GIScience & Remote Sensing

ISSN: 1548-1603 (Print) 1943-7226 (Online) Journal homepage: www.tandfonline.com/journals/tgrs20

Improving chlorophyll-a estimation using Sentinel-2 data: a comparative analysis of augmented datasets

Jinmyeong Lee, Do Hyuck Kwon, Heewon Jeong, Gibeom Nam, Euiho Hwang, Jin Hwi Kim, Kyung Hwa Cho & Hyo Gyeom Kim

To cite this article: Jinmyeong Lee, Do Hyuck Kwon, Heewon Jeong, Gibeom Nam, Euiho Hwang, Jin Hwi Kim, Kyung Hwa Cho & Hyo Gyeom Kim (2025) Improving chlorophyll-a estimation using Sentinel-2 data: a comparative analysis of augmented datasets, GIScience & Remote Sensing, 62:1, 2496551, DOI: 10.1080/15481603.2025.2496551

To link to this article: https://doi.org/10.1080/15481603.2025.2496551

GISCIENCE & REMOTE SENSING
2025, VOL. 62, NO. 1, 2496551
https://doi.org/10.1080/15481603.2025.2496551

Taylor & Francis
Taylor & Francis Group

Improving chlorophyll-a estimation using Sentinel-2 data: a comparative analysis of augmented datasets

Jinmyeong Lee[a], Do Hyuck Kwon[b], Heewon Jeong[c], Gibeom Nam[d], Euiho Hwang[d], Jin Hwi Kim[c], Kyung Hwa Cho[e] and Hyo Gyeom Kim[f]

[a]Environmental Engineering, Chungnam National University, Daejeon, Republic of Korea; [b]Department of Civil Urban Earth and Environmental Engineering, Ulsan National Institute of Science and Technology, Ulsan, Republic of Korea; [c]Future and Fusion Lab of Architectural, Civil and Environmental Engineering, Korea University, Seoul, Republic of Korea; [d]Water Resources Satellite Center, K-water Research Institute, K-water, Daejeon, Republic of Korea; [e]School of Civil, Environmental and Architectural Engineering, Korea University, Seoul 02841, Republic of Korea; [f]Environmental Assessment Group, Korea Environment Institute, Sejong, Republic of Korea

ABSTRACT

Although remote sensing using machine learning techniques can effectively monitor harmful algal blooms, their application is often limited by data availability. The synergetic impacts of rapid urbanization and climate change contribute to the unprecedented occurrence of severe algal blooms, which require sufficient high-concentration data for successful model training. In this study, we evaluated the feasibility of integrating datasets from two different watersheds to estimate chlorophyll-a (Chl-a) concentrations using machine learning models with Sentinel-2 imagery. The original dataset, consisting of data from the Nakdong (ND) River, and two augmented datasets – an integrated dataset combining the Geum (GE) and ND rivers (GEND) and a resampled ND dataset using the synthetic minority oversampling technique for regression with Gaussian noise (ND-SMOGN) – were used to train six machine learning models. Models trained on the augmented datasets, GEND and ND-SMOGN, successfully addressed this underestimation issue for the sample with the highest Chl-a concentration. Among the six algorithms, multilayer perceptron with attention mechanism exploited the highest performance across all indicators with coefficient of determination (R^2) and root mean square error (RMSE) values of 0.93 and 2.76. Model interpretations revealed that models trained on GEND assigned high significance to B03 (560 nm) to B05 (705 nm), aligning with the optical characteristics of Chl-a, whereas models trained on ND and ND-SMOGN also emphasized less relevant bands. This study provides valuable insights into improving model performance, understanding the impacts of data availability, and informing the development of more accurate and reliable environmental management practices.

ARTICLE HISTORY
Received 21 August 2024
Accepted 15 April 2025

KEYWORDS
Cyanobacteria; remote sensing; machine learning; explainable AI; Sentinel-2

1. Introduction

Algal blooms, the proliferation of algae in aquatic environments, can greatly degrade water quality and disrupt the ecological balance (Kwon et al. 2023). Harmful algal blooms (HABs) are of particular concern because they produce toxins, such as microcystin and anatoxin-a, which adversely affect both ecosystem and human health (Jaffari et al. 2024; Kim, Hwa Cho, and Recknagel 2024). The frequency and intensity of HABs are expected to increase owing to rapid urbanization and climate change, underscoring the need for effective monitoring and management strategies (Burford et al. 2020; Gobler 2020). Severe algal blooms have also been recorded at unexpected periods and places (Li et al. 2023; Mertens et al. 2023).

Remote sensing using satellites, aircraft, and drones is highly effective in the wide-scale monitoring of HABs (Kiefer et al. 2015; Kwon et al. 2020; Moses et al. 2012; Pyo et al. 2018; Zhou et al. 2014). Multispectral and hyperspectral images acquired from remote-sensing platforms provide high-resolution optical information about aquatic environments, enabling the determination of water quality (Shin et al. 2024; Soomets et al. 2020). Recently, the integration of machine learning (ML) techniques has significantly advanced the field of remote sensing, facilitating the development of accurate models for HAB monitoring (Chusnah and Chu 2022; Silveira Kupssinskü et al. 2020). Decision tree (DT)-based models such as random forest (RF), and neural network

CONTACT Kyung Hwa Cho khcho80@korea.ac.kr; Hyo Gyeom Kim 1004kyeom@gmail.com
Supplemental data for this article can be accessed online at https://doi.org/10.1080/15481603.2025.2496551

(NN)-based models, such as multilayer perceptron (MLP), were highly effective in capturing the complex nonlinear relationships between Chl-a concentrations and remote sensing reflectance (Aptoula and Ariman 2021; Kolluru and Prakash Tiwari 2022; Mpakairi et al. 2024; Pyo et al. 2022). Furthermore, advanced techniques in deep learning, such as attention mechanisms, has enabled a deeper understanding of the interrelationships among input variables (Vaswani 2017). These advances have led to the development of more sophisticated NN-based models that further enhance the accuracy and reliability of Chl-a concentration predictions from remote sensing data (Hong et al. 2022).

Despite the potential of ML and deep learning techniques in remote sensing, their effectiveness in estimating Chl-a concentrations can be limited by the availability of adequate training data (Li et al. 2023; Yao et al. 2023). However, the need to collect sufficient training data often defeats the primary purpose of remote sensing, which is to minimize or even eliminate the laborious and time-consuming process of field sampling and Chl-a analysis. Even with an adequate number of Chl-a data points, the data are often imbalanced, with a lack of high-concentration samples, due to its nature (Shin et al. 2024). Moreover, the emergence of unprecedented high-concentration algal blooms can complicate the determination of their intensity and frequency in areas where they were previously rare (Kim, Cha, and Hwa Cho 2024). To address these data imbalance problems, several studies have suggested resampling methods, such as under- and over-sampling for improving data quality (Kim et al. 2021, 2023, 2024; Shin et al. 2021). However, while these methods improve data balance, they may distort the inherent characteristics of the dataset, particularly when applied to a limited number of data points (Branco, Torgo, and Ribeiro 2017; Demircioğlu 2024).

Traditionally, remote sensing data from different watersheds have been treated separately due to variations in optical properties, which hinders the utilization of integrated data across diverse environments (Cao et al. 2022; Hong et al. 2022; Mobley 1995; Ogashawara, Mishra, and Gitelson 2017; Yu et al. 2014). However, synthetic data generated through resampling, as an alternative approach, may instead introduce biases, thereby limiting transferability and complicating the interpretation of relationships between reflectance and algal bloom dynamics. To overcome these challenges, this study adopts the approach of integrating real-world data from different watersheds, rather than relying solely on synthetic data generated through resampling. This strategy aims to address data imbalance while maintaining the inherent characteristics of the data, providing a more robust solution for improving Chl-a concentration estimation across diverse aquatic environments.

Although ML models often demonstrate remarkable performance, their black-box nature limits intuitive understanding of the underlying data processes and resulting predictions. To address this, Shapley additive explanations (SHAP) (Lundberg and Lee 2017) have been developed to provide interpretability for models trained on various datasets, including original, resampled, and augmented datasets. SHAP allows for insights into how variable importance changes with data augmentation, offering a detailed assessment of the contributions and impacts of different augmentation strategies on model predictions.

This study aimed to enhance the accuracy and interpretability of Chl-a concentration prediction models, which could inform the development of improved HAB monitoring and management strategies in diverse watershed environments. The main objectives of this study were to: (1) develop ML models for Chl-a concentration estimation using original and augmented datasets; (2) determine the best model structure for estimating Chl-a using sentinel-2 multispectral data; (3) identify the feasible data augmentation strategy; and (4) analyze significant wavelengths for each model using SHAP.

2. Materials and methods

2.1. Site description

We surveyed the watersheds of two major rivers in South Korea: the Geum (GE) (N 35.35°–37.03° N, 126.40°–128.04° E) and Nakdong (ND) rivers (35.00°–37.21° N, 127.49°–129.30° E). The GE (length: 360.70 km, watershed area: 9,912.15 km^2) is located in mid-western South Korea (Figure 1a). More than half of the GE River watershed is surrounded by forests, and ~35% is used for agriculture (Choi, Koh, and Yeol Yoon 2023). The ND River, the longest national river (length: 511 km, watershed area: 23384 km^2) flows in

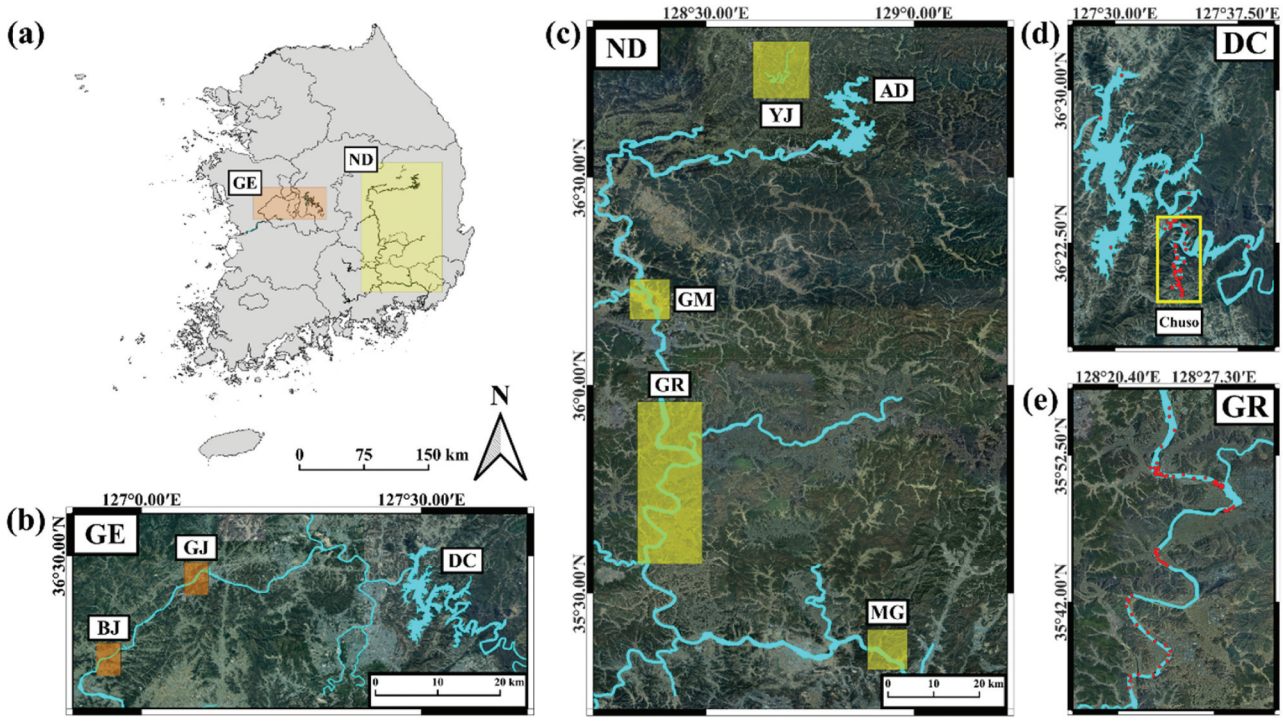

Figure 1. Description of study sites. Locations of (a) Geum (GE) and Nakdong (ND) rivers, (b) subbasins of Baekje Weir (BJ), Gongju Weir (GJ), and Daecheong Lake (DC) in GE; (c) Yeongju Dam (YJ), Gumi weir (GM), Gangjeong-Goryeong Weir (GR), and Mulgeum (MG) in ND. Field monitoring points in (d) DC (GE River) and (e) GR (ND River) are displayed as red points.

southeastern South Korea (Figure 1a). The major land-use types of this watershed are forest (67.50%) and agriculture (23.50%) (Jung et al. 2020). Both rivers are eutrophic in terms of Chl-a and total phosphorous concentrations, with cyanobacterial blooms frequently recorded in summer (Kim et al. 2020). Algal bloom issues have been most frequently reported in reservoirs within the GE River basin, while the ND River, a key source of drinking water, has also been affected by frequent occurrences (Pyo et al. 2022; Kim, Cho, and Recknagel 2024).

A total of eight sub-basins were selected for sample collection: four on the GE [Baekje Weir, Gongju Weir, Sejong Weir, and Daecheong (DC) Lake] and four on the ND [Yeongju Dam (YJ), Gumi Weir (GM), Gangjeong-Goryeong Weir (GR), and Mulgeum] (Figure 1b–c, Table S1). All locations have barriers, such as dams and weirs, to secure water resources and control flood risk by maintaining the water level and regulating river discharge. All sites have been reported to suffer from algal blooms in the summer season (Kim et al. 2019, 2024; Lee, Eun Kim, and Oh Baek 2023; Lee, Park, and Cheon 2018). Most sampling and monitoring were conducted in the Chuso region (Figure 1d) and GR (Figure 1e) in the GE and ND rivers, respectively (Table 1). An additional sub-basin was selected, Andong Lake (AD) along the ND River, to validate the model.

Table 1. Summary of monitoring and sampling events.

River	Subbasin	Number of samples	Monitoring period
Geum River	Baekje Weir	2	2021
	Daecheong Lake	15	2021–2023
	Gongju Weir	3	2021
Nakdong River	Gumi Weir	6	2021
	Gangjeong-Goryeong Weir	115	2020–2023
	Mulgeum	6	2021
	Yeongju Dam	11	2020

2.2. Study overview

This study consisted of four parts: (a) data acquisition, (b) model development, (c) qualitative validation, and (d) model interpretation (Figure 2). The Chl-a concentration and surface reflectance data were collected from field-monitored water samples and Sentinel-2 multispectral images. Three distinct datasets were generated to train the ML models: (1) ND, the original ND River dataset; (2) GEND, the integrated dataset of

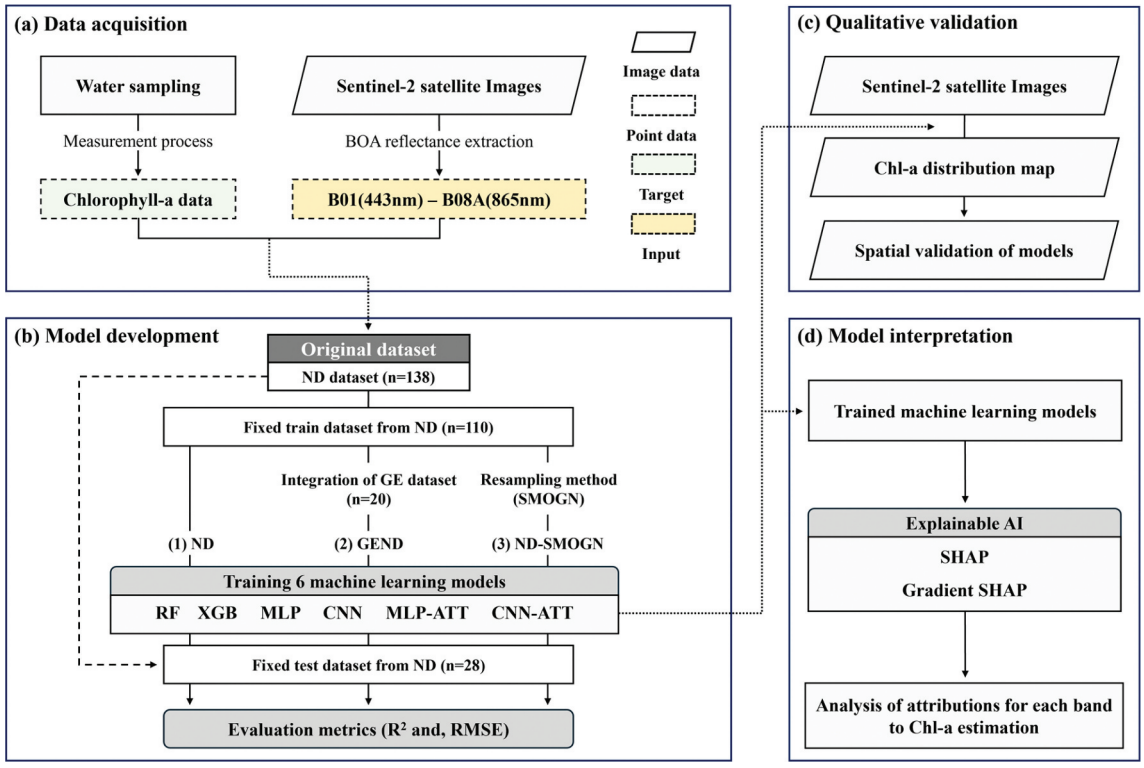

Figure 2. Flowchart for developing and interpreting ML models for chlorophyll-a (chl-a) concentration estimation. Each part indicates the steps for (a) data acquisition, (b) model development, (c) qualitative validation, and (d) model interpretation.

GE and ND river data; and (3) ND-SMOGN, the ND River dataset augmented using the synthetic minority oversampling technique for regression with Gaussian noise (SMOGN) technique. Six ML model algorithms – random forest (RF), extreme gradient boosting (XGB), multilayer perceptron (MLP), convolutional neural network (CNN), MLP with attention mechanism (MLP-ATT), and CNN with attention mechanism (CNN-ATT) – were applied to each dataset to estimate the Chl-a concentration. The trained models were quantitatively validated with two evaluation metrics, followed by further validation via Chl-a concentration distribution maps. Finally, the variable importance was computed to identify significant spectral wavelengths for Chl-a concentration estimation.

2.3. Data acquisition

2.3.1. Field monitoring

Field sampling was conducted from September 2021 to October 2023 at 20 points along the GE River and 138 points along the ND River. Most samples were collected at DC in the GE River (75.00%) and GR in the ND River (83.33%), particularly between July and October (Table 1). All water sampling dates were adjusted to coincide with the Sentinel-2 visit date. Given that algal blooms occur predominantly on the water surface, water samples were collected within 50 cm of the water bodies. The coordinates of the water sample locations were collected using global positioning system (GPS) equipment and used to extract local surface reflectance values from satellite imagery.

The Chl-a concentration was quantified using a standard method (Rice, Bridgewater, and American Public Health Association 2012). Water samples were filtered within 24 h using a 0.70 µm glass fiber filter and homogenized with acetone (9 + 1). After storage and centrifugation, the absorbance of the supernatant was measured at 630, 645, 663, and 750 nm with a Cary 5000 UV-Vis-NIR spectrophotometer (Agilent Inc., USA). In this study, a Chl-a concentration above 50 mg m^{-3} was the criterion for algal blooms, which is indicative of a high risk for harmful cyanobacterial blooms (World Health Organization 2003).

2.3.2. Sentinel-2 multispectral imagery

Sentinel-2, which consists of two satellites (Sentinel-2A and Sentinel-2B), provides observational data on

inland water. Each satellite is equipped with a multispectral instrument (MSI) and has a high revisit time of 5 d (Drusch et al. 2012). The MSI spectral bands consist of four visible (VIS), five visible and near-infrared (VNIR), and four shortwave infrared channels, which can achieve a high level of monitoring performance (Table S2). Sentinel-2 provides two primary reflectance data products: Level-1C (L1C) and Level-2A (L2A) (Sola et al. 2018). The L1C represents raw sensor observations, primarily comprising Top-of-atmosphere reflectance values that account for atmospheric effects (Nazeer et al. 2021). In contrast, the L2A data undergo additional processing, including atmospheric, aerosol, and adjacency corrections, representing bottom-of-atmosphere (BOA) reflectance values (Main-Knorn et al. 2017). In this study, L2A product images were used to estimate the Chl-a concentration on the water surface (Tables S3 and S4). By refining the data to correct for atmospheric disturbances, L2A provides a more accurate depiction of surface reflectance.

As Chl-a is optically active in the VIS and VNIR regions (Gitelson, Schalles, and Hladik 2007; Le et al. 2013), the corresponding spectral bands (B01–B08A) with a 60 m resolution were used. To align satellite data with observed Chl-a concentrations, surface reflectance at field sampling locations was extracted using Sentinel-2 imagery, geographic information systems (GIS), and corresponding geographic coordinates. The extracted BOA reflectance served as the predictor, and the in-situ Chl-a concentration was the prediction target. Furthermore, because vegetation along the riverbanks can interfere with the reflectance data, data collected from river edges were carefully excluded using GIS-based spatial filtering. Additionally, when generating the Chl-a distribution maps, we utilized the Sentinel-2 Scene Classification Layer (SCL) to mask out non-water pixels. Specifically, only pixels classified as water (SCL value = 6) were retained, while all other land or vegetation pixels were assigned a value of zero.

2.3.3. Data augmentation

Addressing the significant lack of high-concentration Chl-a data points in the ND dataset required employing two data augmentation methods. The first approach involves incorporating measurements of the Chl-a concentration and reflectance data from the GE River dataset into the ND dataset (GEND dataset). As combining datasets from different watersheds is generally uncommon in remote sensing due to variations in water quality and optical properties, we assessed this method as a data augmentation strategy.

The second method is SMOGN which generates synthetic data points to improve representation in sparsely populated areas of a dataset. By introducing Gaussian noise, SMOGN ensures the augmented data closely aligns with the original dataset's distribution while maintaining its integrity (Branco, Torgo, and Ribeiro 2017). To determine the appropriate level of perturbation, we conducted sensitivity analyses using perturbation parameter. This parameter, which controls the intensity of Gaussian noise, was experimentally adjusted and set to 0.5 to generate a sufficient amount of synthetic data within the underrepresented range.

To ensure a fair and consistent comparison between the original and augmented datasets, the ND dataset was initially divided into training (80%) and testing (20%) sets, with the test set fixed for evaluating model performance. Only the training set was augmented using two distinct approaches.

2.4. Model development

In this study, we developed DT-based models (RF and XGB) and NN-based models (MLP, CNN, MLP-ATT, and CNN-ATT) to predict the in situ Chl-a concentration using satellite imagery. Using three distinctive datasets (i.e. ND, GEND, and ND-SMOGN), the models were trained with satellite imagery from eight sub-basins and validated with spatial maps of another sub-basin, AD.

2.4.1. Decision tree-based models

The RF and XGB models are ensemble methods that improve model performance by combining multiple decision trees (DTs), but they differ in their ensemble strategies. RF employs bootstrap aggregating (bagging), which involves generating multiple datasets by random sampling with replacement (Breiman 2001). Each DT is trained on a different dataset to introduce diversity; randomness is further enhanced by selecting a subset of variables for splitting at each node (Quinlan 1986). The final prediction is obtained by averaging the outputs of all DTs for regression tasks:

$$Y_{prediction} = \frac{1}{N}\sum_{n=1}^{N} T_n(x) \quad (1)$$

where $Y_{prediction}$ denotes the combined result, N is the number of DTs, $T(x)$ is the output of each DT, and x is the bootstrapped input dataset from the original dataset. This approach increases stability and reduces overfitting by leveraging model diversity (Dietterich 2002).

In contrast, XGB, employs boosting, where DTs are trained sequentially to correct errors made by previous trees (Friedman 2001). Unlike RF, XGB assigns weights to each observation, focusing on samples with higher prediction errors during subsequent iterations. The final prediction is the weighted sum of the outputs from all DTs:

$$Y_{prediction} = \sum_{n=1}^{N} \alpha_n T_n(x) \quad (2)$$

where α_n is the weight of the n-th DT. Additionally, XGB optimizes the objective function by minimizing a combination of the loss function and regularization terms, improving both accuracy and generalization (Chen and Guestrin 2016).

2.4.2. Neural network-based models

The MLP is a basic artificial neural network (ANN) structure consisting of input, hidden, and output layers (Popescu et al. 2009; Yegnanarayana 2009). Nodes in each layer are connected to all nodes in the next layer, enabling data propagation and transformation (Sammartino et al. 2020). The input layer receives the band reflectance as input data (B01–B08A) while hidden layers apply weighted connections and activation functions to extract complex patterns. The output layer generates the model predictions as follows:

$$z^{l}_{j} = \tau\left(\sum_{i=0}^{N} w^{l-1}_{ji} \cdot z^{l-1}{}_{i} + b^{l-1}_{j0}\right) \quad (3)$$

where τ represents the activation function, z^{l}_{j} represents the *jth* node in the *lth* layer, w^{l-1}_{ji} denotes the weight connecting z^{l}_{j} and $z^{l-1}{}_{i}$, and b^{l-1}_{j0} is the bias term of the $l-1th$ layer. If the MLP is composed of 0–*Lth* layers, $z^{0}{}_{i}$ and $z^{L}{}_{j}$ are the input and output of the perceptron, respectively.

The 1D-CNN is an effective algorithm for treating one-dimensional data, such as hyperspectral imagery (Pyo et al. 2022; Shim et al. 2023). 1D-CNN models generally consist of convolution, pooling, and fully connected layers. A 1-D convolution extracts features by sliding a kernel (u) over the input (z), performing element-wise multiplication and summation to generate feature maps (Boureau, Ponce, and Le Cun 2010; Kiranyaz et al. 2021):

$$(z * u)(t) = \sum_{i=-(K-1)/2}^{(K-1)/2} z(t+i) \cdot u(i) \quad (4)$$

where z and u represent the input data and kernel, respectively, t refers to the position of the output data, and K denotes the size of the kernel.

The convolved input was down-sampled through max-pooling, resulting in a refined feature representation. This final feature map was subsequently passed to fully connected layers for prediction.

Despite its simple structure, MLP and 1D-CNN exhibit exceptional performance and are utilized across various fields in prediction and classification tasks (Hong, Yoon, and Hwa Cho 2024; Jeong et al. 2024; Tang et al. 2022), including remote water quality estimation (Kolluru and Prakash Tiwari 2022; Pyo et al. 2018).

2.4.3. Multi-head attention

The attention mechanism is a powerful tool in ML that enables models to focus selectively on relevant parts of the input data, improving their ability to capture intricate relationships (Vaswani 2017). It computes a weighted combination of values (V) based on the relevance of query (Q) and key (K) vectors. The attention mechanism is defined as:

$$Attention(Q, K, V) = softmax\left(\frac{QK^{T}}{\sqrt{d_k}}\right)V \quad (5)$$

where d_k is the dimensionality of the key vectors.

Multi-head attention is an advanced form of the attention mechanism that improves the model's capacity to capture complex relationships within the input data (Vaswani 2017). Unlike single-head attention, the mechanism employs multiple attention heads to independently compute attention scores, enabling the model to simultaneously focus on different aspects of the input. Each head learns distinct Q, K, and V matrices, which are combined to generate the final output:

$$MHA(Q,K,V) = Concat(head_1, head_2, \ldots, head_h)W^O \quad (6)$$

where $head_i = Attention\left(QW_i^Q, KW_i^K, VW_i^V\right)$ and W^Q, W^K, W^V, and W^O are learnable weight matrices. In this study, multi-head attention was implemented with eight attention heads after a fully connected layer. This setup enables the model to learn complex patterns among transformed features derived from spectral bands.

2.4.4. Hyperparameter optimization

Bayesian optimization (BO) of the machine-learning hyperparameters was performed to enhance model performance. The BO has two main components: i) a surrogate model and ii) objective function. Probabilistic models, such as Gaussian processes, are utilized as surrogate models to determine the uncertainty of the model predictions. The objective function is used to approximate the maximum improvement over the current model and iteratively select promising points within the search space (Shahriari et al. 2015).

Bayesian optimization effectively balances exploitation and exploration, enabling the search process to cover a broader range of hyperparameter combinations. As shown in Figure S1, the objective function fluctuates as the algorithm navigates the search space, iteratively increasing and decreasing before converging on the best-performing solution. The search space for the hyperparameters of each model algorithm, along with the optimal parameters of the models, are summarized in Tables S5–S8. The optimization steps were set to 200 for DT- and NN-based models. The number of initial points for exploration was set to one, and the expected improvement was uniformly computed with the objective function, root mean square error (RMSE).

2.4.5. Model performance evaluation metrics

To evaluate the performance of the models, the RMSE and coefficient of determination (R^2) were calculated. The RMSE, calculated from the square root of the mean-squared error, facilitates the transformation of evaluation metrics into units representing the actual values, enhancing intuitive evaluation. The R^2 value quantifies the proportion of variance in the dependent variable that is predictable from the independent variables, ranging from 0 to 1. Models that performed worse than the predicted mean value yielded a negative R^2 value. These metrics were calculated as follows:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y_i})^2} \quad (7)$$

and

$$R^2 = 1 - \frac{SSR}{SST} = 1 - \frac{\sum_{i=1}^{n}\left(y_i - \widehat{y_i}\right)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (8)$$

where y_i, $\widehat{y_i}$, and $\bar{y}$ denote the actual, predicted, and average Chl-a concentrations, respectively.

2.5. Model interpretation

In this study, SHAP and Gradient SHAP were employed to interpret the DT- and NN-based models to assess the importance of the eight input bands, respectively. SHAP, based on game theory, computes the marginal contribution of each participant in a game (Lundberg and Lee 2017). In ML models, it quantifies the contribution of each input variable to the model's output and can be described as:

$$\emptyset_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F|-|S|-1)!}{|F|!}\left[f_{S\cup\{i\}}\left(x_{S\cup\{i\}}\right) - f_S(x_s)\right] \quad (9)$$

where $\emptyset_i$ refers to the SHAP value of the i-th input feature, F represents the universal set of input data, S denotes the subset of the set F excluding the i-th feature, and f and $f(x)$ are the model and predicted value, respectively.

While SHAP provides intuitive insights for black box models, calculating exact SHAP values is computationally expensive as it requires evaluating all possible feature subsets. Gradient SHAP, proposed by Lundberg and Lee (2017), efficiently approximates SHAP values using gradients and baseline values. Gradient SHAP computes feature importance in three steps: (i) sampling between the baseline and input values, (ii) adding noise (e.g. Gaussian noise) to each input sample, and (iii) averaging the gradients of the model with respect to these inputs. The SHAP value is then calculated as:

$$SHAP\ value(x) = \frac{1}{N}\sum_{i=1}^{N}\frac{\partial F}{\partial x_i} \times (x_i - x') \quad (10)$$

where, N is the number of input samples, F refers to the model, and x represents the input value. Here, x is

the i-th input sample and x' denotes the baseline, which is set to a value of zero.

3. Results and discussion

3.1. Chlorophyll-a concentration distribution

The original (ND) and augmented datasets (GEND and ND-SMOGN) exhibited significantly skewed distributions with a right tail (Figure 3, Table S9). These right-tailed distributions are common in real-world data, particularly water quality measurements (Cao et al. 2020). ND exhibited a severe imbalance in Chl-a concentrations, with only three samples exceeding 50 mg m^{-3} out of a total of 138. Notably, the ND dataset contained no samples with Chl-a concentrations between 55 and 160 mg m^{-3}. To address this gap, the GEND and ND-SMOGN datasets were created. Differences in Chl-a distributions between the GE and ND datasets were primarily driven by variations in sampling sites and data preprocessing constraints related to cloud coverage. In the GE dataset, most high-concentration samples were collected from DC, where stagnant water conditions promoted algal blooms (Pyo et al. 2022). In contrast, the ND dataset contained few samples from lakes or reservoirs, as it primarily consisted of riverine samples, resulting in a scarcity of high Chl-a values. Additionally, some high-concentration Chl-a samples from the ND were excluded due to their incompatibility with satellite images. For example, in this study, in the YJ area, several high Chl-a concentrations were observed, but the corresponding surface reflectance values were obscured by cloud cover, preventing their use in the analysis.

The GEND dataset consisted of 20 samples from the GE dataset and 138 samples from the ND dataset. The added GE dataset samples had Chl-a concentrations ranging from 6.90 to 105.78 mg m^{-3}, including seven samples with concentrations exceeding 50 mg m^{-3}. As a result, the GEND dataset contained a total of 10 high-concentration samples (≥50 mg m^{-3}) out of the 158 total data points. The ND-SMOGN dataset comprised 158 data points, including 15 high-concentration samples. As SMOGN generates synthetic data points based on existing data by introducing noise, it has inherent limitations in generating values in the range that is entirely absent from the original dataset (Branco, Torgo, and Ribeiro 2017). Therefore, in the ND dataset, which lacks samples between 55 and 160 mg m^{-3}, SMOGN can only generate data near 55 and 160 mg m^{-3}. Consequently, the ND-SMOGN dataset still exhibited a noticeable gap in the 80–140 mg m^{-3} range, highlighting a key limitation of the SMOGN method.

3.2. Model performance

The performances of the two bio-optical algorithms and six ML models across datasets are summarized in Table 2. The bio-optical algorithms showed the worst

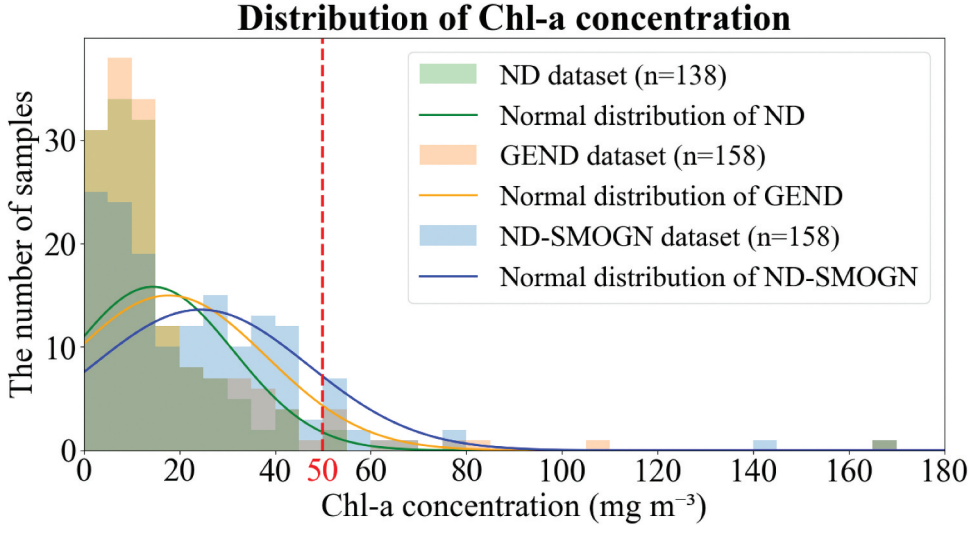

Figure 3. Distribution of three datasets—the original ND River dataset (ND), the integrated dataset of GE and ND River data (GEND), and the ND River dataset augmented using the synthetic minority oversampling technique for regression with Gaussian noise (ND-SMOGN). Solid lines indicate the fitted normal distribution.

Table 2. Summary of the model performance metrics.

Model		Evaluation metrics	ND	GEND	ND-SMOGN
Optical algorithm	2-band ratio	R^2	0.54	0.45	−0.74
		RMSE	7.01	7.65	13.58
	NDCI	R^2	0.12	0.23	−0.55
		RMSE	9.67	9.04	12.81
DT-based	RF	R^2	0.53	0.54	0.55
		RMSE	7.06	7.00	6.90
	XGB	R^2	0.47	0.59	0.36
		RMSE	7.52	6.57	8.27
NN-based	MLP	R^2	0.83	0.93	0.79
		RMSE	4.19	2.77	4.71
	CNN	R^2	0.81	0.90	0.77
		RMSE	4.48	3.21	4.97
	MLP-ATT	R^2	0.85	0.93	0.81
		RMSE	4.04	2.76	4.48
	CNN-ATT	R^2	0.84	0.91	0.80
		RMSE	4.16	3.14	4.60

performance, with the highest R^2 of only 0.54. DT-based models performed slightly better but still exhibited significantly lower performance, with the highest R^2 reaching only 0.59. In contrast, NN-based models demonstrated the best performance across the three datasets, achieving satisfactory R^2 values of at least 0.77. Model performance appeared to decrease with lower model complexity, as evidenced by the low Pearson correlation between the input variables (B01–B08A) and the Chl-a target variable across all three datasets (Figure S2). Notably, none of the spectral bands (B01–B08A) showed a correlation greater than 0.5 with the Chl-a target variable across all datasets. This highlights the inherent challenge of extracting sufficient information on Chl-a from the input variables using simple model structures, emphasizing the need for models capable of capturing more complex patterns.

This trend suggests that ML models capable of capturing more complex nonlinear relationships generally perform better. The MLP-ATT consistently demonstrated the highest performance across all three datasets. Multi-head attention improved predictions by effectively capturing inter-feature dependencies, leading to enhanced model performance across the integrated and augmented datasets (Liang et al. 2024; Tan et al. 2023). Our results also underscore the importance of selecting an appropriate model structure based on the distribution of data and the relationships between input and target variables.

Among the NN-based models, the MLP and MLP-ATT models consistently outperformed the CNN and CNN-ATT models. This result can be attributed to the spectral characteristics of Sentinel-2 data, which provides multispectral data with broad spectral bands and relatively large intervals between wavelengths (Drusch et al. 2012, Nazeer et al. 2021). These large intervals between each band can introduce irrelevant information into feature maps. While bands such as B03–B05 are crucial for Chl-a estimation due to their optical properties, the 1-D convolution kernels in CNN models aggregate data across multiple bands. This aggregation can inadvertently mix less relevant bands, such as B02–B04 or B05–B07, with key spectral regions, diluting the critical signals (Yu, Zhang, and Wang 2021). Consequently, this data characteristic reduces the model's ability to focus on the most informative spectral bands, leading to lower performance compared to MLP-based models.

DT-based models are also capable of capturing complex non-linear relationships; however, they face significant challenges when learning from skewed distributions with rare high-concentration samples. DTs rely on splitting the data into subsets during training by minimizing impurity, a process that often prioritizes the majority class. As a result, minority samples, such as high-concentration Chl-a data in a skewed distribution, are frequently underrepresented in the splits, leading to suboptimal learning. This limitation makes DT-based models particularly unsuited for datasets with imbalanced distributions. Indeed, previous studies have shown that RF can outperform NN-based models when applied to normally distributed datasets, owing to their ability to effectively capture patterns in balanced data (Hong, Yoon, and Hwa Cho 2024; Zang and Ma 2012).

Models trained on the augmented datasets, GEND and ND-SMOGN, successfully addressed this underestimation issue for the sample with the highest Chl-a concentration. For the GEND dataset, which

incorporates real-world high-concentration data collected from another watershed, improvements in both R^2 and RMSE were observed on the fixed test set. In contrast, all models trained on the ND dataset consistently underestimated the test data with the highest Chl-a concentration (Figure 4). This result was attributed to the lack of sufficient high-concentration samples in the training data. Although ND-SMOGN dataset alleviated underestimation of high-concentration samples, it led to a decrease in R^2 and an increase in RMSE. Despite the selection of an appropriate model, the dataset itself can have a significant impact on predictive performance, such as the underestimation of high-concentration data points. While remote sensing data are typically not combined across watersheds due to variations in water quality and optical characteristics (Hong et al. 2022), our results suggest that integrating data from another watershed can be highly effective in addressing problems due to data limitations.

3.3. Spatial and temporal validation of developed models

Chl-a concentration maps of AD, a dataset that was not utilized for model training, were created from five

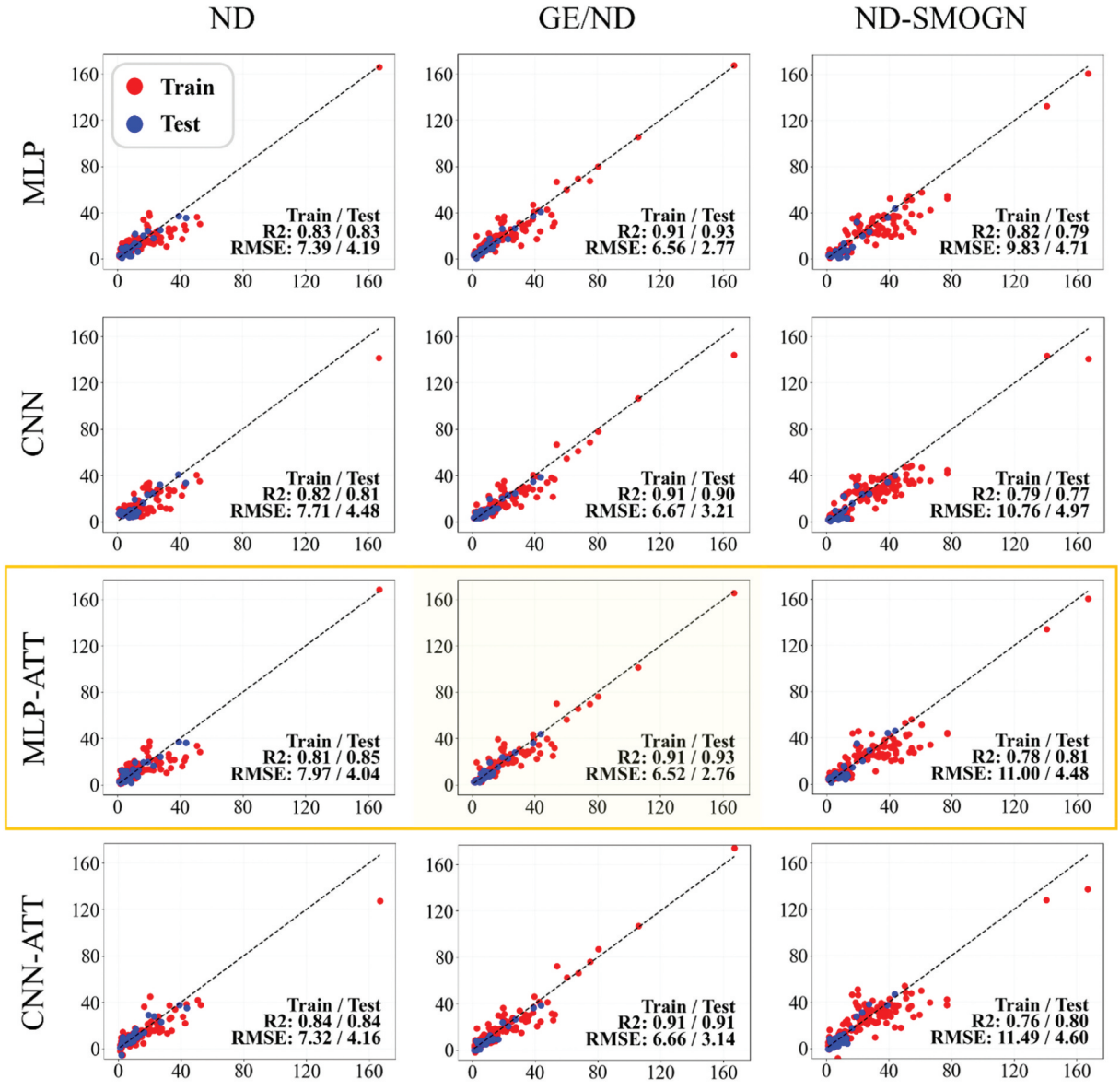

Figure 4. Scatter plots with evaluation metrics for neural network-based models: multilayer perceptron (MLP), multilayer perceptron with multi-head attention (MLP-ATT), one-dimensional convolutional neural network (CNN), and one-dimensional convolutional neural network with multi-head attention (CNN-ATT).

viewpoints based on the best models: MLP-ATT models trained on the ND, GEND, and ND-SMOGN datasets (Figure 5). Dense greenish coloration in the TCI images, indicative of potential algal blooms, was visually identified in the mid-upstream and middle regions of the lake from the third to fifth viewpoints

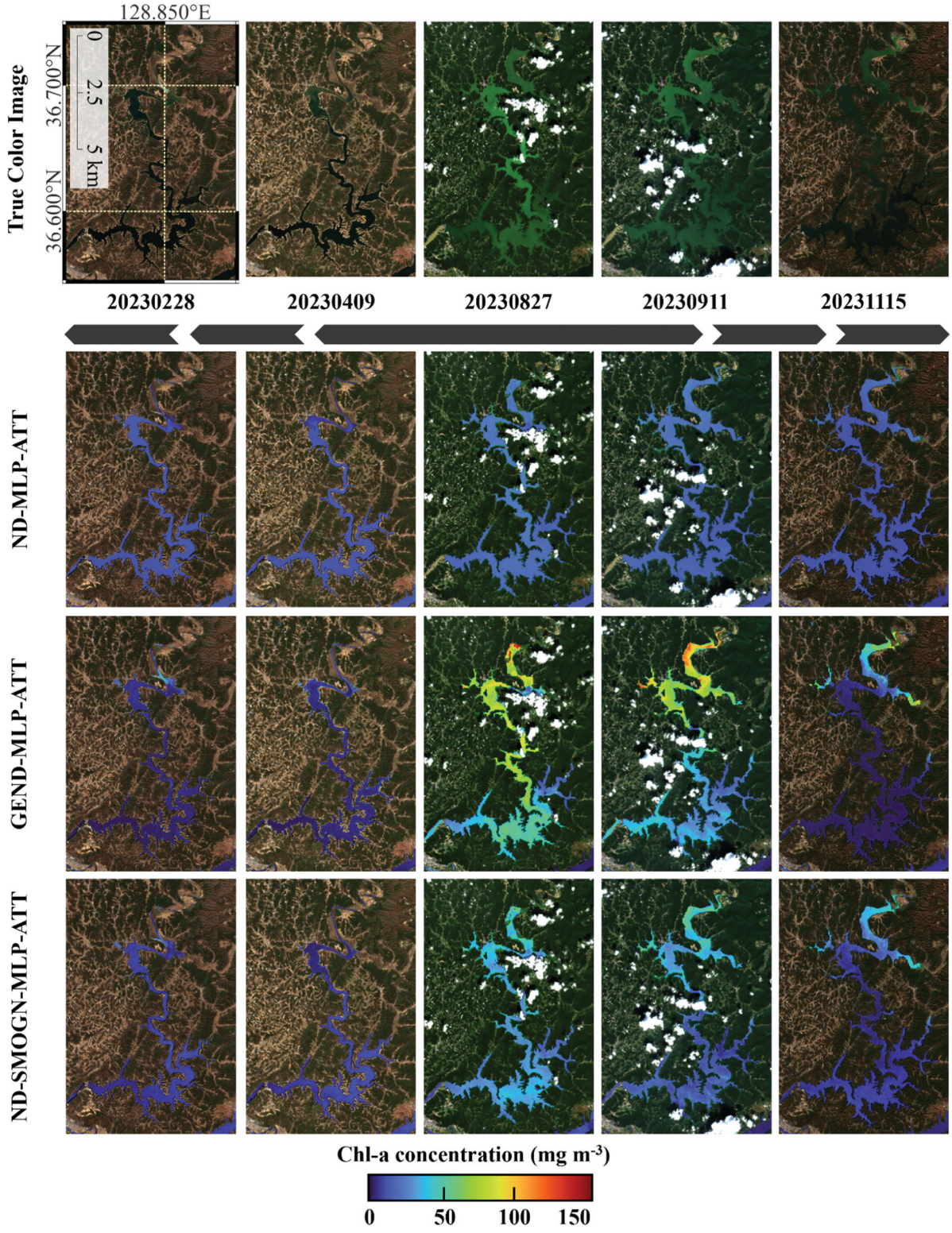

Figure 5. Generated Chl-a concentration distribution map for Andong Lake utilizing the best models: ND-MLP-ATT, GEND-MLP-ATT, and ND-SMOGN-MLP-ATT. The Chl-a concentration range was fixed at 0–150 mg m^{-3}.

(27 August 2023; 11 September 2023; and 15 November 2023, respectively). The GEND-MLP-ATT model, which exhibited the best performance, effectively captured these high-concentration Chl-a areas, aligning well with the visual cues from the TCI images. Although the other two models, ND-SMOGN-MLP-ATT and ND-MLP-ATT, similarly generated Chl-a maps with spatial patterns consistent with suspected algal bloom regions observed in TCI, they were less effective in represent high-concentration areas compared to GEND-MLP-ATT. For example, ND-MLP-ATT consistently predicted low Chl-a concentrations across all five viewpoints, with no pixels exceeding 40 mg m^{-3}. These results demonstrate the transferability of GEND-MLP-ATT, confirming its ability to generalize beyond the training regions and accurately predict Chl-a distributions in a different watershed.

Although the GEND-MLP-ATT model effectively indicated regions with potential algal blooms, it contained various sources of uncertainties. For instance, at the first viewpoint (28 February 2023), the model predicted high Chl-a concentrations in frozen upstream areas. However, algal blooms rarely occur under low-temperature winter conditions, and such cases are typically not monitored, making this uncertainty negligible. Similarly, at the third viewpoint, the model overestimated Chl-a concentrations in upstream regions. This overestimation may be attributed to a specific train data point in Figure 4, where the model overestimated Chl-a concentrations within the 50–60 mg m^{-3} range, resulting in a training bias. To mitigate this issue, acquiring additional validation and test data in this range and re-training the model could enhance its performance.

In addition to these factors, the identification of algal blooms is reported to be significantly affected by several factors, including changes in spectral information from water bodies caused by cloud cover, frozen surfaces, and shadows cast by mountains (Pi et al. 2021; Wu et al. 2020). Although uncertainties exist regarding the predictions of Chl-a at a larger scale, combining datasets from different watersheds improved the prediction accuracy and representation of high-concentrations. Moreover, submerged plants and turbidity in waterbodies can be interfering factors for extracting algal blooms (Fendereski, Creed, and Trick 2024; Klemas 2016), but other factors appear to significantly affect the model applicability regarding study sites with a depth of >20 m and suspended solids concentrations <3 mg L^{-1}. Our findings support the feasibility of integrating data from different watersheds with distinct water quality and optical properties, which could overcome the limitations of data scarcity and enhance model performance.

With the threshold of 50 mg m^{-3}, the developed models can be further utilized to generate maps indicating regions with potential algal blooms (Figure 6). Moreover, based on the pixel size, the total area of regions can be calculated, which can be applied for risk assessment and management cost estimation. Several studies suggest that estimating risk and habitat ranges of cyanobacteria would be a rapid, cost-effective approach to mitigate and manage algal blooms (King et al. 2022; Ma et al. 2021; Weber et al. 2020). Therefore, our approach has significant practical utility, as it allows for the efficient detection and assessment of bloom-prone regions. By replacing traditional methods that often involve labor-intensive and time-consuming field measurements, this framework with a remote monitoring system offers a more streamlined and effective solution for water resource management and environmental monitoring.

3.4. Importance of each band for Chl-a estimation

Important bands identified by SHAP and Gradient SHAP varied by training dataset – ND, GEND, and ND-SMOGN (Figure 7). NN-based models demonstrated the ability to prioritize spectral bands deeply linked to the optical characteristics of Chl-a, particularly the wavelengths surrounding B04 (665 nm). Among the datasets, models trained with the GEND dataset showed a distinct emphasis on B03 (443 nm) to B05 (665 nm). Chl-a exhibits optical properties characterized by high absorption in the spectral bands surrounding B04 (665 nm), a feature that has been experimentally established (Björn et al. 2009). As summarized in Table 3, B04 and B05 are widely employed in algorithms for estimating Chl-a concentrations. Thus, models that assign high importance to B03 to B05 could extract and utilize more relevant information for predicting Chl-a. While models trained on the ND and ND-SMOGN datasets included B03 to B05 as important variables, they also assigned significant importance to other bands, such as B01, B02, and B06. In contrast, models trained on the GEND dataset exhibited a sharper focus, with B03 to B05 being the

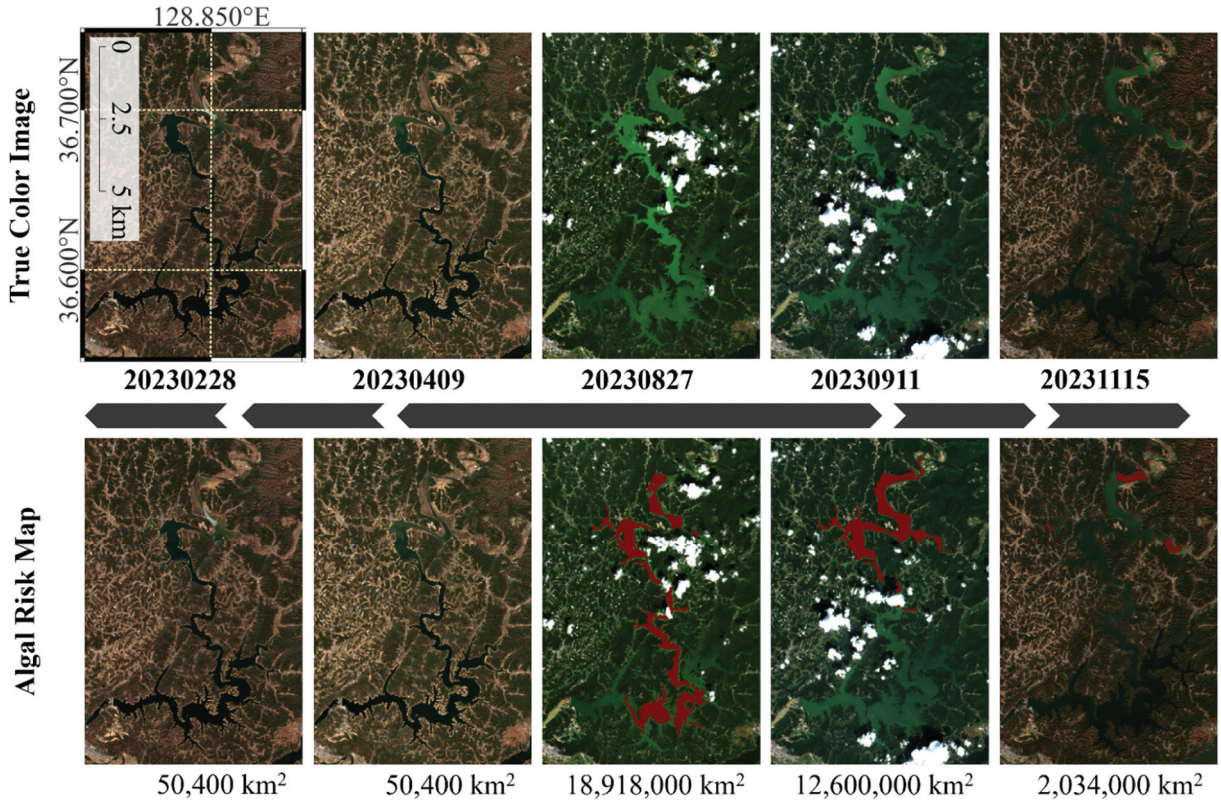

Figure 6. Maps of regions indicating potential algal blooms (> 50 mg m^{-3}) in Andong Lake. Red colored pixels indicate regions with potential algal blooms estimated by the GEND-MLP-ATT model.

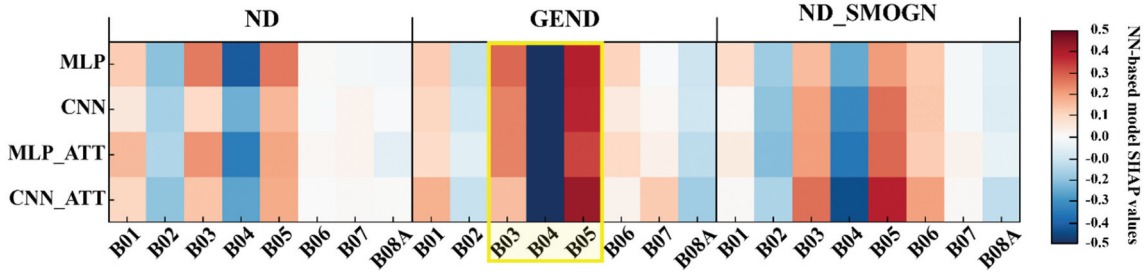

Figure 7. Variable importance analysis for the trained neural network-based models.

Table 3. Bio-optic algorithms and proposed bands for remote chlorophyll-a estimation.

Algorithm	Central wavelength of proposed bands (nm)				Reference
	λ_1	λ_2	λ_3	λ_4	
Ocean Color 4	443	490	510	560	O' Reilly et al. (2000)
Two-band ratio	443	560			Gons (1999); Gilerson et al. (2010); Ha et al. (2017)
	492	560			
	665	709			
	680	709			
	Min (660–690)	Max(690–710)			
Three-band ratio	665	709	754		Dall'olmo, Gitelson, and Rundquist (2003); Gilerson et al. (2010); Ha et al. (2017)
	680	709	754		
Four-band ratio		Band tuning required			Le et al. (2009)
Maximum Chlorophyll Index	665	709	754		Gower et al. (2005)
	680	709	754		
Normalized Difference Chlorophyll Index	665	709			Mishra and Mishra (2012)
	680	709			
Synthetic Chlorophyll Index	560	620	665	681	Shen et al. (2010)

only spectral bands with distinctly high importance. This suggests that the GEND-trained models effectively extracted and leveraged information from the spectral bands most closely associated with Chl-a concentrations.

As for DT-based models, none of the models consistently assigned high importance to all three spectral bands, B03 (560 nm) to B05 (705 nm), that are deeply associated with Chl-a concentrations (Figure S3). In addition to the previously discussed limitations of DT-based models in learning from imbalanced datasets, this inability to feasibly explain critical information from the spectral bands most relevant to Chl-a may further explain the suboptimal performance of DT-based models. These findings support that DT-based models struggle to effectively utilize the key spectral features required for accurate Chl-a prediction from multispectral reflectance data.

When combined with the results of the variable importance analysis, it becomes evident that the integration of the GE dataset effectively compensates for the lack of high-concentration samples in the ND dataset, significantly enhancing model performance. Moreover, the addition of high-concentration data enabled the models to focus more precisely on the spectral bands most closely associated with Chl-a, further improving predictive accuracy. Integrating data from different watersheds not only enhances model performance, but also provides deeper insights into how variable importance shifts with the integration of high-concentration data. These insights contribute to a better understanding of Chl-a concentration estimation and can inform the development of more effective monitoring strategies.

4. Conclusions

This study assessed the feasibility of integrating datasets from different watersheds, including remote sensing reflectance data and corresponding Chl-a concentrations, to address data scarcity challenges in ML model training. Among the six evaluated algorithms, NN-based models consistently outperformed DT-based models, with CNN-ATT achieving the highest accuracy across three datasets. The models trained with the integrated datasets demonstrated their ability to effectively generalize across diverse spatial scales, offering insights into the spectral characteristics of algal blooms. The major findings include:

- Integration of the GE and ND datasets substantially improved model performance, as evidenced by an increase in R^2 from 0.85 to 0.93 and a reduction in RMSE from 4.04 to 2.76 in the GEND-MLP-ATT model compared to the ND-MLP-ATT model.
- The SMOGN technique partially mitigated the scarcity of high-concentration samples and addressed underestimation issues. However, it reduced the number of low-concentration samples, leading to instability in low-concentration predictions and a decline in overall performance metrics.
- Model interpretation of the GEND dataset identified B03 (560 nm) to B05 (705 nm) as critical wavelengths for Chl-a estimation, aligning with established bio-optical principles.
- Models trained on the ND-SMOGN dataset assigned significant importance to less relevant bands, such as B02 (490 nm) and B06 (740 nm), in addition to B03–B05, highlighting potential limitations of synthetic data augmentation.

The integration of remote-sensing reflectance data from watersheds with distinct optical and water quality characteristics offers a promising approach to address data scarcity while maintaining the inherent variability of natural systems. By creating algal bloom risk maps with developed models, this study offers practical tools for identifying bloom-prone regions, enhancing environmental monitoring, and supporting efficient water resource management. Future research will focus on expanding the model to incorporate more diverse sources (e.g. all major river systems in South Korea) while addressing domain shift issues to enhance its generalizability.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This study was supported by the Ministry of Trade, Industry and Energy (MOTIE) and the Korea Institute for Advancement of Technology (KIAT) through the International Cooperative R&D Program [P0028541]. This study was also supported by Ministry

of Environment, under the Development of Ground Operation System for Water Resources Satellite from K-water.

Author contributions

J.L.: Conceptualization, Methodology, Software, Writing.
D.H.K. and H.J. (Ph.D.): Methodology, Software.
G.N. (Ph.D.) and E.H. (Ph.D.): Data curation, Writing-Reviewing.
J.H.K. (Ph.D.): Writing, Software.
K.H.C (Professor), and H.G.K (Ph.D.): Writing-Reviewing and Editing.
All authors read and approved the final manuscript.

Data availability statement

The data that support the findings of this study are available from the coauthor, E.H., upon reasonable request.

References

Aptoula, E., and S. Ariman. 2021. "Chlorophyll-a Retrieval from Sentinel-2 Images Using Convolutional Neural Network Regression." *IEEE Geoscience & Remote Sensing Letters* 19:1–5. https://doi.org/10.1109/LGRS.2021.3070437.

Björn, L. O., G. C. Papageorgiou, R. E. Blankenship, and Govindjee. 2009. "A Viewpoint: Why Chlorophyll A?" *Photosynthesis Research* 99 (2): 85–98. https://doi.org/10.1007/s11120-008-9395-x.

Boureau, Y.-L., J. Ponce, and Y. LeCun. 2010. "A Theoretical Analysis of Feature Pooling in Visual Recognition." *International Conference on Machine Learning (ICML)*, Haifa, Isra 10:111–118.

Branco, P., L. Torgo, and R. P. Ribeiro. 2017. "SMOGN: A Pre-Processing Approach for Imbalanced Regression." *Proceedings of the First International Workshop on Learning with Imbalanced Domains: Theory and Applications*, Skopje, Macedonia 74:36–50.

Breiman, L. 2001. "Random Forests." *Machine Learning* 45 (1): 5–32. https://doi.org/10.1023/A:1010933404324.

Burford, M. A., C. C. Carey, D. P. Hamilton, J. Huisman, H. W. Paerl, S. A. Wood, and A. Wulff. 2020. "Perspective: Advancing the Research Agenda for Improving Understanding of Cyanobacteria in a Future of Global Change." *Harmful Algae* 91:101601. https://doi.org/10.1016/j.hal.2019.04.004.

Cao, Z., R. Ma, H. Duan, N. Pahlevan, J. Melack, M. Shen, and K. Xue. 2020. "A Machine Learning Approach to Estimate Chlorophyll-a from Landsat-8 Measurements in Inland Lakes." *Remote Sensing of Environment* 248:111974. https://doi.org/10.1016/j.rse.2020.111974.

Cao, Z., R. Ma, M. Liu, H. Duan, Q. Xiao, K. Xue, and M. Shen. 2022. "Harmonized Chlorophyll-a Retrievals in Inland Lakes from Landsat-8/9 and Sentinel 2A/B Virtual Constellation Through Machine Learning." *IEEE Transactions on Geoscience & Remote Sensing* 60:1–16. https://doi.org/10.1109/TGRS.2022.3207345.

Chen, T., and C. Guestrin. 2016. "XGBoost: A Scalable Tree Boosting System." *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, California, 785–794.

Choi, H., D.-C. Koh, and Y. Yeol Yoon. 2023. "Spatial Investigation of Water Quality and Estimation of Groundwater Pollution Along the Main Stream in the Geum River Basin, Korea." *Environmental Geochemistry and Health* 45 (8): 6387–6406. https://doi.org/10.1007/s10653-023-01643-3.

Chusnah, W. N., and H.-J. Chu. 2022. "Estimating Chlorophyll-a Concentrations in Tropical Reservoirs from Band-Ratio Machine Learning Models." *Remote Sensing Applications: Society & Environment* 25:100678. https://doi.org/10.1016/j.rsase.2021.100678.

Dall'olmo, G., A. A. Gitelson, and D. C. Rundquist. 2003. "Towards a Unified Approach for Remote Estimation of Chlorophyll-A in Both Terrestrial Vegetation and Turbid Productive Waters." *Geophysical Research Letter* 30 (18). https://doi.org/10.1029/2003GL018065.

Demircioğlu, A. 2024. "The Effect of Data Resampling Methods in Radiomics." *Scientific Reports* 14 (1): 2858. https://doi.org/10.1038/s41598-024-53491-5.

Dietterich, T. G. 2002. "Ensemble Learning." *The Handbook of Brain Theory and Neural Networks* 2 (1): 110–125.

Drusch, M., U. Del Bello, S. Carlier, O. Colin, V. Fernandez, F. Gascon, B. Hoersch, C. Isola, P. Laberinti, and P. Martimort. 2012. "Sentinel-2: ESA's Optical High-Resolution Mission for GMES Operational Services." *Remote Sensing of Environment* 120:25–36. https://doi.org/10.1016/j.rse.2011.11.026.

Fendereski, F., I. F. Creed, and C. G. Trick. 2024. "Remote Sensing of Chlorophyll-a in Clear Vs. Turbid Waters in Lakes." *Remote Sensing* 16 (19): 3553. https://doi.org/10.3390/rs16193553.

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.

Gilerson, A. A., A. A. Gitelson, J. Zhou, D. Gurlin, W. Moses, I. Ioannou, and S. A. Ahmed. 2010. "Algorithms for Remote Estimation of Chlorophyll-a in Coastal and Inland Waters Using Red and Near Infrared Bands." *Optics Express* 18 (23): 24109–24125. https://doi.org/10.1364/OE.18.024109.

Gitelson, A. A., J. F. Schalles, and C. M. Hladik. 2007. "Remote Chlorophyll-a Retrieval in Turbid, Productive Estuaries: Chesapeake Bay Case Study." *Remote Sensing of Environment* 109 (4): 464–472. https://doi.org/10.1016/j.rse.2007.01.016.

Gobler, C. J. 2020. "Climate Change and Harmful Algal Blooms: Insights and Perspective." *Harmful Algae* 91:101731. https://doi.org/10.1016/j.hal.2019.101731.

Gons, H. J. 1999. "Optical Teledetection of Chlorophyll a in Turbid Inland Waters." *Environmental Science & Technology* 33 (7): 1127–1132. https://doi.org/10.1021/es9809657.

Gower, J., S. King, G. Borstad, and L. Brown. 2005. "Detection of Intense Plankton Blooms Using the 709 Nm Band of the MERIS Imaging Spectrometer." *International Journal of Remote Sensing* 26 (9): 2005–2012.

Ha, N. T. T., N. T. P. Thao, K. Koike, and M. T. Nhuan. 2017. "Selecting the Best Band Ratio to Estimate Chlorophyll-a Concentration in a Tropical Freshwater Lake Using Sentinel 2A Images from a Case Study of Lake Ba Be (Northern Vietnam)." *ISPRS International Journal of Geo-Information* 6 (9): 290. https://doi.org/10.3390/ijgi6090290.

Hong, S. M., K. Hwa Cho, S. Park, T. Kang, M. Sung Kim, G. Nam, and J. Pyo. 2022. "Estimation of Cyanobacteria Pigments in the Main Rivers of South Korea Using Spatial Attention Convolutional Neural Network with Hyperspectral Imagery." *GIScience and Remote Sensing* 59 (1): 547–567. https://doi.org/10.1080/15481603.2022.2037887.

Hong, S. M., I.-H. Yoon, and K. Hwa Cho. 2024. "Predicting the Distribution Coefficient of Cesium in Solid Phase Groups Using Machine Learning." *Chemosphere* 352:141462. https://doi.org/10.1016/j.chemosphere.2024.141462.

Jaffari, Z. H., S. Na, A. Abbas, K. Y. Park, and K. Hwa Cho. 2024. "Digital Imaging-In-Flow (FlowCAM) and Probabilistic Machine Learning to Assess the Sonolytic Disinfection of Cyanobacteria in Sewage Wastewater." *Journal of Hazardous Materials* 468:133762. https://doi.org/10.1016/j.jhazmat.2024.133762.

Jeong, H., S. Park, B. Choi, C. Seok Yu, J. Young Hong, T.-Y. Jeong, and K. Hwa Cho. 2024. "Machine Learning-Based Water Quality Prediction Using Octennial in-Situ Daphnia Magna Biological Early Warning System Data." *Journal of Hazardous Materials* 465:133196. https://doi.org/10.1016/j.jhazmat.2023.133196.

Jung, K.-Y., S. Cho, S.-Y. Hwang, Y. Lee, K. Kim, and E. Hye Na. 2020. "Identification of High-Priority Tributaries for Water Quality Management in Nakdong River Using Neural Networks and Grade Classification." *Sustainability* 12 (21): 9149.

Kiefer, I., D. Odermatt, O. Anneville, A. Wüest, and D. Bouffard. 2015. "Application of Remote Sensing for the Optimization of in-Situ Sampling for Monitoring of Phytoplankton Abundance in a Large Lake." *Science of the Total Environment* 527:493–506. https://doi.org/10.1016/j.scitotenv.2015.05.011.

Kim, H. G., Y. Cha, and K. Hwa Cho. 2024. "Projected Climate Change Impact on Cyanobacterial Bloom Phenology in Temperate Rivers Based on Temperature Dependency." *Water Research* 249:120928. https://doi.org/10.1016/j.watres.2023.120928.

Kim, H. G., S. Hong, K.-S. Jeong, D.-K. Kim, and G.-J. Joo. 2019. "Determination of Sensitive Variables Regardless of Hydrological Alteration in Artificial Neural Network Model of Chlorophyll A: Case Study of Nakdong River." *Ecological Modelling* 398:67–76. https://doi.org/10.1016/j.ecolmodel.2019.02.003.

Kim, H. G., S. Hong, D.-K. Kim, and G.-J. Joo. 2020. "Drivers Shaping Episodic and Gradual Changes in Phytoplankton Community Succession: Taxonomic versus Functional Groups." *Science of the Total Environment* 734:138940. https://doi.org/10.1016/j.scitotenv.2020.138940.

Kim, H. G., K. Hwa Cho, and F. Recknagel. 2024. "Bibliometric Network Analysis of Scientific Research on Early Warning Signals for Cyanobacterial Blooms in Lakes and Rivers." *Ecological Informatics* 102503. https://doi.org/10.1016/j.ecoinf.2024.102503.

Kim, H. G., K. Hwa Cho, and F. Recknagel. 2024. "Bibliometric Network Analysis of Scientific Research on Early Warning Signals for Cyanobacterial Blooms in Lakes and Rivers." *Ecological Informatics* 102503. https://doi.org/10.1016/j.ecoinf.2024.102503.

Kim, J. H., S. Byeon, H. Lee, D. H. Lee, M.-Y. Lee, J.-K. Shin, K. Chon, D. Seong Jeong, and Y. Park. 2024. "Deep-Learning and Data-Resampling: A Novel Approach to Predict Cyanobacterial Alert Levels in a Reservoir." *Environmental Research* 263:120135. https://doi.org/10.1016/j.envres.2024.120135.

Kim, J. H., H. Lee, S. Byeon, J.-K. Shin, D. Hoon Lee, J. Jang, K. Chon, and Y. Park. 2023. "Machine Learning-Based Early Warning Level Prediction for Cyanobacterial Blooms Using Environmental Variable Selection and Data Resampling." *Toxics* 11 (12): 955. https://doi.org/10.3390/toxics11120955.

Kim, J. H., J.-K. Shin, H. Lee, D. H. Lee, J.-H. Kang, K. Hwa Cho, Y.-G. Lee, K. Chon, S.-S. Baek, and Y. Park. 2021. "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.

King, T., S. Hundt, K. Hafen, V. Stengel, and S. Ducar. 2022. "Mapping the Probability of Freshwater Algal Blooms with Various Spectral Indices and Sources of Training Data." *Journal of Applied Remote Sensing* 16 (4): 044522–044522. https://doi.org/10.1117/1.JRS.16.044522.

Kiranyaz, S., O. Avci, O. Abdeljaber, T. Ince, M. Gabbouj, and D. J. Inman. 2021. "1D Convolutional Neural Networks and Applications: A Survey." *Mechanical Systems and Signal Processing* 151:107398. https://doi.org/10.1016/j.ymssp.2020.107398.

Klemas, V. V. 2016. *Seafloor Mapping along Continental Shelves*, edited by C. W. Finkl, and C. Makowski, 125–140. Vol. 13. 1st ed. Springer, Cham: Coastal Research Library. https://doi.org/10.1007/978-3-319-25121-9_5.

Kolluru, S., and S. Prakash Tiwari. 2022. "Modeling Ocean Surface Chlorophyll-a Concentration from Ocean Color Remote Sensing Reflectance in Global Waters Using Machine Learning." *Science of the Total Environment* 844:157191. https://doi.org/10.1016/j.scitotenv.2022.157191.

Kwon, D. H., S. Min Hong, A. Abbas, J. Pyo, H.-K. Lee, S.-S. Baek, and K. Hwa Cho. 2023. "Inland Harmful Algal Blooms (HABs) Modeling Using Internet of Things (IoT) System and Deep Learning." *Environmental Engineering Research* 28 (1). https://doi.org/10.4491/eer.2021.280.

Kwon, Y. S., J. Pyo, Y.-H. Kwon, H. Duan, K. Hwa Cho, and Y. Park. 2020. "Drone-Based Hyperspectral Remote Sensing of Cyanobacteria Using Vertical Cumulative Pigment Concentration in a Deep Reservoir." *Remote Sensing of Environment* 236:111517. https://doi.org/10.1016/j.rse.2019.111517.

Le, C., C. Hu, J. Cannizzaro, D. English, F. Muller-Karger, and Z. Lee. 2013. "Evaluation of Chlorophyll-a Remote Sensing Algorithms for an Optically Complex Estuary." *Remote Sensing of Environment* 129:75–89. https://doi.org/10.1016/j.rse.2012.11.001.

Le, C., Y. Li, Y. Zha, D. Sun, C. Huang, and H. Lu. 2009. "A Four-Band Semi-Analytical Model for Estimating Chlorophyll a in Highly Turbid Lakes: The Case of Taihu Lake, China." *Remote Sensing of Environment* 113 (6): 1175–1182.

Lee, D. Y., S. Eun Kim, and K. Oh Baek. 2023. "Modeling of Algal Fluctuations in the Reservoir According to the Opening of Yeongju Dam." *Journal of Korea Water Resources Association* 56 (3): 173–184.

Lee, H.-J., H.-K. Park, and S.-U. Cheon. 2018. "Effects of Weir Construction on Phytoplankton Assemblages and Water Quality in a Large River System." *International Journal of Environmental Research and Public Health* 15 (11): 2348. https://doi.org/10.3390/ijerph15112348.

Li, A., T. Shao, Z. Zhang, W. Fang, W. Li, J. Xu, Y. Jiang, and C. Shu. 2023. "Improvement in Spatiotemporal Chl-A Data in the South China Sea Using the Random-Forest-Based Geo-Imputation Method and Ocean Dynamics Data." *Journal of Marine Science and Engineering* 12 (1): 13. https://doi.org/10.3390/jmse12010013.

Li, N., Y. Zhang, Y. Zhang, K. Shi, H. Qian, H. Yang, Y. Niu, B. Qin, G. Zhu, and R. Iestyn Woolway. 2023. "The Unprecedented 2022 Extreme Summer Heatwaves Increased Harmful Cyanobacteria Blooms." *Science of the Total Environment* 896:165312. https://doi.org/10.1016/j.scitotenv.2023.165312.

Liang, Q., Q. Ma, H. Wu, R. Lai, Y. Zhang, P. Liu, and T. Qi. 2024. "Performance Prediction of Sintered NdFeB Magnet Using Multi-Head Attention Regression Models." *Scientific Reports* 14 (1): 28822.

Lundberg, S. M., and S.-I. Lee. 2017. "A Unified Approach to Interpreting Model Predictions." In *Neural Information Processing Systems (NeurIPS)*, 30. Long Beach, California, USA.

Ma, J., S. Jin, J. Li, Y. He, and W. Shang. 2021. "Spatio-Temporal Variations and Driving Forces of Harmful Algal Blooms in Chaohu Lake: A Multi-Source Remote Sensing Approach." *Remote Sensing* 13 (3): 427. https://doi.org/10.3390/rs13030427.

Main-Knorn, M., B. Pflug, J. Louis, V. Debaecker, U. Müller-Wilm, and F. Gascon. 2017. "Sen2Cor for Sentinel-2." *Image and Signal Processing for Remote Sensing XXIII*, 37–48. Warsaw, Poland: SPIE (The International Society for Optics and Photonics). https://doi.org/10.1117/12.2278218.

Mertens, K. N., M. Retho, S. Manach, M. Laura Zoffoli, A. Doner, M. Schapira, G. Bilien, V. Séchet, T. Lacour, and E. Robert. 2023. "An Unprecedented Bloom of Lingulodinium Polyedra on the French Atlantic Coast During Summer 2021." *Harmful Algae* 125:102426. https://doi.org/10.1016/j.hal.2023.102426.

Mishra, S., and D. R. Mishra. 2012. "Normalized Difference Chlorophyll Index: A Novel Model for Remote Estimation of Chlorophyll-a Concentration in Turbid Productive Waters." *Remote Sensing of Environment* 117:394–406. https://doi.org/10.1016/j.rse.2011.10.016.

Mobley, C. D. 1995. "The Optical Properties of Water." *Handbook of Optics* 1 (43): 43.

Moses, W. J., A. A. Gitelson, R. L. Perk, D. Gurlin, D. C. Rundquist, B. C. Leavitt, T. M. Barrow, and P. Brakhage. 2012. "Estimation of Chlorophyll-a Concentration in Turbid Productive Waters Using Airborne Hyperspectral Data." *Water Research* 46 (4): 993–1004. https://doi.org/10.1016/j.watres.2011.11.068.

Mpakairi, K. S., F. F. Muthivhi, F. Dondofema, L. F. Munyai, and T. Dalu. 2024. "Chlorophyll-a Unveiled: Unlocking Reservoir Insights Through Remote Sensing in a Subtropical Reservoir." *Environmental Monitoring and Assessment* 196 (4): 1–14. https://doi.org/10.1007/s10661-024-12554-w.

Nazeer, M., C. Olayinka Ilori, M. Bilal, J. Elizabeth Nichol, W. Wu, Z. Qiu, and B. Krishna Gayene. 2021. "Evaluation of Atmospheric Correction Methods for Low to High Resolutions Satellite Remote Sensing Data." *Atmospheric Research* 249:105308. https://doi.org/10.1016/j.atmosres.2020.105308.

Ogashawara, I., D. R. Mishra, and A. A. Gitelson. 2017. "Remote Sensing of Inland Waters: Background and Current State-Of-The-Art." In *Bio-Optical Modeling and Remote Sensing of Inland Waters*. 1st ed., 1–24. Elsevier. https://doi.org/10.1016/B978-0-12-804644-9.00001-X.

O' Reilly, J. E., S. Maritorena, D. A. Siegel, M. C. O'Brien, D. Toole, B. G. Mitchell, M. Kahru, F. P. Chavez, P. Strutton, and G. F. Cota. 2000. *SeaWiFS Postlaunch Calibration and Validation Analyses, Part 3*, 9–23. Vol. 11. Greenbelt, MD, USA: NASA Goddard Space Flight Center.

Pi, X., L. Feng, W. Li, J. Liu, X. Kuang, K. Shi, W. Qi, D. Chen, and J. Tang. 2021. "Chlorophyll-a Concentrations in 82 Large Alpine Lakes on the Tibetan Plateau During 2003–2017: Temporal–Spatial Variations and Influencing Factors." *International Journal of Digital Earth* 14 (6): 714–735.

Popescu, M.-C., V. E. Balas, L. Perescu-Popescu, and N. Mastorakis. 2009. "Multilayer Perceptron and Neural Networks." *WSEAS Transactions on Circuits and Systems* 8 (7): 579–588.

Pyo, J. C., M. Ligaray, Y. Sung Kwon, M.-H. Ahn, K. Kim, H. Lee, T. Kang, S. Been Cho, Y. Park, and K. Hwa Cho. 2018. "High-Spatial Resolution Monitoring of Phycocyanin and Chlorophyll-a Using Airborne Hyperspectral Imagery." *Remote Sensing* 10 (8): 1180. https://doi.org/10.3390/rs10081180.

Pyo, J., S. M. Hong, J. Jang, S. Park, J. Park, J. H. Noh, and K. H. Cho. 2022. "Drone-Borne Sensing of Major and Accessory Pigments in Algae Using Deep Learning Modeling." *GIScience and Remote Sensing* 59 (1): 310–332. https://doi.org/10.1080/15481603.2022.2027120.

Quinlan, J. R. 1986. "Induction of Decision Trees." *Machine Learning* 1:81–106. https://doi.org/10.1007/BF00116251.

Rice, E. W., L. Bridgewater, and American Public Health Association. 2012. *Standard Methods for the Examination of Water and Wastewater*. Vol. 10. Washington, DC.

Sammartino, M., B. Buongiorno Nardelli, S. Marullo, and R. Santoleri. 2020. "An Artificial Neural Network to Infer the Mediterranean 3D Chlorophyll-a and Temperature Fields from Remote Sensing Observations." *Remote Sensing* 12 (24): 4123. https://doi.org/10.3390/rs12244123.

Shahriari, B., K. Swersky, Z. Wang, R. P. Adams, and N. De Freitas. 2015. "Taking the Human Out of the Loop: A Review of Bayesian Optimization." *Proceedings of the IEEE* 104 (1): 148–175. https://doi.org/10.1109/JPROC.2015.2494218.

Shen, F., Y.-X. Zhou, D.-J. Li, W.-J. Zhu, and M. Suhyb Salama. 2010. "Medium Resolution Imaging Spectrometer (MERIS) Estimation of Chlorophyll-a Concentration in the Turbid Sediment-Laden Waters of the Changjiang (Yangtze) Estuary." *International Journal of Remote Sensing* 31 (17–18): 4635–4650. https://doi.org/10.1080/01431161.2010.485216.

Shim, J., S. Hong, J. Lee, S. Lee, Y. M. Kim, K. Chon, S. Park, and K. H. Cho. 2023. "Deep Learning with Data Preprocessing Methods for Water Quality Prediction in Ultrafiltration." *Journal of Cleaner Production* 428:139217. https://doi.org/10.1016/j.jclepro.2023.139217.

Shin, J., G. Lee, T. Kim, K. Hwa Cho, S. Min Hong, D. Hyuck Kwon, J. Pyo, and Y. Cha. 2024. "Deep Learning-Based Efficient Drone-Borne Sensing of Cyanobacterial Blooms Using a Clique-Based Feature Extraction Approach." *Science of the Total Environment* 912:169540. https://doi.org/10.1016/j.scitotenv.2023.169540.

Shin, J., S. Yoon, Y. Kim, T. Kim, B. Go, and Y. Cha. 2021. "Effects of Class Imbalance on Resampling and Ensemble Learning for Improved Prediction of Cyanobacteria Blooms." *Ecological Informatics* 61:101202. https://doi.org/10.1016/j.ecoinf.2020.101202.

Silveira Kupssinskü, L., T. Thomassim Guimarães, E. Menezes de Souza, D. C. Zanotta, M. Roberto Veronez, L. Gonzaga Jr, and F. F. Mauad. 2020. "A Method for Chlorophyll-a and Suspended Solids Prediction Through Remote Sensing and Machine Learning." *Sensors (Switzerland)* 20 (7): 2125. https://doi.org/10.3390/s20072125.

Sola, I., A. García-Martín, L. Sandonís-Pozo, J. Álvarez-Mozos, F. Pérez-Cabello, M. González-Audícana, and R. M. Llovería. 2018. "Assessment of Atmospheric Correction Methods for Sentinel-2 Images in Mediterranean Landscapes." *International Journal of Applied Earth Observation and Geoinformation* 73:63–76. https://doi.org/10.1016/j.jag.2018.05.020.

Soomets, T., K. Uudeberg, D. Jakovels, A. Brauns, M. Zagars, and T. Kutser. 2020. "Validation and Comparison of Water Quality Products in Baltic Lakes Using Sentinel-2 Msi and Sentinel-3 OLCI Data." *Sensors (Switzerland)* 20 (3): 742.

Tan, T.-H., Y.-L. Chang, J.-R. Wu, Y.-F. Chen, and M. Alkhaleefah. 2023. "Convolutional neural network with multihead attention for human activity recognition." *IEEE Internet of Things Journal* 11 (2): 3032–3043. https://doi.org/10.1109/JIOT.2023.3294421.

Tang, X.-J., X. Liu, P.-F. Yan, B.-X. Li, H.-Y. Qi, and F. Huang. 2022. "An MLP Network Based on Residual Learning for Rice Hyperspectral Data Classification." *IEEE Geoscience & Remote Sensing Letters* 19:1–5. https://doi.org/10.1109/LGRS.2022.3149185.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. 2017. "Attention is All You Need." In *Advances in Neural Information Processing Systems*, 30. Long Beach, California, USA.

Weber, S. J., D. R. Mishra, S. B. Wilde, and E. Kramer. 2020. "Risks for Cyanobacterial Harmful Algal Blooms Due to Land Management and Climate Interactions." *Science of the Total Environment* 703:134608. https://doi.org/10.1016/j.scitotenv.2019.134608.

World Health Organization. 2003. *Guidelines for Safe Recreational Water Environments: Coastal and Fresh Waters*. Vol. 1. Geneva, Switzerland.

Wu, R., G. Liu, R. Zhang, X. Wang, Y. Li, B. Zhang, J. Cai, and W. Xiang. 2020. "A Deep Learning Method for Mapping Glacial Lakes from the Combined Use of Synthetic-Aperture Radar and Optical Satellite Images." *Remote Sensing* 12 (24): 4020. https://doi.org/10.3390/rs12244020.

Yao, L., X. Wang, J. Zhang, X. Yu, S. Zhang, and Q. Li. 2023. "Prediction of Sea Surface Chlorophyll-a Concentrations Based on Deep Learning and Time-Series Remote Sensing Data." *Remote Sensing* 15 (18): 4486. https://doi.org/10.3390/rs15184486.

Yegnanarayana, B. 2009. *Artificial Neural Networks*. New Delhi, India: PHI Learning Pvt. Ltd.

Yu, G., W. Yang, B. Matsushita, R. Li, Y. Oyama, and T. Fukushima. 2014. "Remote Estimation of Chlorophyll-a in Inland Waters by a NIR-Red-Based Algorithm: Validation in Asian Lakes." *Remote Sensing* 6 (4): 3492–3510. https://doi.org/10.3390/rs6043492.

Yu, J., C. Zhang, and S. Wang. 2021. "Multichannel One-Dimensional Convolutional Neural Network-Based Feature Learning for Fault Diagnosis of Industrial Processes." *Neural Computing & Applications* 33 (8): 3085–3104. https://doi.org/10.1007/s00521-020-05171-4.

Zang, C., and Y. Ma. 2012. "Ensemble Learning: A Survey." In *Ensemble Machine Learning: Methods and Applications*, 1st ed., 978–1. New York, NY, USA: Springer. https://doi.org/10.1007/978-1-4419-9326-7.

Zhou, L., D. A. Roberts, W. Ma, H. Zhang, and L. Tang. 2014. "Estimation of Higher Chlorophylla Concentrations Using Field Spectral Measurement and HJ-1A Hyperspectral Satellite Data in Dianshan Lake, China." *ISPRS Journal of Photogrammetry & Remote Sensing* 88:41–47. https://doi.org/10.1016/j.isprsjprs.2013.11.016.