8.072	2710.73	453.500

Pearson's correlation analysis was performed using a dataset comprising 49,111 samples, which included all water quality and hydraulic parameters out of a total of 82,257 observations. The correlation analysis is presented as a heatmap plot (Fig. S2). Temperature was significantly and inversely correlated with EC and TN. TN and EC showed strong positive correlations. DO and pH were highly correlated with Chl-a, indicating that both these parameters influenced algal outbreaks. This correlation can be attributed to algae producing oxygen through

photosynthesis and hydroxyl ions by utilizing inorganic carbon (bicarbonate) in the water (Liu et al., 2016).

3.2. Comparison of machine learning models

We employed four distinct machine learning models (EN, DT, RF, and GB) to forecast the Chl-a concentration using pH, DO, Temp, EC, TOC, TN, TP, WL, and D-Flow as input variables (Fig. S3). In the DT, RF, and GB models, pH emerged as the most influential variable, with a feature importance value (FIV) of approximately 0.3. DO ranked fifth in importance across all three models, with a FIV of approximately 0.08. Notably, photosynthesis by algae can lead to elevated DO and pH values (Raven et al., 2020). Consequently, as pH and DO are outcomes of algal occurrence rather than factors influencing its growth, these two variables were excluded from further analysis. The GB and RF models exhibited remarkable performances based on their high R^2 values and low RMSE values, whereas the performance of EN was too low to be useful (Table 2). Since the EN is a linear regression model that assumes a linear relationship between the independent and dependent variables (Zou and Hastie, 2005), it may not be suitable for capturing the complex interactions between Chl-a and water quality parameters. Figure 3 shows the Chl-a concentration observed and predicted by each of the four models for a randomly selected subset of 200 data samples from the test set. Notably, for the GB and RF models, the observed and predicted line plots coincided approximately, whereas the EN model showed less alignment between the two lines. Notably, when the Chl-a concentration was high, disparities between the observed and predicted values emerged between the GB and RF models. In these cases, the predicted values were significantly lower than the observed values.

Table 2.

Root mean squared error (RMSE) and determination coefficient (R^2) values of four machine learning models: elastic net (EN), decision tree (DT), random forest (RF), and gradient boosting (GB)

	EN	DT	RF	GB
RMSE	0.724	0.298	0.192	0.187
R^2	0.338	0.888	0.954	0.956

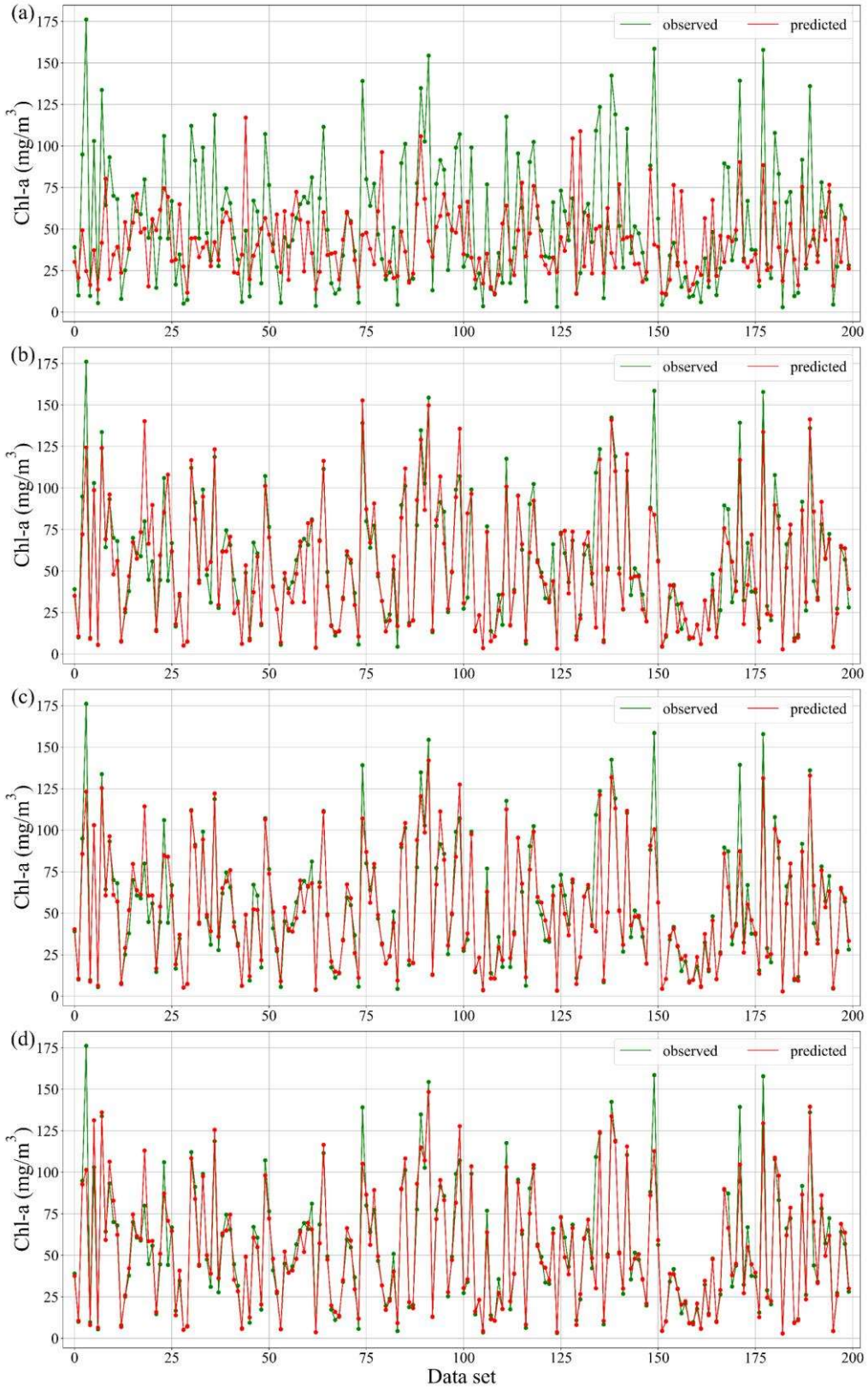


Fig. 3. Line graphs of observed (green line) and predicted (red line) values using the (a) elastic net (EN), (b), decision tree (DT), (c) random forest (RF), and (d) gradient boosting (GB)

models for 200 randomly selected data samples from the test set.

3.3. Feature importance analysis

The investigation into the factors influencing the Chl-a concentration involved assessing the FIV using the GB, RF, and DT models. The EN model was excluded because of its inadequate prediction accuracy. This exploration revealed consistent patterns in the order and FIV across the three models (Fig. 4). D-Flow was the most influential parameter affecting Chl-a concentration, followed by temperature. EC and WL contributed substantially, both with FIV exceeding 0.1.

The optimal conditions for algal growth include low flow rates, elevated temperatures in the presence of sunlight, and nutrient availability. Rivers with shorter water residence times and higher flow rates exhibit lower susceptibility to algal blooms than lakes and reservoirs because the reduced water residence time curbs nutrient uptake by algal cells and hampers algal metabolism (Kim et al., 2021b; Li et al., 2021; Xu et al., 2017; Zamparas and Zacharias, 2014). Water level and flow velocity are intricately connected, with rising water levels in the mainstream leading to a deceleration in flow velocity within the backwater sections of tributaries (Long et al., 2011). The potential for algal blooms may escalate owing to the extended duration of warm water temperatures. Among algae, cyanobacteria exhibit a competitive edge in higher temperature conditions (exceeding 25 °C). Elevated temperatures in surface waters lead to enhanced vertical stratification of the water column, which in turn fosters the development of cyanobacterial blooms (Coffey et al., 2019).

TN, TOC, and TP were found to have a relatively small influence on algal development than the other parameters. The impact of N and P on Chl-a levels has been a topic of controversy in various studies, as they were identified as primary factors significantly correlated with Chl-a in reservoirs in some studies (Gamez et al. 2019; Gupta et al., 2023;

Jargal et al., 2023), but their influence on Chl-a in 15 shallow lakes in China was inconsistent (Lv et al., 2011). The 25th percentile concentrations of TN and TP were 3.817 mg/L and 0.074 mg/L, respectively, exceeding the threshold levels for Microcystis-dominated blooms (TN: 0.80 mg/L, TP: 0.05 mg/L) (Xu et al., 2015). The relatively stable water quality throughout the year, attributed to 70–80% of the water entering the Yeongsan River originating from an upstream sewage treatment plant, may contribute to the diminished importance of nitrogen and phosphorus in this study. Consequently, the role of nutrients in fostering algal growth was less important than that of D-Flow and Temp in this study.

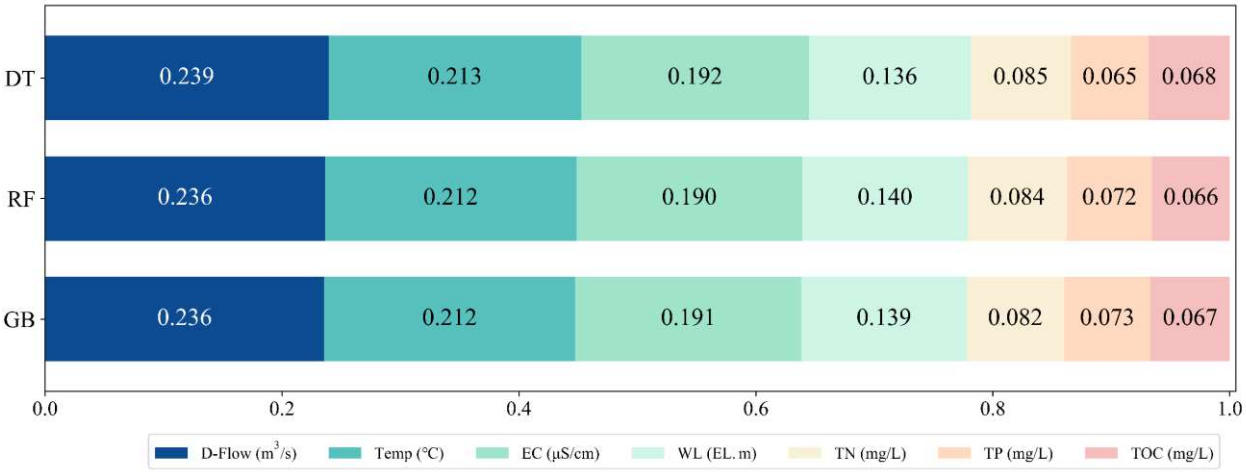


Fig. 4. Feature importance in the prediction of chlorophyll-a (Chl-a) concentrations obtained from the decision tree (DT), random forest (RF), and gradient boosting (GB) models.

3.4. Impact of D-Flow on Chl-a

3.4.1 Feature importance depending on the time span of averaged past D-Flow and Temp values

To delve deeper into the impact and quantify the effects of D-Flow and Temp, recognized as the pivotal factors influencing the Chl-a concentration (Section 3.3), a comprehensive study was conducted employing the GB model, which showed superior accuracy compared to the

other models (Section 3.2). Fourteen GB models were trained, each incorporating the daily mean values of D-Flow and Temp of the previous 1 to 14 days, respectively, as new variables. This analysis focused on assessing how historical D-Flow and Temp averages impacted the current algal concentration, rather than their concurrent influence on algal development. In Fig. 5a, it can be observed that as the historical averaging time span increases, there is a noticeable upward trend in the R^2 value and negated RMSE of the trained model, indicating an enhanced overall performance of the model. Negated RMSE is used instead of the RMSE to ensure that both line graphs exhibit a consistent upward trend. On the other hand, Fig. 5b summarizes the changes in feature importance depending on the time span and highlights the interplay between the averages of the D-Flow and Temp. Notably, the 3-day mean D-Flow shows significant importance, with D-Flow from days 1–5 retaining profound importance. The hydraulic retention time of 3.93 days, corresponding to the average D-Flow (26.508 m^3/sec, as indicated in Table 1), falls within these ranges. Temp was consistently one of the two most influential factors throughout the simulations. This analysis provides valuable insights into the hierarchy of variables that exert the greatest influence on Chl-a concentrations, highlighting the enduring importance of D-Flow and the consistent prominence of Temp.

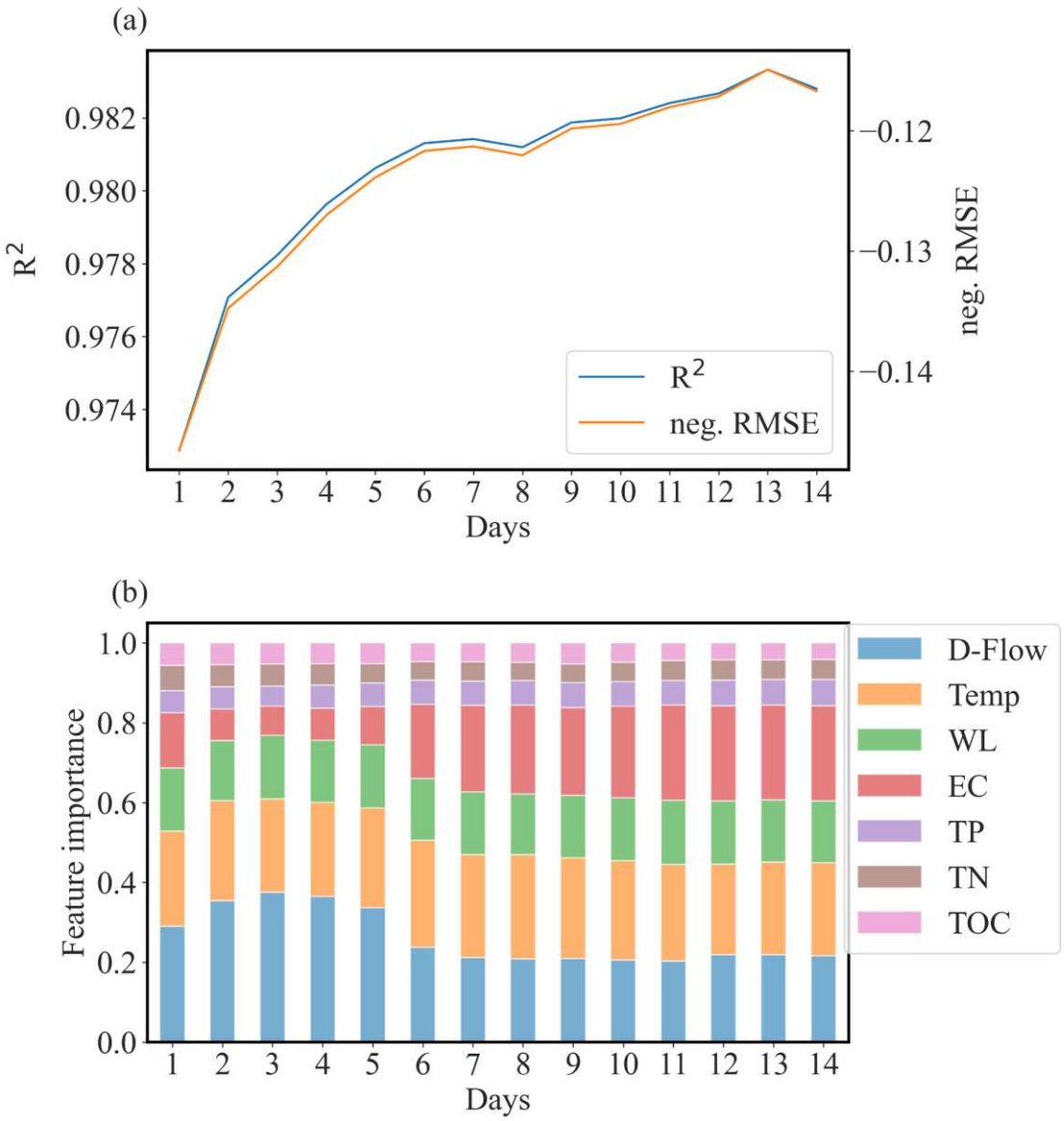


Fig. 5. Gradient boosting (GB) model performance with the time span of past discharge flow rate (D-Flow) and water temperature (Temp): (a) the values of determination coefficient (R^2) and negated root mean squared error (neg. RMSE) and (b) feature importance with averaging of past D-Flow and Temp values, ranging from 1 to 14 days.

3.4.2. Chl-a concentration dependent on the distribution of D-Flow and Temp

The distributions of the D-Flow and Temp were investigated. Figs. 6 and S4 show the respective distributions of D-Flow and Temp under different Chl-a concentrations (Figs 6a–d

and S4a–d for ≤ 20, > 20, > 70, and > 100 mg/m^3, respectively) for two weeks. A Chl-a concentration of 20 mg/m^3 corresponds to level 3 of Korea's Lake Environmental Water Quality Standards, representing a normal level. A concentration of 70 mg/m^3 indicates level 5, indicating poor water quality. At 100 mg/m^3, the water enters the algal bloom stage, which was the most critical phase of Korea's algal warning system before 2015. During this stage, proactive measures have to be implemented, including enhanced water purification, toxicity analyses, public awareness campaigns, algal removal, and discharge adjustments.

As expected, an increase in Chl-a concentration was associated with a reduction in D-Flow values. Significantly, when the algal concentration remained at 20 mg/m^3 or less (indicating good water quality conditions) for more than two weeks, the D-Flow distributions mostly ranged from 12–28 m^3/sec, corresponding to a hydraulic retention time of 3.72 to 8.68 days. In the Chl-a range above 20 mg/m^3, the D-Flow exhibited a broad range, from 6–22 m^3/sec. When Chl-a exceeded 70 mg/m^3, the D-Flow demonstrated a more focused distribution ranging from 6–16 m^3/sec (equivalent to 6.51–17.36 days of hydraulic retention time). Despite a limited dataset for Chl-a concentrations above 100 mg/m^3, D-Flow showed an intensive distribution within the 6–12 m^3/sec range.

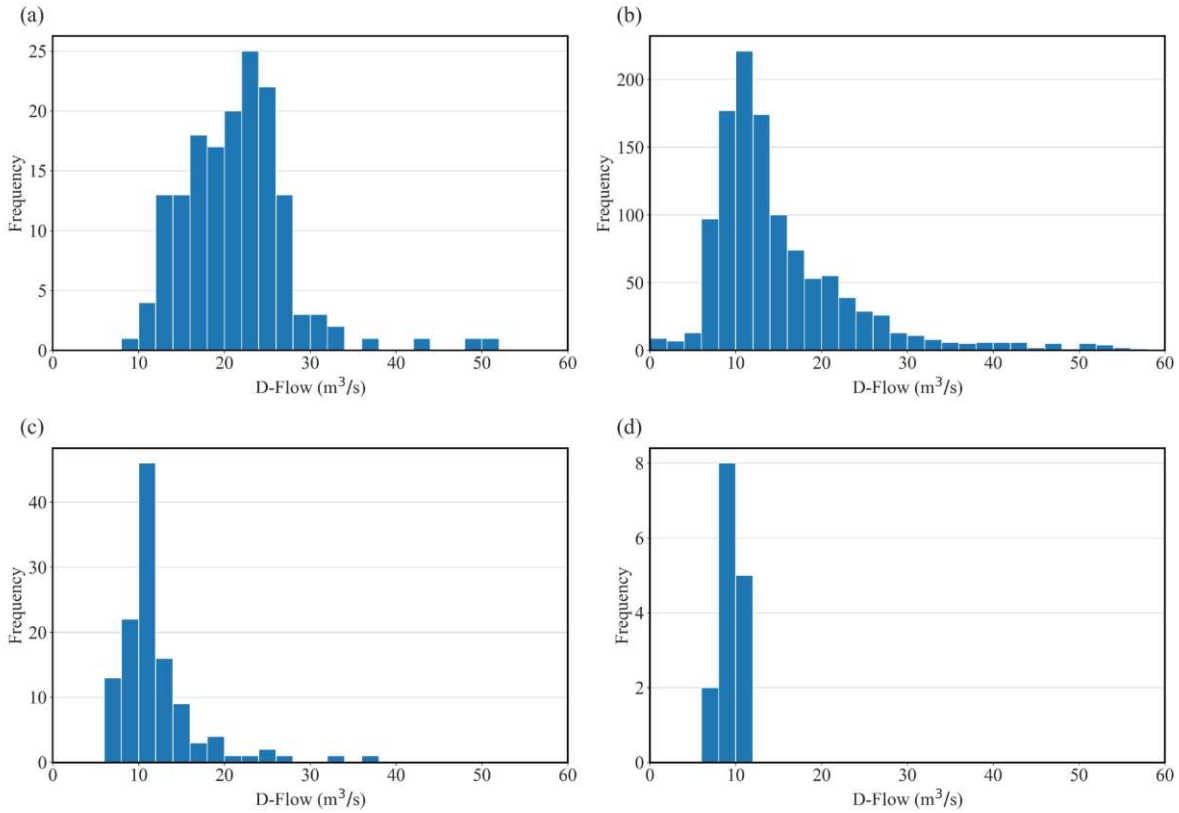


Fig. 6. Distribution of discharge flow rate (D-Flow) under different chlorophyll-a (Chl-a) concentrations for 14 days: (a) ≤ 20 mg/m^3, (b) > 20 mg/m^3, (c) > 70 mg/m^3, and (d) > 100 mg/m^3.

Figure S4 illustrates the Temp distribution corresponding to varying Chl-a concentrations over 14 days. The distribution of Temp when the Chl-a concentrations surpassed 70 mg/m^3 over the 14 days was relatively consistent across seasons. However, there was a notable increase in frequencies at Temp below 10 °C. Instances where Chl-a exceeded 100 mg/m^3 for 14 days were relatively rare, comprising only 15 samples. Nevertheless, the temperature distribution in these cases was more extreme than in the cases above 70 mg/m^3. Remarkably, Temp consistently ranged from 6–10 °C, indicating that occurrences of Chl-a exceeding 100 mg/m^3 primarily transpired during winter. This outcome contradicts the conventional understanding that warmer temperatures foster algal growth more effectively than colder

conditions (Hage et al., 2018; Silva et al., 2021). This result likely stems from the distinctive management practices of the Seungchon Weir and the winter climate of Korea. During winter, the Seungchon Weir experiences minimal inflow, including rainfall. The floodgate of the Seungchon Weir is closed in the winter to store water, which, combined with the dry winter season, contributes to reduced D-Flow.

Figure 7 illustrates the associations between the D-Flow/Temp and Chl-a concentrations over 14 days. As shown in Figure S4, the positive association between D-Flow and Temp, attributed to the management practices of the Seungchon Weir, is also evident in Figure 7. Lower D-Flow and Temp were consistently associated with higher Chl-a concentrations. As mentioned earlier, the observed D-Flow outcomes align with the expectations and findings of existing studies. However, the Temp results contradicted those in the prevailing literature (Chong et al., 2015; Li et al., 2021; Long et al., 2011; Wehr and Descy, 1998), which is due to the low D-Flow during winter. Therefore, it can be deduced that D-Flow had a greater influence on determining the Chl-a concentration than Temp in this study.

Flushing is a proposed algal removal technique that involves a temporary surge in discharge from a weir (Chong et al., 2015; Wehr and Descy, 1998), and its efficacy has been substantiated. Flow control strategies, including flushing, have been explored in the Barwon-Darling River system in Australia, and shown to mitigate algal growth (Mitrovic et al., 2006). Similarly, in the United States, an in-lake mesocosm experiment conducted in a small bay during a harmful algal outbreak demonstrated successful algal reduction through flushing (Hayden et al., 2012).

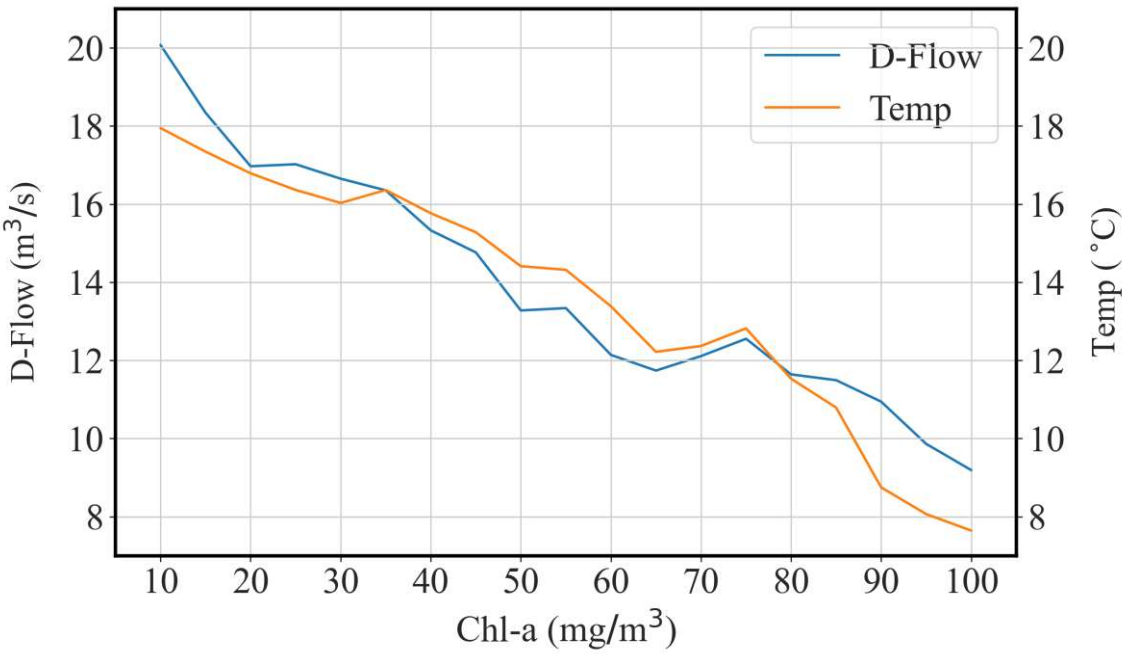


Fig. 7. Discharge flow rate (D-Flow) and water temperature (Temp) under different chlorophyll-a (Chl-a) concentrations for two weeks

3.4.3. Concentration of Chl-a according to the correlation between Temp and D-Flow

D-Flow was a pivotal determinant of the Chl-a concentration in the studied weir-constructed region (Section 3.4.2). Subsequent investigations were conducted on the Chl-a concentration to determine the minimum D-Flow necessary to constrain the Chl-a concentration below critical levels. Alterations based on D-Flow and Chl-a concentrations were assessed across seasons (spring, March–May; summer, June–August; autumn, September–November; and winter, December–February), as shown in Fig. 8.

Considering a Chl-a concentration of 20 mg/m^3 as the benchmark for good water quality, the daily mean flow rates required to achieve a concentration below this level were analyzed. In spring, summer, autumn, and winter, the daily mean flow rates necessary to reduce Chl-a concentrations to below 20 mg/m^3 were approximately 95, 85, 40, and 20 m^3/sec, respectively.

When averaged over 2 days, spring, summer, autumn, and winter required flow rates of approximately 30, 60, 35, and 15 m^3/sec, respectively. This implies that although elevated algal concentrations were observed due to limited D-Flow, Chl-a levels could be maintained within normal ranges despite modest D-Flow. Conversely, the elevated temperatures in summer required greater D-Flow to maintain optimal Chl-a concentrations.

Figure 8 also shows the sensitivity of Chl-a concentration to specific D-Flow ranges. In winter, Chl-a concentrations sharply declined with increasing D-Flow from 11.5 m^3/sec to 16.5 m^3/sec. In spring and fall, similar sharp decreases occurred with D-Flow increases from 11.5 m^3/sec to 26 m^3/sec. Winter exhibited a broader D-Flow range impacting Chl-a concentrations, and a distinct decrease was observed with D-Flow increasing from 11.5 m^3/sec to 50 m^3/sec. Each season exhibited specific D-Flow ranges where Chl-a concentration was particularly sensitive. Therefore, effectively reducing Chl-a concentration can be achieved by increasing D-Flow within the corresponding range for each season.

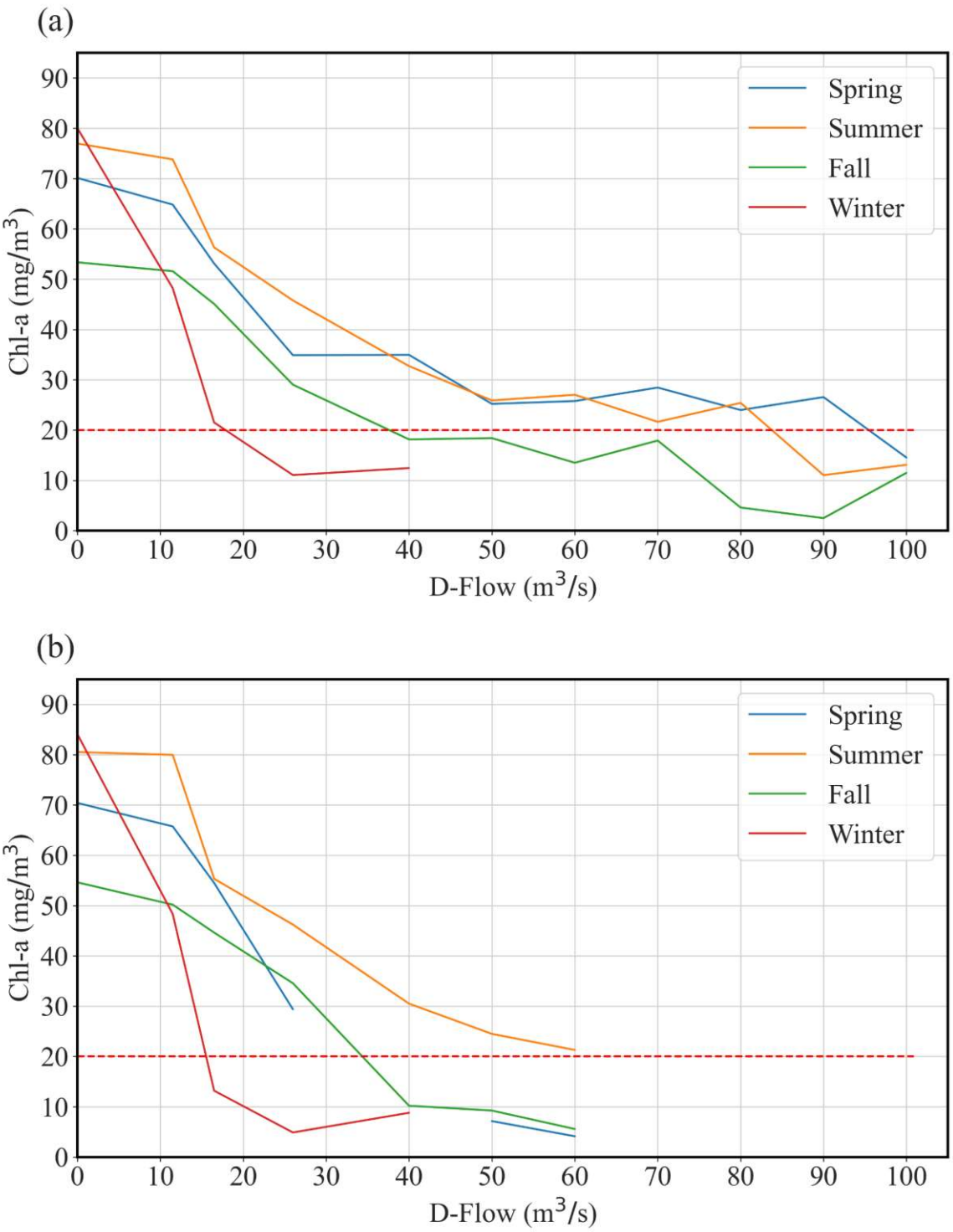


Fig. 8. Chlorophyll-a (Chl-a) concentration in relation to (a) 1-day and (b) 2-day average discharge flow rate (D-Flow) across the four seasons.

To examine the effect of Temp on the Chl-a concentration in more detail, the Chl-a concentration at changing D-Flow was examined by dividing the water temperature into increments of 5 °C. Figure 9 illustrates the relationship between Chl-a concentrations and daily average D-Flow (Fig. 9a for the 1-day and Fig. 9b for the 2-day average discharge flow rate). In Fig. 9a, it can be observed that as the temperature increased, the required D-Flow to achieve a Chl-a concentration of 20 mg/m^3 tended to increase, albeit with notable fluctuations at 15 and 20 °C. In Fig. 9b, a significant increase in the necessary D-Flow was evident at higher temperatures. Specifically, when Temp ranged from 0 to 10 °C, the D-Flow of 15 m^3/sec was necessary to attain the target Chl-a concentration of 20 mg/m^3. In contrast, at temperatures between 25 and 30 °C, the required D-Flow was much higher, at 60 m^3/sec. This finding highlights the substantial impact of D-Flow and Temp on the Chl-a concentration. The elevated Temp demand increased D-Flow to achieve the desired Chl-a level of 20 mg/m^3, signifying good water quality.

In Fig. 9, the sensitivity of Chl-a concentration to specific D-Flow ranges was observed, and these D-Flow ranges were dependent on water temperature. The sharp decrease in Chl-a concentration within specific D-Flow ranges resembled the results illustrated in Fig. 8. At 0–10 °C, a notable drop in Chl-a concentration was observed with the increase of D-Flow from 0 to 16.5 m^3/sec. Between 10–15 °C, the D-Flow ranges leading to the drop in Chl-a concentration extended to 0–26 m^3/sec. In the 25–30 °C range, the D-Flow ranges resulting in the drop of Chl-a concentration further extended to 0–40 m^3/sec. Above 30 °C, such a sharp drop in Chl-a concentration with the increase in D-Flow was not observed. In summary, the increase in water temperature led to the extension of D-Flow ranges affecting the sharp drop in Chl-a concentration, necessitating higher D-Flow for decreasing Chl-a concentration.

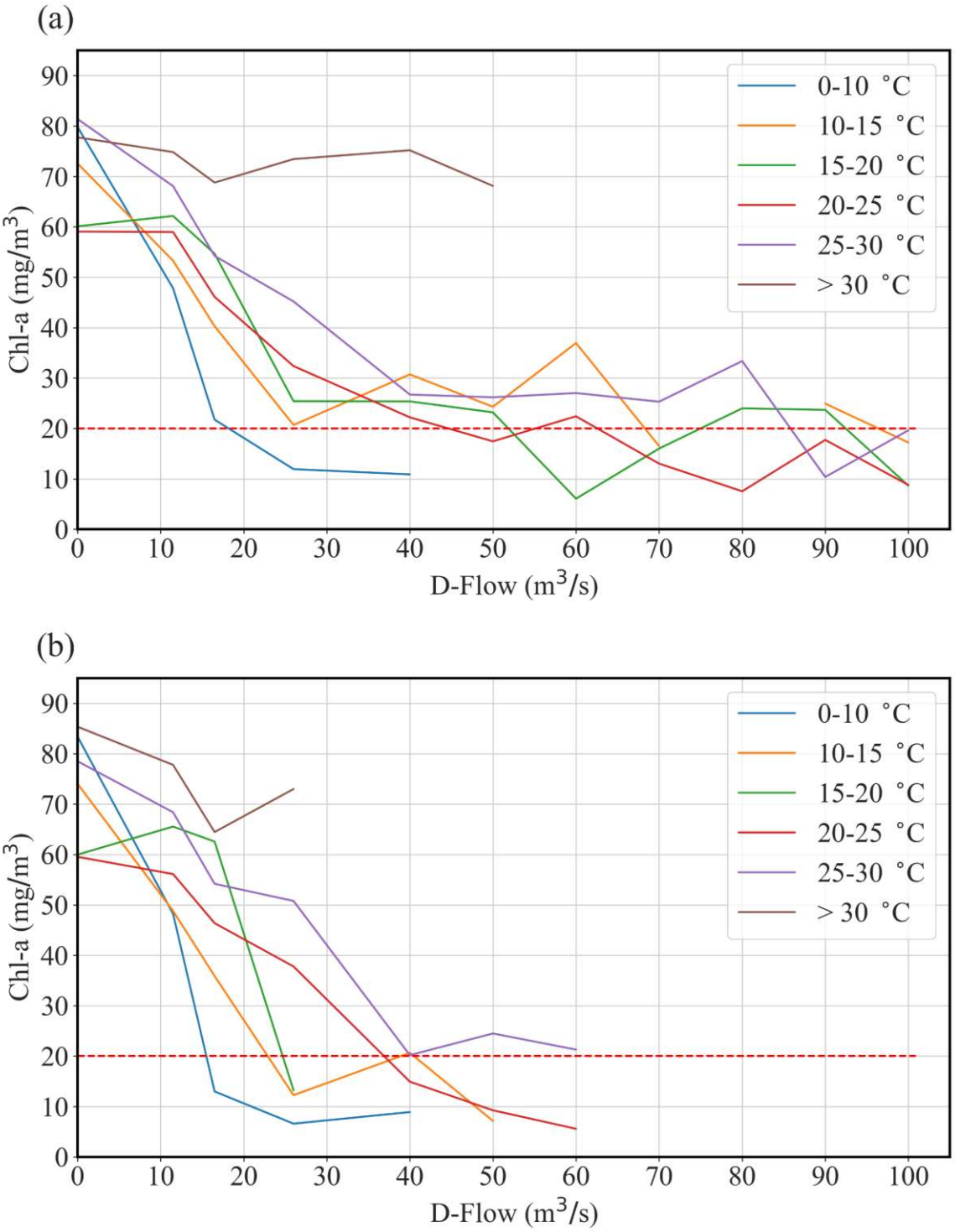


Fig. 9. Chlorophyll-a (Chl-a) concentration in relation to (a) 1-day and (b) 2-day average discharge flow rate (D-Flow) under varying water temperatures.

4. Conclusions

This study harnessed the potential of machine learning algorithms to forecast Chl-a concentrations while delving into the contributors to algal presence in a river ecosystem containing artificially constructed weirs. Four distinct machine learning models (EN, DT, RF, and GB) were deployed to predict Chl-a concentrations and identify the water quality and hydraulic factors influencing these levels. Among the model variants, GB exhibited the highest R^2 and the lowest RMSE values, indicating its superior suitability. The discharge flow rate and water temperature emerged as pivotal factors determining Chl-a concentration. To enhance the model accuracy, we tested the impact of past D-Flow and Temp on Chl-a concentration. Notably, the importance of D-Flow escalated as the averaging timeframe increased. The investigation involved plotting the Chl-a concentration over two weeks alongside the D-Flow and temperature data. Interestingly, higher Chl-a concentrations were observed at lower levels of both D-Flow and temperature, contrary to previous studies that suggested a direct increase in Chl-a with rising temperatures. In the Seungchon Weir, D-Flow emerged as the primary factor influencing Chl-a concentration, outweighing the impact of temperature. While the effect of temperature on Chl-a was relatively small, the required D-Flow for maintaining optimal Chl-a levels (20 mg/m^3) displayed variability in response to temperature fluctuations. Specifically, as the temperature increased, the D-Flow necessary to sustain a favorable Chl-a concentration also increased. The concentration of Chl-a demonstrated distinct sensitivity to varying D-Flow ranges within each season and temperature category. Therefore, the targeted reduction of Chl-a levels can be effectively achieved by strategically adjusting D-Flow within the designated range corresponding to each specific season and temperature condition. This study identifies D-Flow as the primary driver of algal development in artificially constructed river systems, and maintaining an appropriate D-Flow level is essential for sustaining ideal Chl-a concentrations. This study introduced a novel approach for identifying the causes of river algal

blooms and proposed a weir management strategy, along with a minimum D-Flow value, to prevent algal outbreaks. However, additional validation in diverse water bodies is necessary to confirm our hypothesis.

Funding: This research did not receive any specific grants from funding agencies in the public, commercial, or not-for-profit sectors.

Authors' contributions: Hyunju Kim: Writing—original draft, data analysis, and visualization; **Gyesik Lee:** Data analysis, writing—original draft, writing—review and editing; **Chang-Gu Lee:** Writing—review and editing; **Seong-Jik Park:** Conceptualization, writing—original draft, writing—review and editing, and supervision. All authors have read and agreed to the published version of the manuscript.

Data availability: All data generated or analyzed during this study are included in this published article. The datasets used and/or analyzed in the current study are available from the corresponding author upon reasonable request.

References

Ahmed, A.N., Othman, F.B., Afan, H.A., Ibrahim, R.K., Fai, C.M., Hossain, M.S., Ehteram, M., Elshafie, A., 2019. Machine learning methods for better water quality prediction. Journal of Hydrology 578, 124084. https://doi.org/10.1016/j.jhydrol.2019.124084.

Aires, U.R.V., Silva, D.D., Filho, E.I.F., Rodrigues, L.N., Uliana, E.M., Amorim, R.S.S., Ribeiro, C.B.M., Campos, J.A., 2022. Modeling of surface sediment concentration in the Doce River basin using satellite remote sensing. Journal of Environmental Management. 323(1), 116207. https://doi.org/10.1016/j.jenvman.2022.116207.

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 1984. Classification and regression

trees. Wadsworth International Group, Belmont CA.

Breiman, L., 2001. Random forests. Machine Learning 45, 5-32. https://doi.org/10.1023/A:1010933404324.

Çamdevýren, H., Demýr, N., Kanik, A., Keskýn, S., 2005. Use of principal component scores in multiple linear regression models for prediction of Chlorophyll-a in reservoirs. Ecological Modelling 181(4), 581-589. https://doi.org/10.1016/j.ecolmodel.2004.06.043.

Chong, S., Yi, H.S., Hwang, H.S., Kim, H.J., 2015. Modeling the flushing effect of multi-purpose weir operation on algae removal in Yeongsan River. Journal of the Korean Society of Environmental Engineering 37(10), 563-572. https://doi.org/10.4491/KSEE.2015.37.10.563.

Coffey, R., Paul, M.J., Stamp, J., Hamilton, A., Johnson, T., 2019. A review of water quality responses to air temperature and precipitation changes 2: nutrients, algal blooms, sediment, pathogens. JAWRA. 55(4), 844-868. https://doi.org/10.1111/1752-1688.12711.

Deng, T., Chau, K.W., Duan, H.F., 2021. Machine learning based marine water quality prediction for coastal hydro-environment management. Journal of Environmental Management 284, 112051. https://doi.org/10.1016/j.jenvman.2021.112051.

Friedman, J.H., 2001. Greedy function approximation: A gradient boosting machine. Annals of Statistics 29(5), 1189-1232. https://doi.org/10.1214/aos/1013203451.

Friedman, J.H. 2002. Stochastic gradient boosting. CSDA. 38(4), 367-378. https://doi.org/10.1016/S0167-9473(01)00065-2.

Friedman, J.H., Hastie, T., Tibshirani, R., 2010. Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software 33(1), 1-22. https://doi.org/10.18637/jss.v033.i01.

Furnas, M.J., 1990. In situ growth rates of marine phytoplankton: approaches to measurement,

community and species growth rates. Journal of Plankton Research 12(6), 1117-1151. https://doi.org/10.1093/plankt/12.6.1117.

Gamez, T.E., Benton, L., Manning, S.R., 2019. Observations of two reservoirs during a drought in Central Texas, USA: strategies for detecting harmful algal blooms. Ecological Indicators 104, 588–593. https://doi.org/10.1016/j.ecolind.2019.05.022.

Gupta, A., Hantush, M.M., Govindaraju, R.S., 2023. Sub-monthly time scale forecasting of harmful algal blooms intensity in Lake Erie using remote sensing and machine learning. Science of the Total Environment, 900, 165781. https://doi.org/10.1016/j.scitotenv.2023.165781.

Hage, A., Luckett, N., Holbrook, G.P., 2018. Phycoremediation of Municipal Wastewater by the Cold- Adapted Microalga Monoraphidium sp. Dek19. Water Environmental Research 90(11), 1938-1946. https://doi.org/10.2175/106143017X15131012188060.

Hayden, N.J., Roelke, D.L., Brooks, B.W., Grover, J.P., Neisch, M.T., Valenti Jr, T.W., Prosser, K.N., Gable, G.M., Umphres, G.D., Hewitt, N.C., 2012. Beyond hydraulic flushing: Deep water mixing takes the harm out of a haptophyte algal bloom. Harmful Algae, 20, 42-57. https://doi.org/10.1016/j.hal.2012.07.006.

Hong, S.H., Ndingwan, A.M., Yoo, S.C., Lee, C.G., Park, S.J., 2020. Use of calcined sepiolite in removing phosphate from water and returning phosphate to soil as phosphorus fertilizer. J. Environ. Manage. 270, 110817. https://doi.org/10.1016/j.jenvman.2020.110817.

Hong, S.M., Abbas, A., Kim, S., Kwon, D.H., Yoon, N., Yun, D., Lee, S., Pachepsky, Y., Pyo, J.C., Cho, K.H., 2023. Autonomous calibration of EFDC for predicting chlorophyll-a using reinforcement learning and a real-time monitoring system. Environmental Modelling & Software, 168, 105805. https://doi.org/10.1016/j.envsoft.2023.105805.

Jargal, N., Lee, E.H., An, K.G., 2023. Monsoon-induced response of algal chlorophyll to trophic state, light availability, and morphometry in 293 temperate reservoirs. Journal of Environmental Management 337, 117737. https://doi.org/10.1016/j.jenvman.2023.117737.

Kang, J.K., Seo, E.J., Lee, C.G., Jeong, S., Park, S.J., 2022. Application of response surface methodology and artificial neural network for the preparation of Fe-loaded biochar for enhanced Cr (VI) adsorption and its physicochemical properties and Cr (VI) adsorption characteristics. Environmental Science and Pollution Research 29(40), 60852-60866. https://doi.org/10.1007/s11356-022-20009-3

Keller, S., Maier, P.M., Riese, F.M., Norra, S., Holbach, A., Börsig, N., Wilhelms, A., Moldaenke, C., Zaake A., Hinz, S., 2018. Hyperspectral data and machine learning for estimating CDOM, chlorophyll a, diatoms, green algae and turbidity. International Journal of Environmental Research and Public Health 15(9), 1881. https://doi.org/10.3390/ijerph15091881.

Kim, J., Jones, J.R., Seo, D., 2021a. Factors affecting harmful algal bloom occurrence in a river with regulated hydrology. Journal of Hydrology: Regional Studies. 33, 100769. https://doi.org/10.1016/j.ejrh.2020.100769.

Kim, J.H., Shin, J.K., Lee, H., Lee, D.H., Kang, J.H., Cho, K.H., Lee, Y.G., Chon, K., Park, Y., 2021b. Improving the performance of machine learning models for early warning of harmful algal blooms using an adaptive synthetic sampling method. Water Research 207, 117821. https://doi.org/10.1016/j.watres.2021.117821.

Kwak, J., 2021. A study on the 3-month prior prediction of Chl-a concentration in the Daechong Lake using hydrometeorological forecasting data. Journal of Wetlands Research 23(2), 144-153. https://doi.org/10.17663/JWR.2021.23.2.144.

Lee, H.J., Kim, H.J., Choi, K.S., 2017. Investigation and monitoring of causes of algal blooms

in the four major rivers. Water for Future. 50(6), 20-25.

Lee, Y.J., Im, E.C., Lee, G., Hong, S.C., Lee, C.G., Park, S.J. 2024. Comparison of ammonia volatilization in paddy and field soils fertilized with urea and ammonium sulfate during rice, potato, and Chinese cabbage cultivation. Atmospheric Pollution Research, 15(4), 102049. https://doi.org/10.1016/j.apr.2024.102049.

Li, J., Yin, W., Jia, H., Xin, X., 2021. Hydrological management strategies for the control of algal blooms in regulated lowland rivers. Hydrological Processes 35(6), e14171. https://doi.org/10.1002/hyp.14171.

Lian, A., Han, D., Song, X., Yang, S., 2021. Impacts of storm events on Chlorophyll-a variations and controlling factors for algal bloom in a river receiving reclaimed water. Journal of Environmental Management 297(1), 113376. https://doi.org/10.1016/j.jenvman.2021.113376.

Liu, Y., Wang, Y., Zhang, J., 2012. New machine learning algorithm: Random forest. In:Liu, B., Ma, M., Chang, J.(eds) Information Computing and Applications. ICICA 2012. Lecture Notes in Computer Science. Springer, Berlin, Heidelberg.

Liu, N., Yang, Y., Li, F., Ge, F., Kuang, Y., 2016. Importance of controlling pH-depended dissolved inorganic carbon to prevent algal bloom outbreaks. Bioresources And Technology 220, 246-252. https://doi.org/10.1016/j.biortech.2016.08.059.

Long, T.Y., Wu, L., Meng, G.H., Guo, W.H., 2011. Numerical simulation for impacts of hydrodynamic conditions on algae growth in Chongqing Section of Jialing River, China. Ecological Modelling 222(1), 112-119.

Lu, H., Ma, X., 2020. Hybrid decision tree-based machine learning models for short-term water quality prediction. Chemosphere. 249, 126169. https://doi.org/10.1016/j.chemosphere.2020.126169.

Lv, J., Wu, H., Chen, M., 2011. Effects of nitrogen and phosphorus on phytoplankton

composition and biomass in 15 subtropical, urban shallow lakes in Wuhan, China. Limnologica 41, 48–56. https://doi.org/10.1016/j.limno.2010.03.003.

Ly, Q.V., Tong, N.A., Lee, B.M., Nguyen, M.H., Trung, H.T., Le Nguyen, P., Hoang, T.H., Hwang, Y., Hur, J., 2023. Improving algal bloom detection using spectroscopic analysis and machine learning: A case study in a large artificial reservoir, South Korea. Science of The Total Environment. 901, 166467. https://doi.org/10.1016/j.scitotenv.2023.166467

Makhotin, I., Koroteev, D., Burnaev, E., 2019. Gradient boosting to boost the efficiency of hydraulic fracturing. Journal of Petroleum Exploration and Production Technology 9, 1919-1925. https://doi.org/10.1007/s13202-019-0636-7.

Mitrovic, S.M., Chessman, B.C., Bowling, L.C., Cooke, R.H., 2006. Modelling suppression of cyanobacterial blooms by flow management in a lowland river. River Research Applications. 22(1), 109-114. https://doi.org/10.1002/rra.875.

Mng'ong'o, M., Munishi, L.K., Blake, W., Comber, S., Hutchinson, T.H., Ndakidemi, P.A., 2022. Towards sustainability: Threat of water quality degradation and eutrophication in Usangu agro-ecosystem Tanzania. Marine Pollution Bulletin, 181, 113909. https://doi.org/10.1016/j.marpolbul.2022.113909.

Ogutu, J.O., Schulz-Streeck, T., Piepho, H.P., 2012. Genomic selection using regularized linear regression models: ridge regression, lasso, elastic net and their extensions. BMC proceedings. 6, S10. https://doi.org/10.1186/1753-6561-6-S2-S10.

Park, Y., Cho, K.H., Park, J., Cha, S.M., Kim, J.H., 2015. Development of early-warning protocol for predicting chlorophyll-a concentration using machine learning models in freshwater and estuarine reservoirs, Korea. Science of the Total Environment 502, 31-41. https://doi.org/10.1016/j.scitotenv.2014.09.005.

Park, Y., Lee, H.K., Shin, J.K., Chon, K., Kim, S., Cho, K.H., Kim, J.H., Baek, S.S., 2021. A machine learning approach for early warning of cyanobacterial bloom outbreaks in a

freshwater reservoir. Journal of Environmental Management 288, 112415. https://doi.org/10.1016/j.jenvman.2021.112415.

Raven, J.A., Gobler, C.J., Hansen, P.J., 2020. Dynamic CO2 and pH levels in coastal, estuarine, and inland waters: Theoretical and observed effects on harmful algal blooms. Harmful Algae, 91, 101594. https://doi.org/10.1016/j.hal.2019.03.012.

Shin, C.M., Min, J.H., Park, S.Y., Choi, J., Park, J.H., Song, Y.S., Kim, K., 2017. Operational water quality forecast for the Yeongsan River using EFDC model. Journal of the Korean Society of Water Environment 33(2), 219-229. https://doi.org/10.15681/KSWE.2017.33.2.219.

Silva, L., Calleja, M.L., Ivetic, S., Huete-Stauffer, T., Roth, F., Carvalho, S., Morán, X.A.G., 2021. Heterotrophic bacterioplankton responses in coral-and algae-dominated Red Sea reefs show they might benefit from future regime shift. Science of the Total Environment 751, 141628. https://doi.org/10.1016/j.scitotenv.2020.141628.

Suggett, D.J., Prášil, O., Borowitzka, M.A., 2010. Chlorophyll a fluorescence in aquatic sciences: methods and applications, Springer.

Water Resources Management Information System, 2018. Water quality standards. http://www.wamis.go.kr/wke/wke_wqbase_lst.do (accessed 26 August 2023)

Wehr, J.D., Descy, J.P., 1998. Use of phytoplankton in large river management, Journal of Phycology 34(5), 741–749. https://doi.org/10.1046/j.1529-8817.1998.340741.x.

Xu, H., Paerl, H.W., Qin, B., Zhu, G., Hall, N.S., Wu, Y., 2015. Determining critical nutrient thresholds needed to control harmful cyanobacterial blooms in eutrophic Lake Taihu, China. Environmental Science and Technology 49(2), 1051-1059. https://doi.org/10.1021/es503744q

Xu, H., Paerl, H.W., Zhu, G., Qin, B., Hall, N.S., Zhu, M., 2017. Long-term nutrient trends and

harmful cyanobacterial bloom potential in hypertrophic Lake Taihu, China. Hydrobiologia, 787, 229-242. https://doi.org/10.1007/s10750-016-2967-4.

Yang, Q., Liu, G., Hao, Y., Zhang, L., Giannetti, B.F., Wang, J., Casazza, M., 2019. Donor-side evaluation of coastal and marine ecosystem services. Water Research 166, 115028. https://doi.org/10.1016/j.watres.2019.115028.

Yaqub, M., Ngoc, N.M., Park, S., Lee, W., 2022. Predictive modeling of pharmaceutical product removal by a managed aquifer recharge system: Comparison and optimization of models using ensemble learners. Journal of Environmental Management. 324(15), 116345. https://doi.org/10.1016/j.jenvman.2022.116345.

Yu, P., Gao, R., Zhang, D., Liu, Z.P., 2021. Predicting coastal algal blooms with environmental factors by machine learning methods. Ecological Indicators 123, 107334. https://doi.org/10.1016/j.ecolind.2020.107334.

Zamparas, M., Zacharias, I., 2014. Restoration of eutrophic freshwater by managing internal nutrient loads. A review. Science of the Total Environment 496, 551-562. https://doi.org/10.1016/j.scitotenv.2014.07.076.

Zhou, S., Shao, Y., Gao, N., Deng, Y., Li, L., Deng, J., Tan, C., 2014. Characterization of algal organic matters of Microcystis aeruginosa: biodegradability, DBP formation and membrane fouling potential. Water Research 52, 199-207. https://doi.org/10.1016/j.watres.2014.01.002.

Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. Journal of the royal statistical society series B: Statistical Methodology 67(2), 301-320. https://doi.org/10.1111/j.1467-9868.2005.00503.x.

Supplementary Information

Algae Development in Rivers with Artificially Constructed Weirs: Dominant Influence of Discharge Over Temperature

Hyunju Kim [a], Gyesik Lee [b,*], Chang-Gu Lee [c], Seong-Jik Park [d,*]

a Faculty of Liberal Education, Seoul National University, Seoul, 08826, Republic of Korea

b School of Computer Engineering and Applied Mathematics, Hankyong National University, Anseong, 17579, Republic of Korea

c Department of Environmental and Safety Engineering, Ajou University, Suwon 16499, Republic of Korea

d Department of Bioresources and Rural System Engineering, Hankyong National University, Anseong, 17579, Republic of Korea

⁎ ***Corresponding author***: S. J. Park

E-mail address: parkseongjik@hknu.ac.kr

ORCID: 0000-0003-2122-5498

G. Lee

E-mail address: gslee@hknu.ac.kr

Table S1. Best parameters for each model (elastic net (EN), decision tree (DT), random forest (RF), and gradient boosting (GB)) found by the grid search

Model	Best parameter
EN	alpha: 0.001, l1_ratio: 0.1
DT	max_depth: 25, min_sample_leaf: 4
RF	max_depth: 22, min_sample_leaf: 1, n_estimators: 500
GB	max_depth: 22, min_sample_leaf: 16, n_estimators: 500

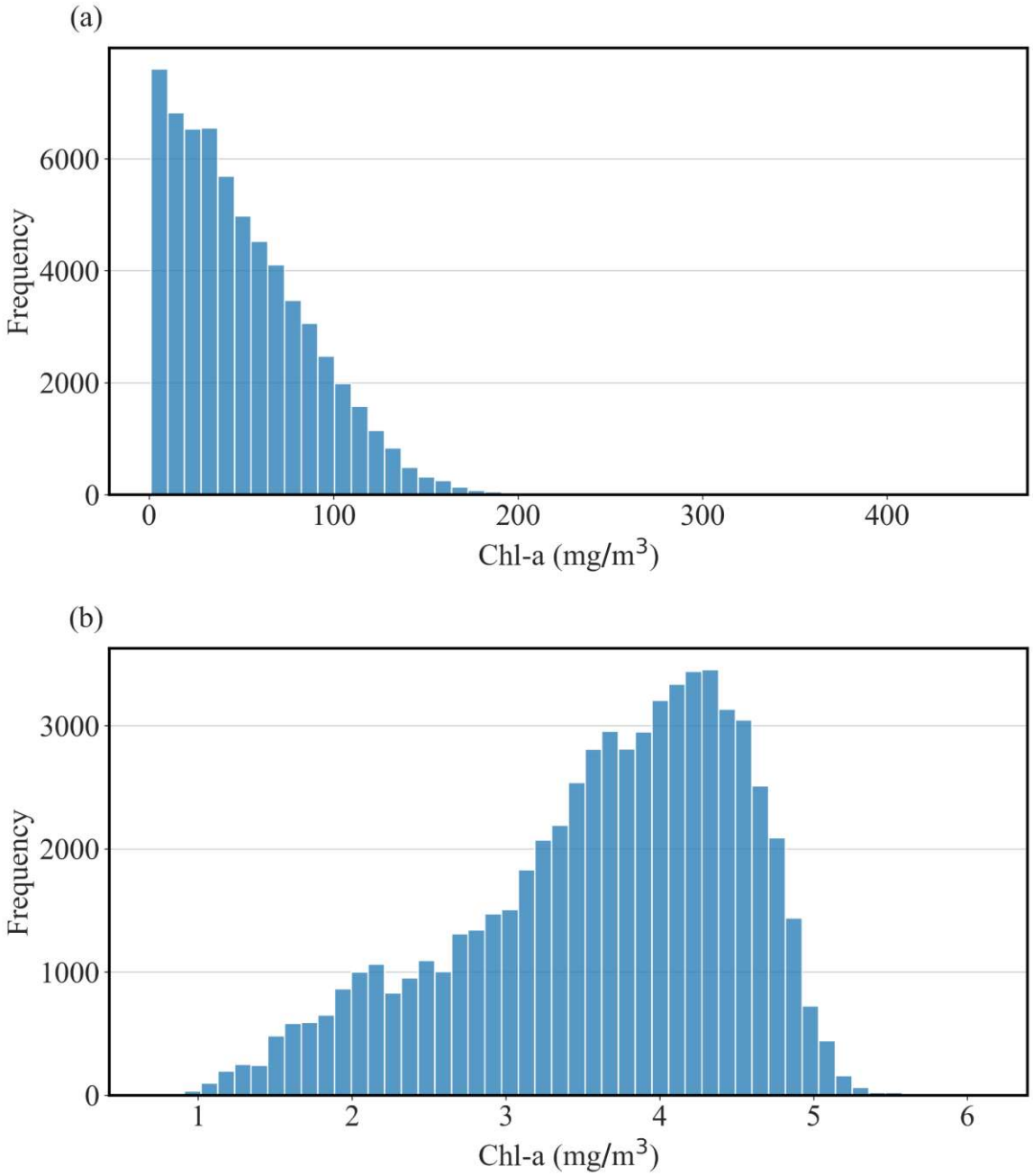


Fig. S1. Log transformation of the data set: (a) distribution of the original data set and (b) distribution after log transformation.

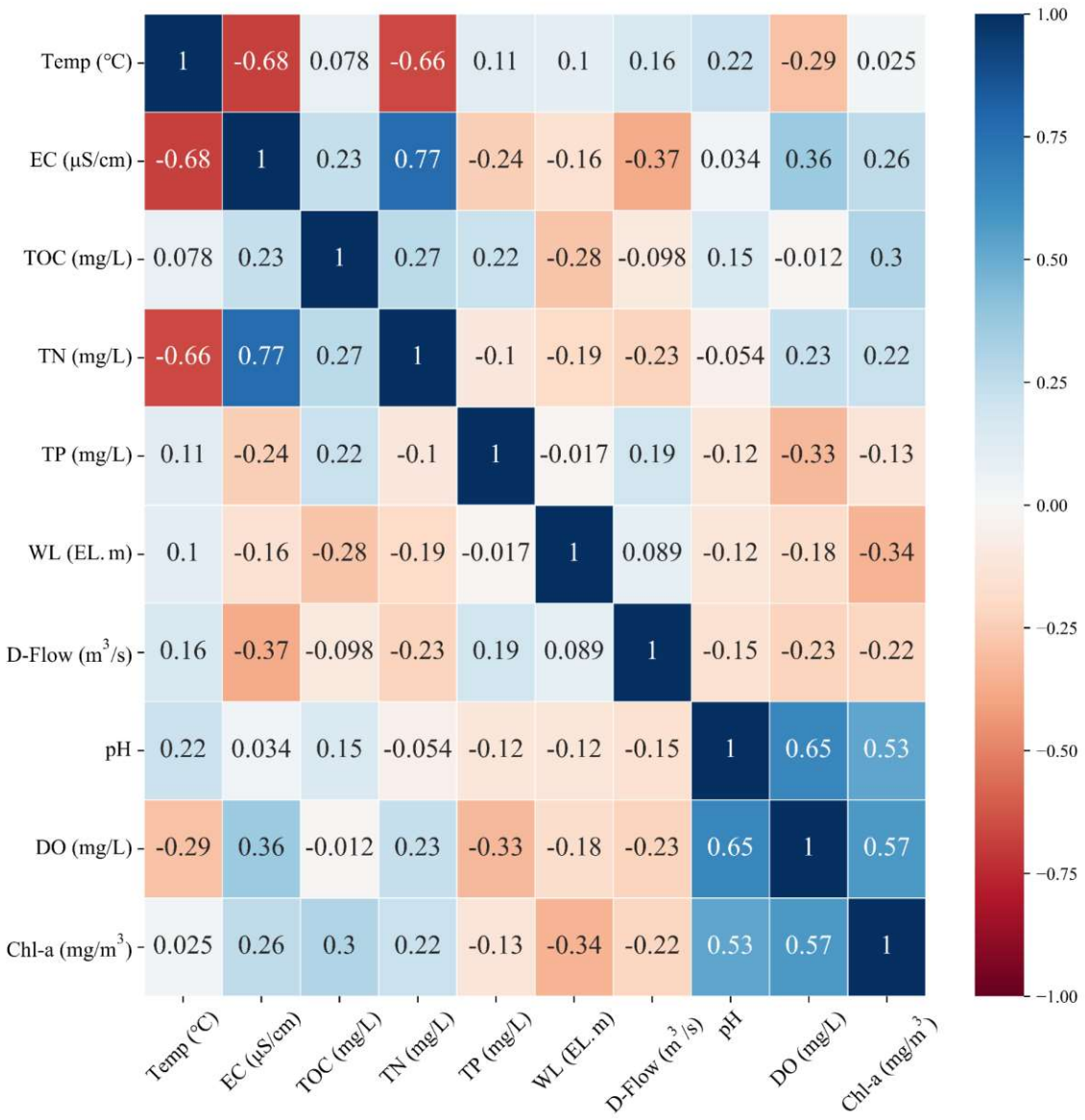


Fig. S2. Correlation analysis of water quality and hydraulic condition parameters including water temperature (Temp), electrical conductivity (EC), total organic carbon (TOC), total nitrogen (TN), total phosphorus (TP), water level (WL), discharge flow rate (D-Flow), pH, and dissolved oxygen (DO), and chlorophyll-a (Chl-a).

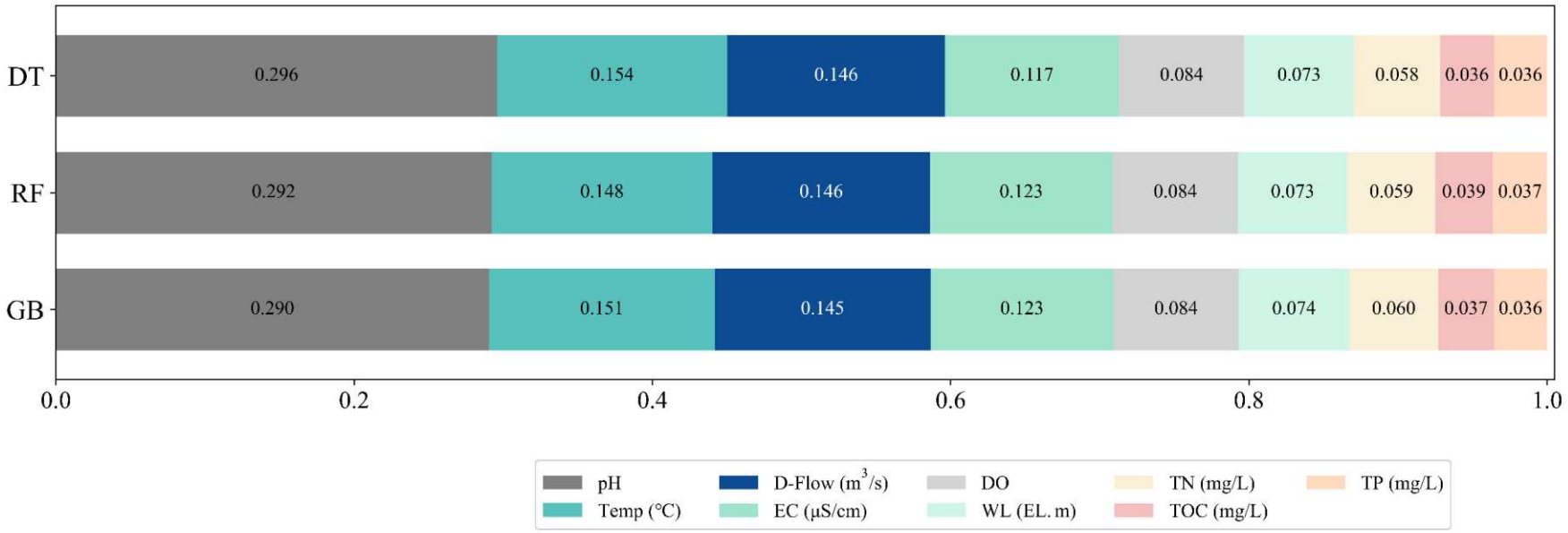


Fig. S3. Feature importance of pH, discharge flow rate (D-Flow), dissolved oxygen (DO), total nitrogen (TN), total phosphorus (TP), temperature (Temp), electric conductivity (EC), water level (WL), and total organic carbon (TOC) in the prediction of Chl-a concentration obtained from decision tree (DT), random forest (RF), and gradient boosting (GB) models.

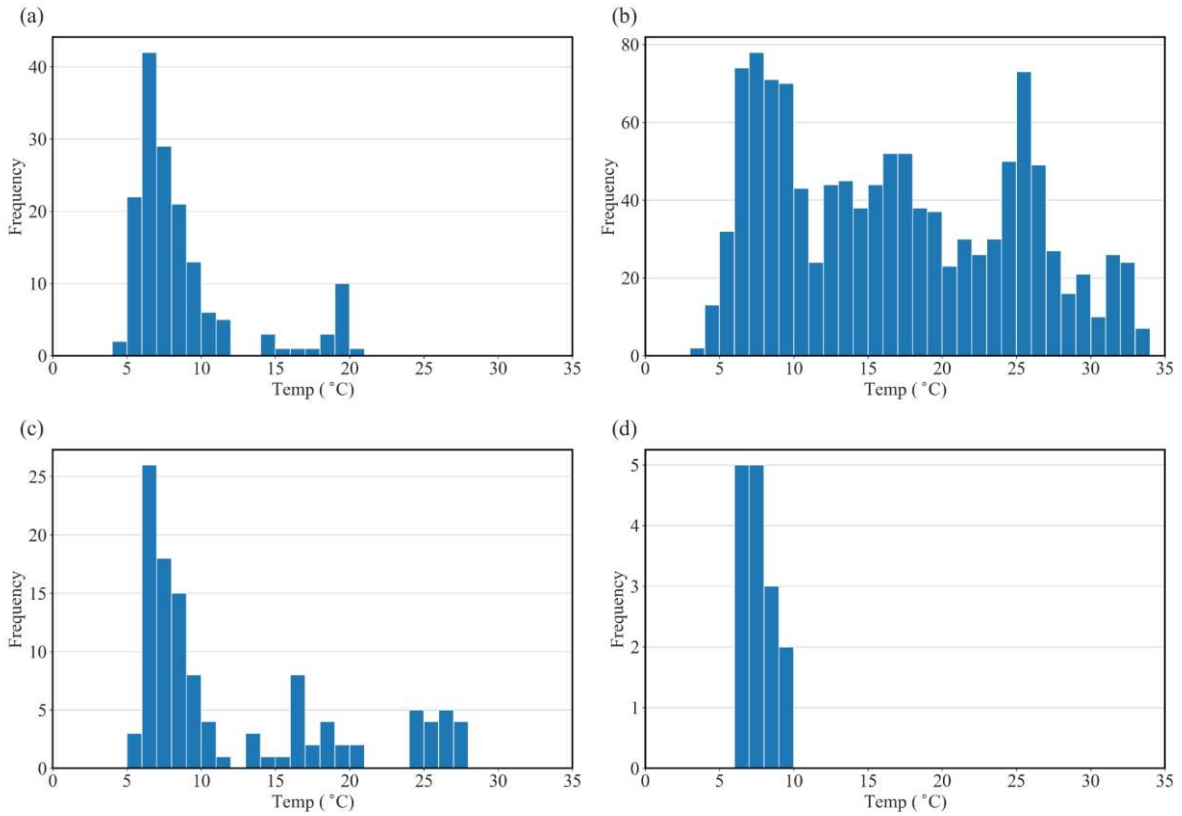


Fig. S4. Distribution of water temperature (Temp) under different chlorophyll-a (Chl-a) concentrations prolonged for two weeks: (a) ≤ 20 mg/m^3, (b) > 20 mg/m^3, (c) > 70 mg/m^3, and (d) > 100 mg/m^3