Korean Journal of Remote Sensing, Vol. 39, No. 5-1, 2023, pp. 655-667
https://doi.org/10.7780/kjrs.2023.39.5.1.15

ISSN 1225-6161 (Print)
ISSN 2287-9307 (Online)

Research Article

머신러닝 기반 위성영상과 수질·수문·기상 인자를 활용한 낙동강의 Chlorophyll-a 농도 추정

박소련[1] · 손상훈[2] · 배재구[3] · 이도이[3] · 서동주[4] · 김진수[5]*

Estimation of Chlorophyll-a Concentration in Nakdong River Using Machine Learning-Based Satellite Data and Water Quality, Hydrological, and Meteorological Factors

Soryeon Park[1] · Sanghun Son[2] · Jaegu Bae[3] · Doi Lee[3] · Dongju Seo[4] · Jinsoo Kim[5]*

Abstract: Algal bloom outbreaks are frequently reported around the world, and serious water pollution problems arise every year in Korea. It is necessary to protect the aquatic ecosystem through continuous management and rapid response. Many studies using satellite images are being conducted to estimate the concentration of chlorophyll-a (Chl-a), an indicator of algal bloom occurrence. However, machine learning models have recently been used because it is difficult to accurately calculate Chl-a due to the spectral characteristics and atmospheric correction errors that change depending on the water system. It is necessary to consider the factors affecting algal bloom as well as the satellite spectral index. Therefore, this study constructed a dataset by considering water quality, hydrological and meteorological factors, and sentinel-2 images in combination. Representative ensemble models random forest and extreme gradient boosting (XGBoost) were used to predict the concentration of Chl-a in eight weirs located on the Nakdong river over the past five years. R-squared score (R^2), root mean square errors (RMSE), and mean absolute errors (MAE) were used as model evaluation indicators, and it was confirmed that R^2 of XGBoost was 0.80, RMSE was 6.612, and MAE was 4.457. Shapley additive expansion analysis showed

Received October 10, 2023; Revised October 15, 2023; Accepted October 18, 2023; Published online October 31, 2023

[1] 부경대학교 지구환경시스템과학부 공간정보시스템공학전공 석사과정생(Master Student, Major of Spatial Information Engineering, Division of Earth Environmental System Science, Pukyong National University, Busan, Republic of Korea)
[2] 부경대학교 지구환경시스템과학부 공간정보시스템공학전공 박사수료생(PhD Candidate, Major of Spatial Information Engineering, Division of Earth Environmental System Science, Pukyong National University, Busan, Republic of Korea)
[3] 부경대학교 지구환경시스템과학부 공간정보시스템공학전공 박사과정생(PhD Student, Major of Spatial Information Engineering, Division of Earth Environmental System Science, Pukyong National University, Busan, Republic of Korea)
[4] 현강이엔지(주) 대표(CEO, Hyun Kang Engineering Co., Ltd., Busan, Republic of Korea)
[5] 부경대학교 지구환경시스템과학부 공간정보시스템공학전공 교수(Professor, Major of Spatial Information Engineering, Division of Earth Environmental System Science, Pukyong National University, Busan, Republic of Korea)

* Corresponding author: Jinsoo Kim (jinsookim@pknu.ac.kr)

Copyright © 2023 by The Korean Society of Remote Sensing. This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (https://creativecommons.org/licenses/by-nc/4.0/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

that water quality factors, suspended solids, biochemical oxygen demand, dissolved oxygen, and the band ratio using red edge bands were of high importance in both models. Various input data were confirmed to help improve model performance, and it seems that it can be applied to domestic and international algal bloom detection.

Keywords: Sentinel-2, Chlorophyll-a, Random forest, XGBoost, Machine learning, Algal bloom, SHAP

요약: 전 세계적으로 녹조 대발생은 빈번하게 보고되고 있으며, 국내에서도 매년 녹조로 인한 심각한 수질 오염 문제가 발생하고 있다. 지속적인 관리와 신속한 대응을 통한 수생태계 보호가 필요하다. 녹조 발생의 지표인 chlorophyll-a (Chl-a) 농도를 예측하기 위해 위성 영상을 이용한 연구가 많이 이루어지고 있다. 하지만 수계에 따라 변하는 분광특성과 대기 보정 오류로 인해 정확한 Chl-a 산출에 어려움이 있어 최근 머신러닝 모델을 활용하고 있다. 위성 분광지수 뿐만 아니라 녹조에 영향을 미치는 인자들에 대한 복합적인 고려가 필요하다. 따라서, 본 연구는 수질, 수문 및 기상 인자와 Sentinel-2 영상을 복합적으로 고려하여 데이터셋을 구축하였다. 최근 5년간 낙동강에 위치한 8개 보 구간의 Chl-a 농도 예측에 대표적인 앙상블 모델 random forest (RF)와 extreme gradient boosting (XGBoost)을 활용하였다. 모델 평가 지표로 r-squared score (R^2), root mean square errors (RMSE), mean absolute errors (MAE)를 사용하였으며, XGBoost의 R^2가 0.810, RMSE가 6.612, MAE가 4.457로 유의미한 결과를 얻은 것을 확인하였다. Shapley additive explanations (SHAP) 분석을 통해 두 모델 모두 수질 인자 suspended solids (SS), biochemical oxygen demand (BOD), dissolved oxygen (DO)과 red edge 밴드를 활용한 밴드비가 높은 중요도를 보인 것을 알 수 있었다. 다양한 입력 데이터는 모델 성능 향상에 도움을 주는 것을 확인할 수 있었으며, 국내외 녹조 탐지에 적용될 수 있을 것으로 보인다.

주요어: Sentinel-2, Chlorophyll-a, Random forest, XGBoost, 머신러닝, 녹조, SHAP

1. 서론

녹조는 수온, 영양염류, 체류시간 등 환경조건에 영향을 받아 부영양화된 수체(하천, 호소, 강 등)에서 남조류가 대량 증식하여 물색이 녹색으로 변하는 현상을 말한다(Park et al., 2018). 녹조의 대발생은 전세계에서 빈번하게 보고되고 있으며(Smetacek and Zingone, 2013), 최근 몇 년 동안 기후변화로 인한 잦은 가뭄과 마른 장마로 인해 발생 빈도와 강도가 증가하고 있다(Wang et al., 2018; Ho et al., 2019). 녹조 발생으로 인한 수질 악화는 경제적·생태적·심미적 피해 등 다양한 피해를 미치며(Pretty et al., 2003), 일부 남조류는 독소를 생성하여 인간과 동물에 건강상 피해를 유발한다(Jeon et al., 2015). 국내에서는 4대강 중 하나인 낙동강이 매년 지속적으로 녹조가 가장 많이 발생하며, 수질 악화로 인한 심각한 문제를 직면하고 있다(Lee and Kim, 2021). 선제적인 관리와 신속한 대응을 위해서는 주기적인 모니터링이 필요하다.

녹조 발생과 수질의 영양 상태를 나타내는 지표로 chlorophyll-a (Chl-a) 농도가 사용되고 있다(Boyer et al., 2009; Gregor and Maršálek, 2004). 위성영상을 기반으로 한 녹조 탐지 연구는 광범위한 지역에 대한 관측이 가능하여 효율적인 모니터링 방법으로 활용되고 있다(Zhang et al., 2019; Li et al., 2021). 많은 연구에서 녹조가 발생한 탁한 물과 깨끗한 물의 분광특성이 뚜렷하다는 점과 Chl-a의 분광특성을 고려한 밴드비 조합 알고리즘과 스펙트럼 지수가 적용되었다(Byeon et al., 2021; Saberioon et al., 2020; Ansper and Alikas, 2019). Chl-a는 Blue, Red 밴드에서 반사도가 낮고 Green, Red edge 밴드에서 반사도가 높게 나타나는 특징이 있다(Jensen, 2006). Shi et al. (2022b)은 Sentinel-2 multi spectral instrument (MSI) 영상을 기반으로 6개의 단일 밴드와 6개의 밴드 조합을 활용하여 중국의 차간호의 Chl-a 농도를 예측하였으며, Rodríguez-López et al. (2020)은 Landsat-8 operational land imager (OLI) 영상을 기반으로 11개의 식생지수를 활용하여 칠레 라자호의 Chl-a 농도를 예측하였다. 그러나 위성 기반 탐지 기법은 대기 보정으로 인한 오류와 수질에 따라 변하는 수계의 복잡한 분광특성으로 정확한 Chl-a 농도를 추정하는 것이 어렵다는 문제점이 있다(Kim and Yom, 2018).

이러한 한계를 극복하기 위해 녹조 탐지의 많은 연구에서 머신러닝 알고리즘이 활용되고 있다(Shi et al., 2022a; Kim et al., 2022, Park et al., 2021). 머신러닝 알고리즘은 구조가 간단하고 각 인자간 사전 정보가 없거나 부족한 경우에도 적용이 가능하며, 모델 구축 및 계산을 짧은 시간에 수행할 수 있다는 장점이 있다(Kim et al., 2021). Li et al. (2021)은 6개의 단일 밴드와 4개의 밴드 조합을 세 가지 모델 support vector regression (SVR), linear regression, catboost에 적용하여 중국 전역에 위치한 45개 호수들의 Chl-a 농도 예측을 수행하였다. Kim et al. (2022)은 random forest (RF), light gradient boosting machine (LGBM) 외 총 5가지 모델에 6개의 단일 밴드와 4개의 밴드 조합을 활용하여 국내 4대강 유역의 28개 하천과 호소의 Chl-a 농도 예측을 성능을 비교 및 분석하였다. Nguyen et al. (2021)은 Gaussian process regression, extreme gradient boosting (XGBoost) 외 3개의 머신러닝 모델을 비교하여 베트남 TAR 저수지의 Chl-a 농도 예측 시 입력 데이터로 수질 인자와 위성 영상을 별개로 사용하여 모델 성능을 비교하였다. 최근 머신러닝을 활용한 Chl-a 농도 예측 연구에서는 입력 변수로 수질인자 뿐만 아니라 기상, 수문 인자 등 다양한 영향인자들의 복합적 분석이 활발히 진행 중이다. Park et al. (2015)은 수질, 기상 인자를 기반으로 SVR과 artificial neural network를 활용하여 주암저수지와 영산저수지의 Chl-a 농도를 예측하였다. Lee et al. (2020)은 낙동강 중류 지역, Lee and Kim (2021)은 낙동강 하류 지역의 Chl-a 농도를 수질과 수문 인자를 기반으로 RF, decision tree, elastic net, gradient boosting을 활용하여 예측하였다.

머신러닝을 활용한 Chl-a 농도 예측 연구는 활발히 진행되고 있지만 다양한 녹조 발생 영향인자와 위성 데이터를 융합하여 머신러닝을 적용한 연구는 미흡하다. 녹조 발생 원인을 하나의 기준으로 살펴보기는 어려우므로, 보다 정확한 Chl-a 농도 예측을 위해서는 다양한 입력자료를 고려한 연구가 시도될 필요가 있다. 본 연구에서는 낙동강 수계에 위치한 8개 보의 Chl-a 농도를 대표적인 앙상블 머신러닝 알고리즘인 RF와 XGBoost를 적용하여 결과를 비교 및 분석하였다. 수질, 수문, 기상 자료와 Chl-a 농도 추정 시 중요한 Red edge 밴드를 탑재한 Sentinel-2 MSI를 활용하여 수행하였다.

2. 연구자료 및 방법

2.1. 연구 지역

본 연구에서는 Fig. 1과 같이 낙동강 본류에 위치한 8개의 보(상주보, 낙단보, 구미보, 칠곡보, 강정고령보, 달

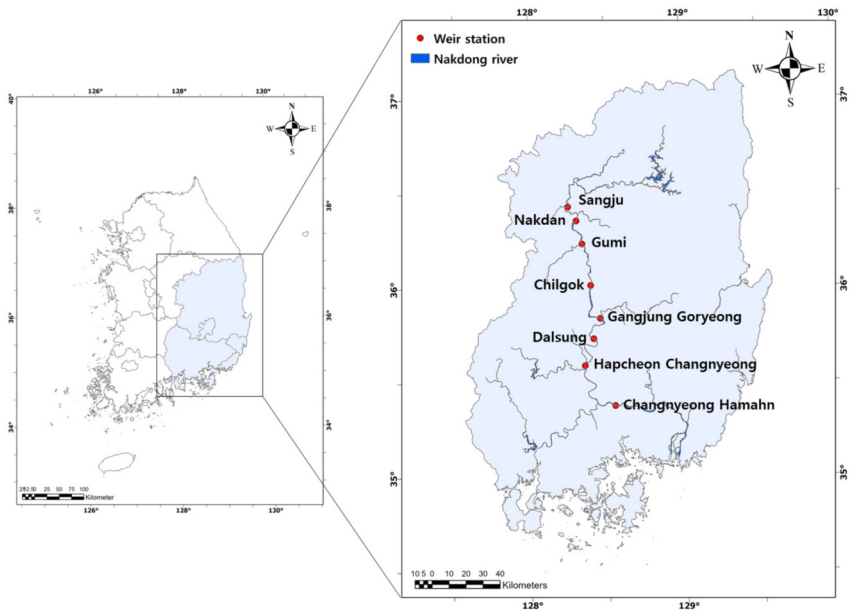


Fig. 1. Research area: 8 weirs position of Nakong river.

성보, 합천창녕보, 창녕함안보)를 대상지역으로 선정하였다. 낙동강은 2012년 다기능 보 건설로 인해 체류시간의 변화로 폐쇄성 수역의 특성을 나타내고 있으며, 녹조 발생과 같은 수질환경적 변화가 이루어지고 있다(Lee et al., 2014; Cho et al., 2018). 녹조가 자주 발생하는 여름 외의 계절 자료까지 포함하여 알고리즘의 성능을 평가하기 위해 2018년 1월부터 2022년 12월까지 지난 5년을 연구기간으로 설정하였다.

2.2. 연구 데이터

2.2.1. 위성 영상

본 연구에서는 유럽 우주국(European Space Agency)이 운용하여 Copernicus에서 제공하는 Sentinel-2 MSI 센서 자료를 사용하였다. 2015년 6월에 발사된 Sentinel-2A 위성과 2017년 3월 발사된 Sentinel-2B 위성의 각 재방문 주기는 10일로 이루어져 있으나 두 위성을 함께 사용할 시 5일로 단축시킬 수 있다(Gascon et al., 2017). Sentinel-2 MSI 센서는 10 m, 20 m, 60 m의 공간해상도 데이터를 제공하며, 가시광선 (visible, VIS), 근적외(near-infrared, NIR) 및 단파적외(short-wave infrared, SWIR) 영역에 걸쳐 총 13개의 넓은 범위(443–2,190 nm)의 다중 스펙트럼 영상을 제공한다(Table 1).

Table 1. Sentinel-2 MSI band information

Sentinel-2 band	Central wavelength (nm)	Resolution (m)
Band 1 – Costal aerosol	443	60
Band 2 – Blue	490	10
Band 3 – Green	560	10
Band 4 – Red	665	10
Band 5 – Red edge	704	20
Band 6 – Red edge	740	20
Band 7 – Red edge	783	20
Band 8 – NIR	842	10
Band 8A – Red edge	865	20
Band 9 – Water vapor	945	60
Band 10 – Cirrus	1375	60
Band 11 – SWIR	1610	20
Band 12 – SWIR	2190	20

2018년 1월부터 2022년 12월까지 다운로드한 Sentinel-2 Level-1C (L1C) 영상은 Sentinel 2 Correction (Sen2Cor) 대기 보정 프로세서를 사용하여 L2A 영상으로 대기 보정 후 사용하였다. 10, 20, 60 m의 밴드 조합을 위해 10 m의 공간해상도로 리샘플링(resampling) 하였다. 높은 정확도의 Chl-a 농도 예측을 위해 현장 측정 날짜와 ±1일 이내의 구름이 없는 영상으로 수집하였다.

본 연구에서는 선행연구를 기반으로 Table 2와 같이 6개의 단일 밴드와 7개의 밴드 조합으로 총 13개의 입력변수가 선정되었다. 많은 연구에서 Red edge 밴드에서 반사도가 매우 높게 나타나는 특성을 적용한 밴드비 알고리즘이 높은 기여도를 보였으며, 이를 고려하여 다양한 입력 변수를 모델에 사용하였다(Li et al., 2021; Kim et al., 2022).

2.2.2. 수질 및 기상, 수문 자료

수질 자료는 환경부 물환경정보시스템(http://water.nier.go.kr)에서 제공하는 보 지점의 상류에서 매주 1회 측정하는 수질 측정망의 데이터를 사용하였다. 기상 자료는 기상청의 기상자료개방포털(https://data.kma.go.kr)에서 제공하는 보 지점 10 km 내외의 가장 근접한 기상관측지점의 데이터를 사용하였으며, 수문 자료는 물정보포털(https://www.water.or.kr)에서 제공하는 보 지점

Table 2. Sentinel-2 MSI input variables used to machine learning

Data	Index	Reference
Spectral bands	Rrs (443), Rrs (492), Rrs (560), Rrs (665), Rrs (704), Rrs (740)	
Two-band ratio	Rrs (492) / Rrs (560)	Moses et al. (2009)
	Rrs (560) / Rrs (665)	Ha et al. (2017)
	Rrs (704) / Rrs (665)	Gurlin et al. (2011)
	Rrs (740) / Rrs (665)	Gitelson et al. (2008)
	Rrs (740) / Rrs (704)	Li et al. (2021)
	[Rrs (704) – Rrs (665)] / [Rrs (704) + Rrs (665)]	Gitelson et al. (2008)
Three-band ratio	$[Rrs(665)^{-1} - Rrs(704)^{-1}] \times Rrs(740)$	Gitelson et al. (2011)

Table 3. Water quality, meteorological, hydrology input variables used to machine learning

Category	Variable name	Variable description	Resolution
Water quality	Chl-a	Chlorophyll-a (mg/m^3)	Weekly
	PH	Potential of hydrogen	
	DO	Dissolved oxygen (mg/L)	
	BOD	Biochemical oxygen demand (mg/L)	
	COD	Chemical oxygen demand (mg/L)	
	Temp	Water temperature (°C)	
	Cell	Cyanobacteria cells (cells/ml)	
	TN	Total nitrogen (mg/L)	
	TP	Total phosphorus (mg/L)	
	TOC	Total organic carbon (mg/L)	
	Conductivity	Electric conductivity (μS/cm)	
	NH_3-N	Ammonia nitrogen (mg/L)	
	NO_3-N	Nitrate nitrogen (mg/L)	
	DTP	Dissolved total phosphorus (mg/L)	
	PO4-P	Phosphate phosphorus(mg/L)	
	SS	Suspended solids (mg/L)	
Meteorological	Temp_avg	1-day average of air temperature (°C)	Daily
	Temp_high	Maximum air temperature (°C)	
	Temp_low	Minimum air temperature (°C)	
	Wind_speed	1-day average of wind speed (m/s)	
Hydrological	Water level	Water level of the weir (EL.m)	
	Pondage	Water storage capacity (100,000,000 m^3)	
	Storage efficiency	Water storage rate (%)	
	Rainfall	Rainfall of weir region (mm)	
	Inflow	Inflow amount of water (m^3/s)	
	Outflow	Outflow amount of water (m^3/s)	

에 따른 수위 관측지점에서 측정된 자료를 사용하였다. 본 연구에 활용한 자료는 Table 3과 같이 Chl-a, 수소이온농도(pH), 용존 산소(DO), 생화학적 산소요구량(BOD), 화학적 산소요구량(COD), 수온(temp), 남조류 세포수(cell), 총질소(TN), 총인(TP), 총유기탄소(TOC), 전기전도도(conductivity), 암모니아성 질소(NH_3-N), 질산성 질소(NO_3-N), 용존총인(DTP), 인산염인(PO_4-P), 부유물질(SS), 평균 기온(temp_avg), 최고 기온(temp_high), 최저 기온(temp_low), 평균 풍속(wind_speed), 수위(water level), 수량(pondage), 저수율(storage efficiency), 강우량(rainfall), 유입량(inflow), 방류량(outflow) 이다. 매일 측정되는 기상, 수문 자료들은 주간 단위로 측정되는 수질 자료와 시간적 해상도를 일치시켰으며, 종속 변수인 Chl-a가 결측인 날의 자료는 제외하였다.

2.3. 연구 방법

2.3.1. Random Forest

RF는 여러 개 의사결정나무(decision tree)의 각 모델 결과에 대해 다수나 평균값을 사용하여 하나의 예측 결과를 도출하는 분류, 회귀 분석 등에 사용되는 앙상블(ensemble) 알고리즘 중 하나이다(Breiman, 2001). 부트스트랩(bootstrap) 데이터 샘플링 방법을 적용하여 생성된 데이터셋의 훈련 결과를 결합하는 배깅(bootstrap aggregation, bagging) 기법을 기초로 한다(Kim et al., 2021). 부트스트랩은 랜덤으로 데이터를 생성하고 중복을 허용하여 복원 추출을 통해 원본 데이터와 같은 크기의 데이터셋을 생성하는 방법이다. 부트스트랩을 통해 복원 추출되지 않은 데이터 셋을 out-of-bag이라고 하며, 모델 학습 결과 out-of-bag error를 통해 인자 중요도 분석이

가능하다(Nguyen et al., 2021). RF는 훈련 시 독립 변수 간의 높은 비선형성을 잘 처리하고 이상치에 강하여 수질 예측 연구에서 많이 활용되고 있다(Wang et al., 2021). 따라서 본 연구에서 RF를 예측 모델로 사용하였다.

2.3.2. XGBoost

XGBoost는 extreme gradient boosting의 약자로 기존 gradient boost 모델의 과적합문제를 개선하고 병렬 학습이 지원되도록 구조를 변형하고 알고리즘이다(Chen and Guestrin, 2016). Gradient boost는 여러 개의 약한 의사결정나무를 조합하여 강력한 예측 모델을 만드는 대표적인 앙상블 모델 중 하나이다. 순차적으로 학습하며 이전 학습 결과에 따른 가중치를 부여하여 모델의 성능을 향상시키는 boosting 기법을 기초로 한다(Cao et al., 2020). 오류를 최소화하기 위해 2차 미분 계산과 정규화(regularization)를 통해 모델의 일반화를 개선하였다(Nguyen et al., 2021). 모델의 병렬처리를 통해 기존 gradient boost 기법보다 학습 속도가 빠르며, 내부적으로 결측치 처리가 가능하다. XGBoost는 분류와 회귀 분석에 모두 사용이 가능하며 예측 정확도가 뛰어나 많은 연구에서 활용되고 있다(Cao et al., 2020; Lee et al., 2021). 이러한 장점을 기반으로 XGBoost를 본 연구에서 예측 모델로 사용하였다.

2.3.3. 자료 및 모델 구축

낙동강 8개 보 Chl-a 농도 예측을 위해 수질 인자 15개, 기상 인자 4개, 수문 인자 6개, 위성 인자 13개 총 38개의 독립 변수가 선정되었다. 총 757개의 데이터셋을 8:2의 비율로 랜덤하게 훈련 데이터(n=605), 테스트 데이터(n=152)로 나누었으며, RF와 XGBoost에 동일한 데이터셋을 적용하여 모델의 성능을 비교 및 분석하였다.

머신러닝 모델의 성능을 향상시키기 위해서는 최적의 모델 hyperparameter를 찾는 과정이 매우 중요하다(Yang and Shami, 2020). 본 연구에서는 다양한 조건의 hyperparameter 조합으로 모델의 성능을 최적화하는 grid-search cross validation 기법을 통해 최적의 hyperparameter 조합을 선정하였다.

2.3.4. 평가지표

본 연구에서는 모델 성능 비교를 위해 R^2, RMSE, MAE 3개의 평가지표를 이용하였다. R^2은 0에서 1 사이의 값을 가지며 1에 가까울수록 모델이 자료를 잘 설명한다고 할 수 있다. RMSE와 MAE는 실제 값과 예측 값의 차이를 나타내기 때문에 값이 작을수록 모델이 좋은 성능을 보인다고 할 수 있다. 평가지표의 수식은 식(1-3)과 같으며, 수식 내 $\hat{y}_i$는 예측 값, y_i는 실제 값, $\bar{y}$는 관측치 평균 값을 나타낸다.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (1)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n} | y_i - \hat{y}_i | \quad (2)$$

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(y_i - \hat{y}_i)^2}{n}} \quad (3)$$

2.3.5. SHAP

머신러닝 모델이 많은 연구에서 활용되면서 성능 비교 및 분석뿐만 아니라, 모델에 대한 해석 가능성과 설명 가능성에 대한 중요도가 높아지고 있다(Arrieta et al., 2020). 모델의 훈련과 예측 과정을 제공하여 사용자의 이해를 돕는 explainable artificial intelligence (XAI) 방법

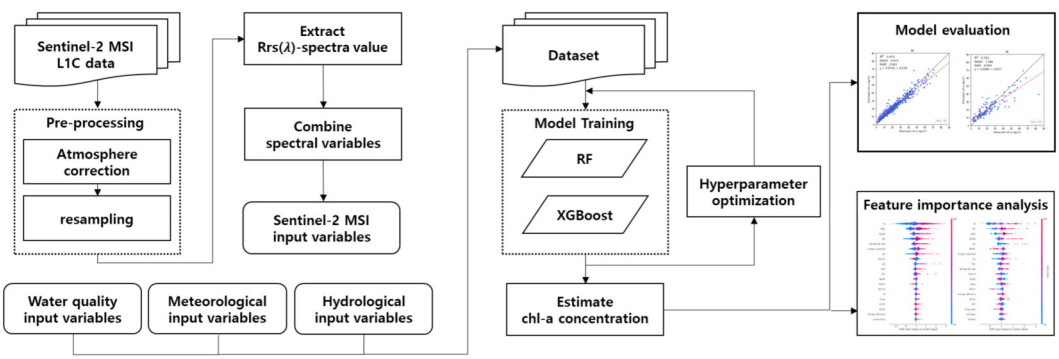


Fig. 2. Flow chart of chlorophyll-a estimation.

중 하나인 shapley additive explanations (SHAP)를 많은 연구에서 활용하고 있다(Kim et al., 2022; Shi et al., 2022b; Werther et al., 2022). SHAP는 예측 모델의 출력 결과를 게임 이론 개념 기반의 shapley value를 이용한 기법이다 (Lundberg and Lee, 2017; Shapley, 1953). SHAP은 독립 변수에 대한 기여도를 모든 데이터셋 뿐만 아니라 개별 데이터셋에 대해 평가가 가능하다. 또한 독립 변수의 개별 값이 종속 변수에 대한 상관성 분석을 음과 양으로 영향력을 제공하여 기존의 특성 중요도 기법보다 정확한 분석이 가능하다(Lee et al., 2021). 본 연구에서는 RF와 XGBoost의 결과 해석에 트리 기반의 모델에서 사용되는 Tree SHAP를 적용하여 인자 중요도 분석을 수행하였다. 연구의 전반적인 흐름은 Fig. 2와 같다.

3. 연구결과 및 토의

3.1. 모델 성능 비교

본 연구에서는 낙동강 보 8개 지점을 대상으로, 머신러닝 모델인 RF와 XGBoost을 적용하여 Chl-a 농도 예측을 수행하였다. 모델 성능 지표로 R^2, RMSE, MAE를 활용하였으며, 테스트 결과 비교는 Table 4와 같다. RF는 R^2이 0.763, MAE가 4.924 mg/m^3, RMSE가 7.386 mg/m^{3}으로 나타났으며, XGBoost는 R^2이 0.810, MAE가 4.457 mg/m^3, RMSE가 6.612 mg/m^{3}로 나타났다. 전체적으로 RF 모델보다 XGBoost 모델의 결과가 더 높게 나타났다. 두 모델 모두 전체적으로 Chl-a 농도 값을 과소 추정하는 경향을 보였으며, 특히 RF는 높은 Chl-a 농도인 80 mg/m^3 값에 대해 과소 추정했다(Fig. 3). 80 mg/m^3 이상의 높은 Chl-a 농도를 가진 데이터셋의 양이 적어 나타난 것으로 보이며, 높은 Chl-a 농도를 가진 데이터의 양을 추가하면 개선될 것으로 보인다.

Table 4. Model performance comparison with R^2, MAE, RMSE

Model	R^2	MAE (mg/m^3)	RMSE (mg/m^3)
RF	0.763	4.924	7.386
XGBoost	0.810	4.457	6.612

유사한 연구사례로 Kim et al. (2022)은 4대강 유역 78개 하천과 호소의 Chl-a 농도 예측에 5개 모델을 활용하여 6개의 단일 스펙트럼과 4개의 밴드비 알고리즘을 사용하였으며, LGBM의 R^2이 0.75, RMSE가 15.15 mg/m^3, MAE가 9.49 mg/m^{3}의 결과로 가장 좋은 성능을 보였다. 두 번째로 좋은 성능을 보인 모델은 RF이며, 100 mg/m^3 이상의 높은 Chl-a 농도를 과소 추정하는 경향이 있었다. Gradient boost 모델을 활용하여 연구한 Lee and Kim

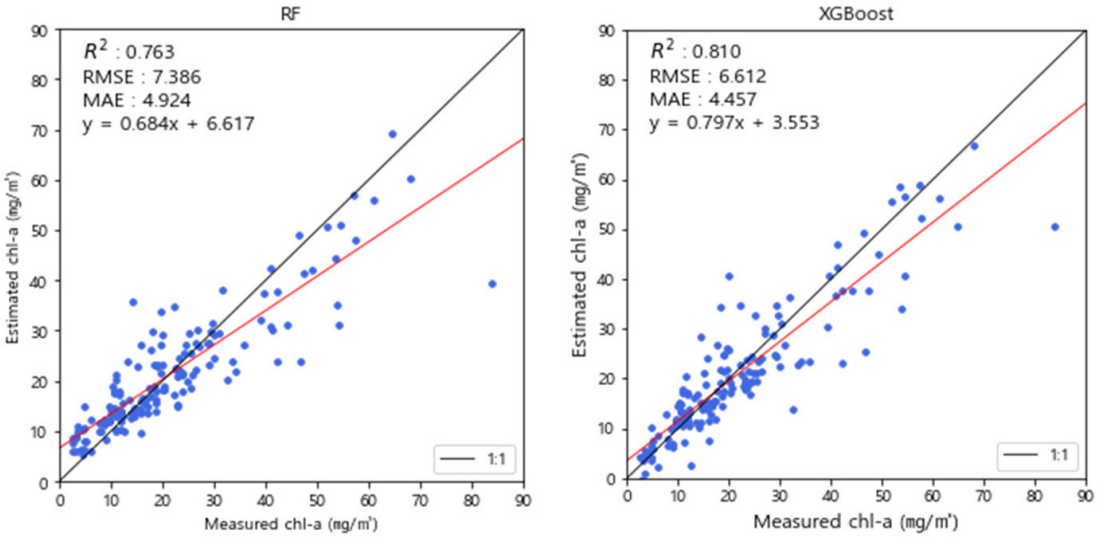


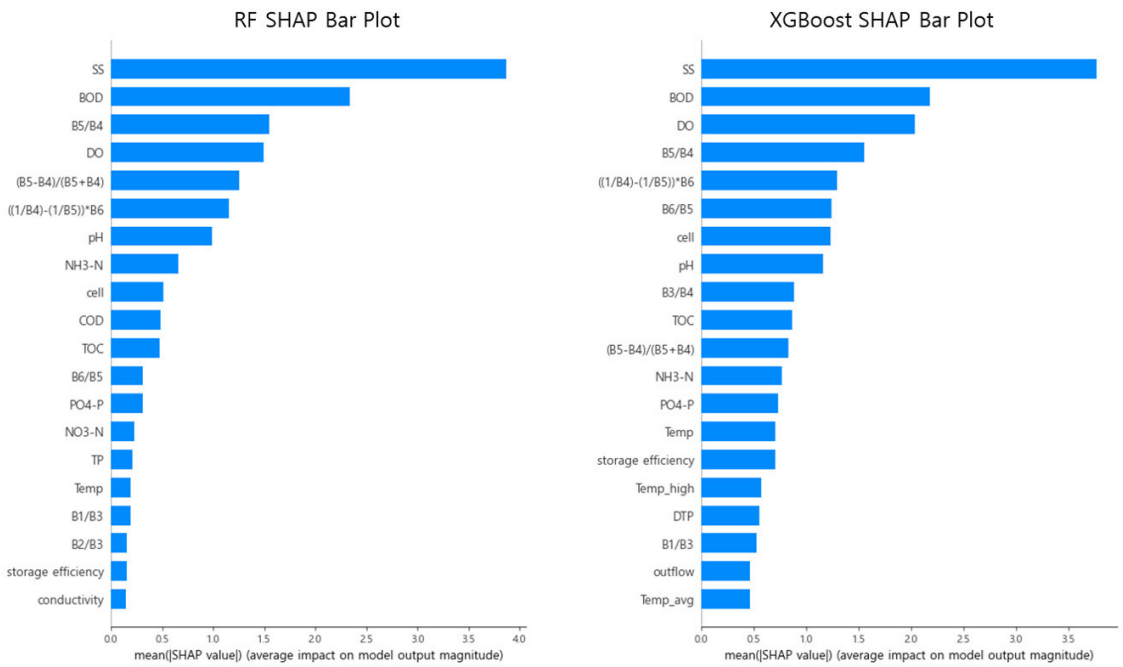
Fig. 3. The relationship between measured and estimated Chl-a by machine learning algorithms (a) RF and (b) XGBoost. The black solid is with the 1:1 line.

(2021)은 수질, 수문 인자를 입력데이터로 낙동강 하류에 위치한 합천창녕보와 창녕함안보의 Chl-a 농도 예측을 수행한 결과 R^2이 0.76–0.78로 나타났으며, RMSE는 7.51–7.99로 나타났다. 본 연구에서 비교적 좋은 정확도를 나타낸 것으로 보아 Chl-a 농도 예측에 위성 영상과 수질, 기상, 수문 인자는 복합적으로 모델 성능 향상에 영향을 미치는 것으로 보이며, 향후 다른 4대강(한강, 금강, 영산강) 유역에 대한 Chl-a 농도 예측에 적용할 수 있을 것이다.

3.2. SHAP 분석 결과

모델 예측 결과에 대한 특징 중요도 분석에 SHAP 기법을 적용하여 개별 데이터, 전체 데이터로 나누어 수행하였다. Fig. 4는 개별 데이터의 shapley value의 절대값 평균으로 Chl-a 농도 추정에 영향을 미치는 정도를 나타낸 그래프이다. 각 모델 예측에 높은 중요도를 보인 상위 20개에 대한 결과이다. 두 모델에서 공통적으로 높은 기여도를 보인 인자는 SS, BOD, DO, B5/B4, ((1/B4)−(1/B5))*B6으로 나타났다. 단일 분광밴드보다 밴드비 알고리즘이 Chl-a 예측에 더 중요한 역할을 한다는 것을 알 수 있다. Kim et al. (2022) 연구에서 모델에 가장 높은 중요도를 가진 B5/B4가 본 연구에 사용된 모델에서도 높은 성능을 보였으며, 모델의 성능에 영향을 미치는 인자에 대한 추가적인 연구를 통해 모델의 성능을 개선할 수 있을 것이다.

Fig. 5는 개별 데이터가 모델 예측에 미친 영향력을 양(+)과 음(-)으로 나타낸 그래프이다. y축은 입력변수를 x축은 shapley value를 나타내며, SHAP value가 0.0보다 작으면 예측 값을 감소시키고, 크면 예측 값을 증가시킨다는 것을 의미한다. 빨간색은 특징 중요도가 큰 값을 파란색은 특징 중요도가 작은 값을 나타낸다. RF의 경우 SS 값이 클수록 Chl-a 농도가 높아지는 경향을 가지며, NH_3-N 값이 작을수록 Chl-a 농도가 높아지는 경향을 가지는 것을 알 수 있다. XGBoost는 RF와 달리 B6/B5에서 강한 음의 상관계를 가지는 것을 확인 할 수 있었다. Chl-a와 높은 상관성을 가지는 중요한 인자로 알려진 SS는 Chl-a와 함께 수질을 모니터링하는 지표 중 하나로 사용되고 있다(Li et al., 2019). 또한, 최근 위성 영상을 활용한 많은 연구에서 SS를 탐지하여 수질 모니터링을 수행하고 있다.



(a) (b)

Fig. 4. SHAP summary bar plot for Chl-a estimating (a) RF and (b) XGBoost: B1=Rrs (443), B2=Rrs (492), B3=Rrs (560), B4=Rrs (665), B5=Rrs (704), and B6=Rrs (740).

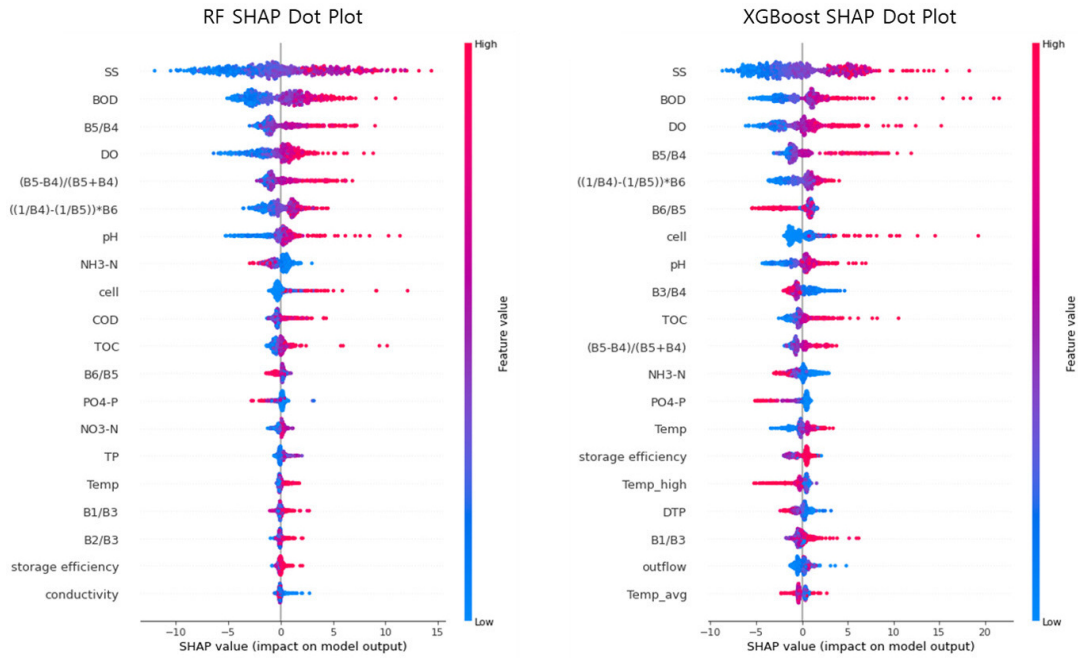


Fig. 5. SHAP summary dot plot for Chl-a estimating (a) RF and (b) XGBoost: B1=Rrs (443), B2=Rrs (492), B3=Rrs (560), B4=Rrs (665), B5=Rrs (704), and B6=Rrs (740).

4. 결론

본 연구에서는 낙동강 본류에 위치한 8개의 다기능 보 지점의 2018년부터 2022년까지 5년 간의 Chl-a 농도 예측을 수행하였다. 수질, 수문, 기상 인자 25개와 Sentinel-2의 13개의 인자 데이터를 결합하여 입력데이터로 구축하였으며, 대표적인 앙상블 알고리즘 중 RF와 XGBoost 모델의 성능을 비교 및 분석하였다. 또한, XAI 기법 중 하나인 SHAP을 이용하여 모델 예측 결과를 해석하였다. 두 모델의 Chl-a 농도 예측 테스트 결과를 살펴보면 RF의 R^2, RMSE, MAE의 값은 각각 0.763, 7.386, 4.924였으며, XGBoost의 R^2, RMSE, MAE의 값은 각각 0.810, 6.612, 4.457로 XGBoost가 상대적으로 좋은 성능을 보였다. 두 모델 모두 높은 Chl-a 농도에 대해 과소 추정하는 경향을 보였다. SHAP 분석 결과 RF는 SS, BOD, B5/B4, DO, (B5-B4)/(B5+B4)가 XGBoost는 SS, BOD, DO, B5/B4, ((1/B4)-(1/B5))*B6가 높은 중요도를 보였다. 수질 인자 SS, BOD, DO는 두 모델에서 높은 중요도를 보였으며, 위성 인자로는 높은 반사도를 가지는 B5와 낮은 반사도를 가지는 B4의 특징을 조합한 밴드비가 높은 중요도를 보였다. 본 연구를 통하여 Chl-a 농도 예측 시 수질, 기상, 수문 인자와 위성 데이터의 융합은 모델의 성능향상에 도움을 주는 것을 확인할 수 있다. 추가적인 인자 중요도 분석을 통해 최적의 입력 데이터 조합 선정이 필요할 것으로 보인다.

사사

본 연구는 과학기술정보통신부의 “초소형위성 군집 시스템의 활용지원시스템 및 활용기술개발(과제번호: 2021M1A3A4A11032019)”의 일환으로 수행되었습니다.

Conflict of Interest

No potential conflict of interest relevant to this article was reported.

References

Ansper, A., and Alikas, K., 2019. Retrieval of chlorophyll *a* from Sentinel-2 MSI data for the European Union water framework directive reporting purposes. *Remote Sensing, 11*(1), 64. https://doi.org/10.3390/rs11010064

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A. et al., 2020. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion, 58*, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Boyer, J. N., Kelble, C. R., Ortner, P. B., and Rudnick, D. T., 2009. Phytoplankton bloom status: chlorophyll *a* biomass as an indicator of water quality condition in the southern estuaries of Florida, USA. *Ecological Indicators, 9*(6), S56–S67. https://doi.org/10.1016/j.ecolind.2008.11.013

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

Byeon, Y., Seo, M., Jin, D., Jung, D., Woo, J., Jeon, U., and Han, K. S., 2021. Green algae detection in the middle·downstream of Nakdong River using high-resolution satellite data. *Korean Journal of Remote Sensing, 37*(3), 493–502. https://doi.org/10.7780/kjrs.2021.37.3.10

Cao, Z., Ma, R., Duan, H., Pahlevan, N., Melack, J., Shen, M., and Xue, K., 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

Chen, T., and Guestrin, C., 2016. XGBoost: A scalable tree boosting system. In *Proceedings of the 2016 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, CA, USA, Aug. 13–17, pp. 785–794. https://doi.org/10.1145/2939672.2939785

Cho, H., Lim, H., and Kim, S., 2018. Comparison of water quality before and after four major river project for water monitoring stations located near 8 weirs in Nakdong River. *Journal of Agriculture & Life Science, 52*(6), 89–101. https://doi.org/10.14397/jals.2018.52.6.89

Gascon, F., Bouzinac, C., Thépaut, O., Jung, M., Francesconi, B., Louis, J. et al., 2017. Copernicus Sentinel-2A calibration and products validation status. *Remote Sensing, 9*(6), 584. https://doi.org/10.3390/rs9060584

Gitelson, A. A., Dall'Olmo, G., Moses, W., Rundquist, D. C., Barrow, T., Fisher, T. R. et al., 2008. A simple semi-analytical model for remote estimation of chlorophyll-*a* in turbid waters: Validation. *Remote Sensing of Environment, 112*(9), 3582–3593. https://doi.org/10.1016/j.rse.2008.04.015

Gitelson, A. A., Gao, B. C., Li, R. R., Berdnikov, S., and Saprygin, V., 2011. Estimation of chlorophyll-a concentration in productive turbid waters using a hyperspectral imager for the coastal ocean—The Azov Sea case study. *Environmental Research Letters, 6*(2), 024023. https://doi.org/10.1088/1748-9326/6/2/024023

Gregor, J., and Maršálek, B., 2004. Freshwater phytoplankton quantification by chlorophyll a: A comparative study of in vitro, in vivo and in situ methods. *Water Research, 38*(3), 517–522. https://doi.org/10.1016/j.watres.2003.10.033

Gurlin, D., Gitelson, A. A., and Moses, W. J., 2011. Remote estimation of chl-*a* concentration in turbid productive waters — Return to a simple two-band NIR-red model?. *Remote Sensing of Environment, 115*(12), 3479–3490. https://doi.org/10.1016/j.rse.2011.08.011

Ha, N. T. T., Thao, N. T. P., Koike, K., and Nhuan, M. T., 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

Ho, J. C., Michalak, A. M., and Pahlevan, N., 2019. Widespread global increase in intense lake phytoplankton blooms since the 1980s. *Nature, 574*(7780), 667–670. https://doi.org/10.1038/s41586-019-1648-7

Jensen, J., 2006. *Remote sensing of the environment: An earth resource perspective* (2nd ed.). Pearson Education India.

Jeon, B. S., Han, J., Kim, S. K., Ahn, J. H., Oh, H. C., and Park, H. D., 2015. An overview of problems cyanotoxins produced by cyanobacteria and the solutions thereby. *Journal of Korean Society of Environmental Engineers, 37*(12), 657–667. https://doi.org/10.4491/KSEE.2015.37.12.657

Kim, D. H., and Yom, J. H., 2018. Machine learning based estimation of chlorophyll-a concentrations in the Nakdong River using satellite imagery. In *Proceedings of the 2018 Korean Society of Surveying, Geodesy, Photogrammetry and Cartography*, Seoul, Republic of Korea, Apr. 19–20, pp. 231–236.

Kim, S. H., Park, J. H., and Kim, B., 2021. Prediction of cyanobacteria harmful algal blooms in reservoir using machine learning and deep learning. *Journal of Korea Water Resources Association, 54*(12S), 1167–1181. https://doi.org/10.3741/JKWRA.2021.54.S-1.1167

Kim, Y. W., Kim, T., Shin, J., Lee, D. S., Park, Y. S., Kim, Y. et al., 2022. Validity evaluation of a machine-learning model for chlorophyll a retrieval using Sentinel-2 from inland and coastal waters. *Ecological Indicators, 137*, 108737. https://doi.org/10.1016/j.ecolind.2022.108737

Lee, J. J., Lee, Y. R., Lim, D. H., and Ahn, H. C., 2021. A study on the employee turnover prediction using XGBoost and SHAP. *The Journal of Information Systems, 30*(4), 21–42. https://doi.org/10.5859/KAIS.2021.30.4.21

Lee, S. H., Kim, B. R., and Lee, H. W., 2014. A study on water quality after construction of the weirs in the middle area in Nakdong River. *Journal of Korean Society of Environmental Engineers, 36*(4), 258–264. https://doi.org/10.4491/KSEE.2014.36.4.258

Lee, S. M., and Kim, I. K., 2021. A Comparative study on the application of boosting algorithm for Chl-a estimation in the downstream of Nakdong River. *Journal of Korean Society of Environmental Engineers, 43*(1), 66–78. https://doi.org/10.4491/KSEE.2021.43.1.66

Lee, S. M., Park, K. D., and Kim, I. K., 2020. Comparison of machine learning algorithms for Chl-a prediction in the middle of Nakdong River (focusing on water quality and quantity factors). *Journal of the Korean Society of Water and Wastewater, 34*(4), 277–288. https://doi.org/10.11001/jksww.2020.34.4.277

Li, Q., Eu, S., Lee, E. J., Lee, Y. E., Kim, M. S., and Im, S. J., 2019. Water quality analysis of in-stream and reservoir water in erosion control dams in the Nakdong River basin. *Journal of Korean Society of Forest Science, 108*(3), 329–340. https://doi.org/10.14578/jkfs.2019.108.3.329

Li, S., Song, K., Wang, S., Liu, G., Wen, Z., Shang, Y. et al., 2021. Quantification of chlorophyll-a in typical lakes across China using Sentinel-2 MSI imagery with machine learning algorithm. *Science of the Total Environment, 778*, 146271. https://doi.org/10.1016/j.scitotenv.2021.146271

Lundberg, S. M., and Lee, S. I., 2017. A unified approach to interpreting model predictions. In *Proceedings of the 2017 31st Conference on Neural Information Processing Systems*, Long Beach, CA, USA, Dec. 4–9, pp. 4768–4777.

Moses, W. J., Gitelson, A. A., Berdnikov, S., and Povazhnyy, V., 2009. Estimation of chlorophyll-a concentration in case II waters using MODIS and MERIS data—Successes and challenges. *Environmental Research Letters, 4*(4), 045005. https://doi.org/10.1088/1748-9326/4/4/045005

Nguyen, H. Q., Ha, N. T., Nguyen-Ngoc, L., and Pham, T. L., 2021. Comparing the performance of machine learning algorithms for remote and in situ estimations of chlorophyll-a content: A case study in the Tri An Reservoir, Vietnam. *Water Environment Research, 93*(12), 2941–2957. https://doi.org/10.1002/wer.1643

Park, S., Lee, S., Yun, Y., Shin, D., Park, S., and Lee, Y., 2018. Estimation of chlorophyll-a concentration for inland water using red-edge band of Sentinel-2 and RapidEye. *The Geographical Journal of Korea, 52*(3), 445–454.

Park, Y., Cho, K. H., Park, J., Cha, S. M., and 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. et al., 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

Pretty, J. N., Mason, C. F., Nedwell, D. B., Hine, R. E., Leaf, S., and Dils, R., 2003. Environmental costs of freshwater eutrophication in England and Wales. *Environmental Science & Technology, 37*(2), 201–208. https://doi.org/10.1021/es020793k

Rodríguez-López, L., Duran-Llacer, I., Gonzalez-Rodriguez, L., Abarca-del-Rio, R., Cárdenas, R., Parra, O. et al., 2020. Spectral analysis using LANDSAT images to monitor the chlorophyll-a concentration in Lake Laja in Chile. *Ecological Informatics, 60*, 101183. https://doi.org/10.1016/j.ecoinf.2020.101183

Saberioon, M., Brom, J., Nedbal, V., Souček, P., and Císař, P., 2020. Chlorophyll-a and total suspended solids retrieval and mapping using Sentinel-2A and machine learning for inland waters. *Ecological Indicators, 113*, 106236. https://doi.org/10.1016/j.ecolind.2020.106236

Shapley, L. S., 1953. *A value for n-person games*. RAND Corporation. https://doi.org/10.7249/p0295

Shi, J., Shen, Q., Yao, Y., Li, J., Chen, F., Wang, R. et al., 2022a. Estimation of chlorophyll-a concentrations in small water bodies: Comparison of fused Gaofen-6 and Sentinel-2 sensors. *Remote Sensing, 14*(1), 229. https://doi.org/10.3390/rs14010229

Shi, X., Gu, L., Jiang, T., Zheng, X., Dong, W., and Tao, Z., 2022b. Retrieval of chlorophyll-a concentrations using Sentinel-2 MSI imagery in Lake Chagan based on assessments with machine learning models. *Remote Sensing, 14*(19), 4924. https://doi.org/10.3390/rs14194924

Smetacek, V., and Zingone, A., 2013. Green and golden seaweed tides on the rise. *Nature, 504*, 84–88. https://doi.org/10.1038/nature12860

Wang, R., Kim, J. H., and Li, M. H., 2021. Predicting stream water quality under different urban development pattern scenarios with an interpretable machine learning approach. *Science of the Total Environment, 761*, 144057. https://doi.org/10.1016/j.scitotenv.2020.144057

Wang, S., Li, J., Zhang, B., Spyrakos, E., Tyler, A. N., Shen, Q. et al., 2018. Trophic state assessment of global inland waters using a MODIS-derived Forel-Ule index. *Remote Sensing of Environment, 217*, 444–460. https://doi.org/10.1016/j.rse.2018.08.026

Werther, M., Odermatt, D., Simis, S. G., Gurlin, D., Jorge, D. S., Loisel, H. et al., 2022. Characterising retrieval uncertainty of chlorophyll-*a* algorithms in oligotrophic and mesotrophic lakes and reservoirs. *ISPRS Journal of Photogrammetry and Remote Sensing, 190*, 279–300. https://doi.org/10.1016/j.isprsjprs.2022.06.015

Yang, L., and Shami, A., 2020. On hyperparameter optimization of machine learning algorithms:

Theory and practice. *Neurocomputing, 415*, 295–316. https://doi.org/10.1016/j.neucom.2020.07.061

Zhang, H., Qiu, Z., Devred, E., Sun, D., Wang, S., He, Y., and Yu, Y., 2019. A simple and effective method for monitoring floating green macroalgae blooms: A case study in the Yellow Sea. *Optics Express, 27*(4), 4528–4548. https://doi.org/10.1364/OE.27.004528