Impact of Spectral Resolution on Quantifying Cyanobacteria in Lakes and Reservoirs: A Machine-Learning Assessment

Kiana Zolfaghari, *Member, IEEE*, Nima Pahlevan, *Member, IEEE*, Caren Binding, *Member, IEEE*, Daniela Gurlin, *Member, IEEE*, Stefan G.H. Simis, *Member, IEEE*, Antonio Ruiz Verdú, *Member, IEEE*, Lin Li, *Member, IEEE*, Christopher J. Crawford, *Member, IEEE*, Andrea VanderWoude, *Member, IEEE*, Reagan Errera, *Member, IEEE*, Arthur Zastepa, *Member, IEEE*, and Claude R. Duguay, *Member, IEEE*

***Abstract*—Cyanobacterial harmful algal blooms are an increasing threat to coastal and inland waters. These blooms can be detected using optical radiometers due to the presence of phycocyanin (PC) pigments. The spectral resolution of best-available multispectral sensors limits their ability to diagnostically detect PC in the presence of other photosynthetic pigments. To assess the role of spectral resolution in the determination of PC, a large ($N = 905$) database of colocated *in situ* radiometric spectra and PC are employed. We first examine the performance of selected widely used machine-learning (ML) models against that of benchmark algorithms for hyperspectral remote sensing reflectance (R_{rs}) spectra resampled to the spectral configuration of the Hyperspectral Imager for the Coastal Ocean (HICO) with a full-width at half-maximum (FWHM) of < 6 nm. Results show that the multilayer perceptron (MLP) neural network applied to HICO spectral configurations (median errors < 65%) outperforms other ML models. This model is subsequently applied to R_{rs} spectra resampled to the band configuration of existing satellite instruments and of the one proposed for the next Landsat sensor. These results confirm that employing MLP models to estimate PC from hyperspectral data delivers tangible improvements compared with retrievals from multispectral data and benchmark algorithms (with median errors between ~73% and 126%) and shows promise for developing a globally applicable cyanobacteria measurement approach.**

***Index Terms*—Cyanobacteria harmful algal bloom (Cyano HAB), hyperspectral, machine learning (ML), neural network, phycocyanin (PC), spectral resolution.**

Manuscript received June 15, 2021; revised August 18, 2021 and September 16, 2021; accepted September 18, 2021. This work was supported in part by the Global Water Futures Program (Transformative Sensor Technologies and Smart Watersheds Project), in part by the National Aeronautics and Space Administration (NASA) Research Opportunities in Space and Earth Science (ROSES) under Grant 80NSSC20M0235 and Grant 80NSSC21K0499, and in part by the United States Geological Survey (USGS) Landsat Science Team under Award 140G0118C0011. *(Corresponding author: Kiana Zolfaghari.)*

Kiana Zolfaghari is with the Department of Geography and Environmental Management, University of Waterloo, Waterloo, ON N2L 3G1, Canada (e-mail: kzolfagh@uwaterloo.ca).

Nima Pahlevan is with the NASA Goddard Space Flight Center, Greenbelt, MD 20771 USA, and also with Science Systems and Applications Inc., Lanham, MD 20706 USA (e-mail: nima.pahlevan@nasa.gov).

Caren Binding and Arthur Zastepa are with Environment and Climate Change Canada, Canada Centre for Inland Waters, Burlington, ON L7S 1A1, Canada (e-mail: caren.binding@canada.ca; arthur.zastepa@canada.ca).

Daniela Gurlin is with the Wisconsin Department of Natural Resources, Madison, WI 53707 USA (e-mail: daniela.gurlin@wisconsin.gov).

Stefan G. H. Simis is with the Plymouth Marine Laboratory, Plymouth PL1 3DH, U.K. (e-mail: stsi@pml.ac.U.K.).

Antonio Ruiz Verdú is with the Image Processing Laboratory, University of Valencia, 46980 Valencia, Spain (e-mail: antonio.ruiz@uv.es).

Lin Li is with the Department of Earth Sciences, Indiana University–Purdue University, Indianapolis, IN 46202 USA (e-mail: ll3@iupui.edu).

Christopher J. Crawford is with the U.S. Geological Survey Earth Resources Observation and Science (EROS) Center, Sioux Falls, SD 57198 USA (e-mail: cjcrawford@usgs.gov).

Andrea VanderWoude and Reagan Errera are with the Great Lakes Environmental Research Laboratory, NOAA, Ann Arbor, MI 48108 USA (e-mail: andrea.vanderwoude@noaa.gov; reagan.errera@noaa.gov).

Claude R. Duguay is with the Department of Geography and Environmental Management, University of Waterloo, Waterloo, ON N2L 3G1, Canada, and also with H2O Geomatics Inc., Waterloo, ON N2L 1S7, Canada (e-mail: crduguay@uwaterloo.ca).

Digital Object Identifier 10.1109/TGRS.2021.3114635

I. Introduction

CYANOBACTERIAL harmful algal blooms (Cyano HABs; see Table I for a list of terms and acronyms) are a major threat to water quality and public health in coastal and inland waters [1]. Several common, bloom-forming species are able to accumulate at the water surface and produce odorous compounds, decreasing the esthetic value of the water and hampering recreational activities. The most notorious bloom-forming species exhibit strains which produce toxins that affect animals and humans [2], posing a particular risk in such surface accumulations. Cyanobacteria constitute various species which differ significantly in cell size, morphology, and toxicity [3]. Therefore, implementing standards for Cyano HAB monitoring and assessment is challenging. Furthermore, due to limited capabilities and resources available to agencies, there is no routine assessment and monitoring of Cyano HABs in many inland waters of the world. However, frequent monitoring of water quality at regional and global scales is still required to predict when and where outbreaks may occur. Thus, policymakers and water resources managers can take proactive measures to mitigate the adverse impacts of water pollution [4]. Conventional monitoring, including shore-based or ship surveys and buoy stations, is relatively costly, time-consuming, and labor intensive and requires high technical skills. More imperative is that these methods can seldom capture the spatial distribution of cyanobacteria, especially when these form patchy, often wind-driven, surface scums [5]. As a result, discrete observations often lack sufficient spatial and temporal information for decision-making. Remote sensing-based monitoring, on the other hand, has the potential

TABLE I
LIST OF TERMS AND ACRONYMS

Terms/Acronyms	Description
$a_{\text{chl}}(440)$	Chlorophyll absorption at 440 nm
AISA	Airborne Imaging Spectrometer for Applications
ANN	Artificial Neural Network
CASI-2	Compact Airborne Spectrographic Imager-2
Chl*a*	(concentration of) Chlorophyll-*a*
Cyano HAB	Cyanobacterial Harmful ‘Algal’ Bloom
CI	Cyanobacteria Index
EnMAP	Environmental Mapping and Analysis Program
FLEX	FLuorescence Explorer
HICO	Hyperspectral Imager for the Coastal Ocean
HyspIRI	Hyperspectral Infrared Imager
Landsat TM	Landsat Thematic Mapper
LNext	Landsat Next
MERIS	MEdium Resolution Imaging Spectrometer
MDN	Mixture Density Network
ML	Machine Learning
MODIS	Moderate Resolution Imaging Spectroradiometer
MPH	Maximum Peak-Height
MSI	MultiSpectral Instrument
OAC	Optically Active Constituent
OLCI	Ocean and Land Colour Instrument
OLI	Operational Land Imager
OWT	Optical Water Type
PACE	Plankton, Aerosol, Cloud, ocean Ecosystem
PLSR	Partial Least-Squares Regression
PC	(concentration of) Phycocyanin
PRISMA	PRecursore IperSpettrale della Missione Applicativa
R_{rs}	Remote Sensing Reflectance, Aquatic Reflectance
SeaWiFS	Sea-viewing Wide Field-of-view Sensor
SBG	Surface Biology and Geology
SS	Spectral Shape
SVR	Support Vector Regression
XGBoost	eXtreme Gradient Boosting

to provide a high degree of spatial and temporal resolution over extensive spatial scales and direct strategic field-based monitoring. Therefore, remote sensing monitoring is complementary to field-based monitoring, with near-real-time capabilities, for measuring Cyano HAB magnitude, extent, frequency, and duration [6]–[13].

Remote sensing techniques which rely on optical characteristics of accessory photosynthetic pigments can facilitate the detection and mapping of freshwater cyanobacteria using optical sensors [2]. Phycocyanin (PC) has been used as an indicator of cyanobacteria presence due to the distinct double optical characteristic of an absorption peak at around 620 nm and fluorescence peak at 650 nm [14]–[17]. However, different studies have relied on optical features at various wavelengths to obtain cyanobacteria abundance. Yan *et al.* [18] provide a comprehensive overview of remote sensing PC retrieval algorithms. These algorithms represent three main categories: empirical band-ratio algorithms, semianalytical algorithms, and baseline algorithms. Vincent *et al.* [19] proposed an empirical spectral-ratio approach, using a combination of Landsat Thematic Mapper (TM) visible, near-, and mid-infrared spectral bands (1, 3, 4, 5, and 7) to estimate PC in Lake Erie. The dark-object subtraction method [20] was applied on each band to reduce the effects of atmospheric haze. Therefore, this study does not include a diagnostic PC optical feature and correlations are demonstrated between cyanobacteria in abundance and the color of water. Li *et al.* [21] estimated chlorophyll-a (Chl*a*) and PC from hyperspectral airborne imaging spectrometer for applications (AISA) imagery for a mesotrophic reservoir in Central Indiana. Spectral indices derived from AISA reflectance spectra were regressed against measured pigment concentrations. The authors found the highest correlation between PC and a reflectance trough at 628 nm. Simis *et al.* [22] presented a semianalytical algorithm for retrieval of PC in turbid, cyanobacteria-dominated waters. The algorithm was suggested for application to sensors that record reflectance in the wavelengths of the PC absorption peak (around 620 nm), Chl*a* absorption (around 675 nm), and far red ($>$705 nm) and near-infrared (between 760 and 800 nm). Hunter *et al.* [23] compared the performance of analytically based algorithms (including the algorithm developed in [22]) against empirical band-ratio algorithms for retrieving PC using Compact Airborne Spectrographic Imager-2 (CASI-2) and hyperspectral AISA imagery collected over two inland waters subject to blooms of toxin-producing cyanobacteria. Their results suggested that the performance of analytically based algorithms is equal if not superior to that of more widely used empirical algorithms. Le *et al.* [24] proposed a semianalytical four-band algorithm for PC estimation in Lake Taihu using field-based hyperspectral reflectance measurements. To optimize the position of four bands, they used an iterative approach and located the first band between 615 and 630 nm, the second band around 650 nm, and the other two bands between 660 and 750 nm. Matthews *et al.* [25] proposed the maximum peak-height (MPH) algorithm for detecting Chl*a*, cyanobacteria blooms, surface scum, and floating vegetation in coastal and inland waters. This baseline subtraction procedure calculated the height of the prominent peak across the red and near-infrared medium resolution imaging spectrometer (MERIS) bands between 664 and 885 nm. Matthews and Odermatt [26] improved this algorithm for the detection of cyanobacteria in clear, oligotrophic waters. Wynne *et al.* [27] used a spectral shape (SS) approach that is based on the reflectance trough at 681 nm (SS (681), with bands located at 665, 681, and 709 nm), to derive a cyanobacteria index (CI). The top-of-atmosphere reflectance converted to Rayleigh surface reflectance (although without removing aerosol scattering) is used in the calculation of CI. Wynne *et al.* [6] used this index as a robust estimate of cyanobacteria cell counts in western Lake Erie, where blooms are primarily composed of microcystis. The CI approach is based primarily on the impact of reduced fluorescence yield of cyanobacteria rather than quantifying PC absorption, but to reduce the issue of false cyanobacteria bloom detection, SS (665) (with bands at 620, 665, and 681 nm) is used as an exclusion criterion. Mishra *et al.* [11] quantified annual and seasonal Cyano HAB biomass magnitude in Florida and Ohio (USA) lakes from

MERIS data using CI and the introduced exclusion criterion (a multiple shape algorithm).

There are many multispectral optical sensors (e.g., MODIS aboard the Aqua and Terra satellites and sensors aboard the current Landsat satellite series) whose spectral configurations cannot capture the distinct PC absorption feature at 620 nm [28]. Although these sensors might obtain Chl*a* optical features to quantify phytoplankton biomass, this pigment is prevalent in most phytoplankton species and is not specific to cyanobacteria [29]. Therefore, these multispectral sensors are suboptimal for separating waters dominated by cyanobacteria from those dominated by other phytoplankton species [30]. Kutser *et al.* [30] suggested that the MERIS band configuration allows detection of the PC absorption feature that is characteristic of waters dominated by cyanobacteria and this was confirmed by Ruiz-Verdú *et al.* [31]. Mishra *et al.* [2] discussed the effect of Chl*a* in degrading the performance of band-ratio-based cyanobacteria detection algorithms applied to multispectral sensors. Simis *et al.* [32] provided insights on the influence of other phytoplankton pigments (e.g., Chl*b*, Chl*c*, and pheophytin) on the estimation of PC, especially at low concentrations, using a band-ratio algorithm with MERIS based on [22]. There are other optically active constituents (OACs) in natural water bodies, such as chromophoric dissolved organic matter (CDOM), and inorganic suspended particles, that can also confound the algorithms due to their overlapping absorption and/or backscattering spectral signatures with PC spectral features [33]. Therefore, successful retrieval of PC from the aforementioned algorithms of multispectral satellites depends on the PC range of values and presence of other phytoplankton pigments whose signals either overlap with PC or are not well captured with current multispectral sensors.

The increasingly available aquatic remote sensing missions with enhanced spectral capabilities encourage the development of novel approaches to estimate water quality parameters such as PC. Missions include the current PRecursore IperSpettrale della Missione Applicativa (PRISMA) hyperspectral mission and the upcoming Environmental Mapping and Analysis Program (EnMAP), Plankton, Aerosol, Cloud, ocean Ecosystem (PACE), FLuorescence Explorer (FLEX), and surface biology and geology (SBG) satellites [34]–[37]. Hyperspectral data facilitate the analysis of a reflectance curve obtained in the visible and near-infrared (VNIR) in aquatic applications. This spectral curve contains valuable information on the concentration and composition of water constituents [38], [39] and can be used to enhance the retrieval of PC through differentiating its spectral signature from other OACs in optically complex waters. By employing suitable techniques, hyperspectral data can capture the discriminative optical features of PC pigments in detail and be used for quantitative mapping of cyanobacteria during bloom conditions [30]. Hyperspectral data, with the detailed spatial and spectral resolutions, and frequent temporal coverage, can complement conventional remote sensing observations [40]. Extending the use of hyperspectral datasets in monitoring different variables requires employing techniques that are able to address their collinearity and data redundancy [41]. Different parametric and nonparametric approaches have been tested in the literature to retrieve OACs using hyperspectral remote sensing observations. Popular parametric regression methods, such as partial least-squares regression (PLSR), have demonstrated adequate results for estimating OACs, such as Chl*a* from hyperspectral data in Long Bay, SC, USA [42]. Ryan and Ali [42] used PLSR to identify spectral bands that are more sensitive to Chl*a* compared with other OACs. The iterative stepwise elimination PLS (ISE-PLS), which is a combination of PLS and a wavelength selection function, outperformed other empirical and semianalytical approaches used in [43] to estimate Chl*a* from aquatic reflectance in the Seto Inland Sea, Japan. Machine-learning (ML) and nonlinear regression algorithms have also been used in the remote sensing of water quality. Nonparametric ML techniques, such as support vector regression (SVR), have been shown to capture the complex relationship between radiometric and *in situ* water quality data. Sun *et al.* [5] employed SVR to estimate PC from hyperspectral data collected in large cyanobacteria-dominated, turbid lakes in China. Pyo *et al.* [44] applied SVR, as well as a feed-forward artificial neural network (ANN), to the hyperspectral data collected from the Baekje reservoir located at the Geum River in South Korea to achieve atmospheric correction and retrieve PC and Chl*a*. Pahlevan *et al.* [45] introduced a mixture density network (MDN) to estimate Chl*a* across different bio-optical regimes in inland and coastal waters. The algorithm was applied to an *in situ* hyperspectral radiometric dataset resampled to Sentinel-2's MultiSpectral Instrument (MSI) and the Sentinel-3 Ocean and Land Color Instrument (OLCI) bands. The MDN algorithm was also adapted to retrieve hyperspectral phytoplankton absorption properties and implemented on images of the Hyperspectral Imager for the Coastal Ocean (HICO) [46]. Decision tree-based ML algorithms such as eXtreme Gradient Boosting (XGBoost) have also gained popularity in the remote sensing community. Ghatkar *et al.* [47] developed an XGBoost model for bloom onset detection and classification of its species in the Arabian Sea and Bay of Bengal waters using MODIS-Aqua data.

The primary objective of this study is to assess the impact of spectral resolution on PC estimation by employing various ML regression techniques. Heritage, existing, and planned hyperspectral and multispectral satellite missions are used to quantify their advantages and limitations as a function of spectral resolution for PC mapping. The aspects of spectral resolution under investigation include band spacing, width, and spectrum coverage. Paired field-measured PC and hyperspectral reflectance data were utilized to train and test selected ML models whose performances were compared against those of benchmark empirical methods. For this study, HICO and PRISMA spectral band configurations, with full-width at half-maximum (FWHM) of <6 nm [48] and ≤10 nm [49], respectively, represent our hyperspectral data. The spectral band settings of OLCI, MSI, operational land imager (OLI), and the proposed science measurement requirements for the future Landsat Next instrument and mission (referred to as LNext hereinafter) are the multispectral datasets. In the following sections, we provide: 1) a description of the bio-optical and limnological data collected at the study sites; 2) the development and evaluation of the performance of several ML algorithms to estimate PC from hyperspectral data resampled to HICO spectral bands; 3) an application of the top-performing ML algorithm to simulated PRISMA, OLCI, MSI, OLI, and LNext reflectance data to quantify the performance loss due to the reduced spectral capability; and 4) a comparison of the performance of selected state-of-the-art PC algorithms against the top-performing ML algorithm. The performance assessments among different band configurations and algorithms are discussed based on a subset of optical water types (OWTs), following [50] and [51].

II. Methods and Datasets

A. Study Sites

For this study, field-based measurements of hyperspectral aquatic reflectance, PC, and Chl*a* were compiled for a number of inland waters: Fremont Lakes, Indiana reservoirs, Lake Erie, South African reservoirs, Spain lakes, and Dutch lakes. The Fremont Lakes are located in the Fremont State Lakes Recreational Area, about 4.82-km west of Fremont, Nebraska, USA. The highly variable biogeochemical conditions found in the Fremont Lakes are typical of many turbid productive inland, estuarine, and coastal waters. This makes these lakes ideal candidates for the development of remote sensing algorithms to estimate PC in optically complex waters. Detailed information on the Fremont Lakes can be found in [52] and [53]. The second set of inland waters are three central Indiana (USA) reservoirs: Eagle Creek Reservoir (39°51′ N, 86°18.3′ W), Geist Reservoir (39°55′ N, 85°56.7′ W), and Morse Reservoir (40°6.4′ N, 86°2.3′ W). These are selected because of their importance in supplying drinking water to residents surrounding the Indianapolis metropolitan area and their severe eutrophication that results in toxic cyanobacteria blooms. More information about this data can be found in [54] and [55]. The third study site, Lake Erie, has persistent degraded water quality because of recurring algal blooms. It is the shallowest and most biologically productive amongst the North American Laurentian Great Lakes. The Great Lakes Water Quality Agreement led to binational efforts to reduce the phosphorus loadings into the lake in order to reduce phytoplankton biomass. However, a reduction of phosphorus has not occurred and Cyano HABs are still a persistent annual event, especially across the western basin [13]. The fourth set of inland waters are Loskop Dam (25° 25.07′ S, 29° 21.53′ E), Hartbeespoort Dam (25° 44.38′ S, 27° 51.55′ E), and Theewaterskloof Dam (34° 4.68′ S, 19° 17.35′ E); three reservoirs selected to capture the diverse bloom conditions in South African inland waters. Further details on these reservoirs can be found in [54] and [55]. The fifth study region includes 62 Spanish lakes and reservoirs distributed throughout the country, representing a large variety of trophic states and environmental conditions. More information about these study sites can be found in [31] and [32]. Finally, the sixth set of inland waters are Lake Loosdrecht (52° 11.7′ N, 5° 3.1′ E) and Lake IJsselmeer (52°45′ N, 5°20′ E) located in the Netherlands. Lake Loosdrecht, which originated from peat excavation, is a well-mixed, eutrophic, and turbid lake. Lake IJsselmeer is the largest lake in the Netherlands with an area of 1190 km^2 and a mean depth of 4.4 m. The water column in the lake is usually fully mixed but surface scums of cyanobacteria occur. Physical and biological characteristics of these lakes are described in [22] and [58]. More information about the field-based data collection methods as well as paired field-based PC, Chl*a*, and radiometric data ($N = 905$) collected in these study regions is provided below.

B. Field-Based Measurements of PC and Chlorophyll-a

In the Fremont Lakes, water samples were collected at each station with 1-L amber High Density Poly Ethylene (HDPE) bottles at a depth of 0.5 m and stored iced in the dark. Sample filtration was started on the same day of collection and used 25-mm GF/C filters to collect sufficient volumes of phytoplankton particles in conditions with low-to-moderate PC concentrations. These filters retained the relatively large-sized cyanobacteria typically found in inland waters effectively and made it possible to filter volumes of 150–500 mL of water at the same time. The filters were immediately frozen and shipped to two different laboratories on dry ice for the analysis of PC at the end of the field season. PC was extracted through repeated homogenization in a 50-mM phosphate buffer [59], [60], as detailed for the water samples from the central Indiana reservoirs, for a small selection of samples and through homogenization in a lysozyme reaction mixture [61], [62] for most of the samples. The extracts were centrifuged to clarify the samples and the supernatants were analyzed using a TD700-fluorometer or 10AU-fluorometer (Turner Designs, Inc.) depending on the laboratory. The Fremont Lakes water samples were additionally filtered through 47-mm GF/F filters and analyzed fluorometrically after extraction in ethanol [63], [64] as described in [52].

Water samples from the central Indiana reservoirs were collected using 1-L amber HDPE bottles, temporarily stored in cold and dark coolers, and filtered and frozen immediately after being transported to the laboratory before measuring PC. The measurement of PC was performed using a homogenization method with a tissue grinder [59], [60]. Samples were filtered through 0.7-μm pore size glass fiber filters (Millipore APFF). The filters were then transferred to 50-mL polycarbonate centrifuge tubes, broken up in 50-mM sodium phosphate buffer (pH 7.0 + 0.2) using a stainless-steel spatula, and subjected to two rounds of grinding and centrifuging. PC of the upper supernatant was measured using a TD700-fluorometer (Turner Designs, Inc.) which had been calibrated against PC solutions made with a Sigma-Aldrich P6161 PC standard.

Surface samples (at approximately 0.75 m) were collected from Lake Erie using a Niskin bottle sampler (General Oceanic's Model 1010) from eight monitoring sites established by NOAA's Great Lakes Environmental Research Laboratory (GLERL). Samples were stored in the dark and transported to GLERL. Upon arrival, aliquots were filtered in the dark using 47-mm GF/F filters and immediately frozen at −20 °C; volumes ranged between 50 and 400 mL. Within 24 h of collection, Chl*a* was extracted using N, N-dimethylformamide [65] and measured on a 10AU-fluorometer (Turner Designs, Inc.). PC analysis begun 24 h after collection and was extracted the following protocols established in [66]. Briefly, filters were placed in a phosphate buffer and subjected to a freeze-thaw cycle and then stored at −20 °C for at least 16 h. The following day, the samples were sonicated (Fisher FS110H) at 10 °C for 20 min and subsequently placed in the dark at 5 °C for at least 12 h. The next day, the samples were centrifuged at 4700 r/min for 20 min at 7 °C and brought to room temperature prior to conducting a reading using an Aquafluor (Turner Designs, Inc).

Part of water samples in Lake Erie were collected and processed by Environment and Climate Change Canada (ECCC) survey cruises. Water samples at each station were taken from the surface using a horizontal Van Dorn sampler and filtered on the same day of collection. For PC, 400 mL was filtered onto 47-mm GF/C filters and immediately frozen at −80 °C. In the lab, PC was analyzed according to methods described in [59]. Briefly, PC was extracted in a phosphate buffer at −4 °C for 24 h and subsequently placed at +4 °C for another 24 h. The extract was centrifuged to remove filter and cell debris and then the supernatant absorbance was measured spectrophotometrically at 455, 564, 592, and 750 nm using potassium phosphate buffer as a blank. The absorbance values were scatter-corrected by subtracting the absorbance at 750 nm. The surface water samples filtered onto 47-mm GF/C filters were also analyzed for Chl*a*, determined spectrophotometrically after extraction in acetone according to the methods of the National Laboratory for Environmental Testing [67].

Water samples in South African reservoirs were collected from the surface using 1- or 5-L opaque plastic containers. Samples were filtered through Whatman GF/F filters at low pressure on the same day of collection and Chl*a* was measured spectrophotometrically using a 90% boiling ethanol as the extraction solution, following methods in [68]. A combination of freeze-thaw and enzymatic degradation was used for PC extraction. PC measurements were performed spectrophotometrically based on [69]. Further details on steps for PC extraction and measurements are described in [57] and references therein.

For lakes in Spain and the Netherlands, water samples were taken from the surface in shallow, turbid lakes, and from the first optical depth layer in vertically stratified lakes [32]. Chl*a* samples were extracted with acetone and measured using gradient high-performance liquid chromatography (HPLC) following the protocols in [70]. Two PC extraction methods were used, including freeze-thaw based on [59] and mechanical grinding [71]. Following PC extraction, the concentrations were calculated spectrophotometrically based on [69]. More details can be found in [22] and [32].

The log distribution of PC data collected from each study site as well as the log distribution of data from all sites are shown in Fig. 1 (top). The histogram of all PC data collected from all sites nearly follows a log-normal distribution, with an overall mean and standard deviation of 58.58 and 124.11 mg/m^3, respectively. The frequency distribution and statistics of colocated Chl*a* and PC measurements are illustrated in Fig. 1 (middle). The bottom plots in Fig. 1 show the log distribution of PC to Chl*a* ratio (PC:Chl*a*). This ratio indicates the presence and abundance of cyanobacteria relative to total phytoplankton biomass [22]. Cyanobacteria is dominant when PC:Chl$a \geq 1$. PC and Chl*a* pigments are strongly correlated. Therefore, the performance of PC retrieval algorithms is assessed based on this ratio in Section III-C2.

C. In Situ Radiometry Measurements

The remote sensing reflectance, R_{rs}, is computed as the upwelling radiance emerging from the water column, L_w, divided by the total incident downwelling irradiance, $E_d(0^+)$, just above the water [72] according to (1). The depth dependency of L_w has been dropped as it is defined only at the upper side of the air–water interface [73] and the wavelength dependencies have been dropped for brevity

$$R_{\mathrm{rs}} = \frac{L_w}{E_d(0^+)}. \tag{1}$$

The *in situ* measurements of R_{rs} in the Fremont Lakes followed the method of [74]. There, a pair of intercalibrated Ocean Optics USB2000 UV-NIR spectrometers (Ocean Insight, Orlando, FL, USA) was employed to measure upwelling radiance below the water surface, $L_u(0^-)$, and $E_d(0^+)$, acquiring hyperspectral measurements from 400 to 900 nm at less than 1-nm intervals at the same time. The measurements from the two spectrometers were related through calibration scans of a white Spectralon reflectance target (Labsphere, Inc., North Sutton, NH, USA) at the start of each set of measurements and the upward radiance transmittance of the water surface was accounted based on the relationship $L_w = tL_u(0^-)n^{-2}$ [73] under the assumption of a constant upward Fresnel transmittance of the air–water interface, t, of ∼0.975 [73] and a water temperature and wavelength specific refractive index of water, n, [75] to calculate R_{rs} [52].

This closely resembled the measurements in the three central Indiana reservoirs where dual Ocean Optics USB4000 UV-NIR spectrometers (Ocean Insight) were used to measure underwater remote sensing reflectance, r_{rs}, in 2010 from 350 to 900 nm at 1-nm intervals. The measurement steps are described in [54] and [55]. Briefly, an optical fiber equipped with a cosine collector, attached to a first spectrometer, was mounted on a 2-m-high pole and pointed upward to measure the real-time incident $E_d(0^+)$. Simultaneously, a 25° field-of-view optical fiber, attached to a second radiometer, was dipped ∼2 cm below the water surface via a 2-m-long pole to measure $L_u(0^-)$ at nadir. The measurements from the two spectrometers were related through calibration scans of a gray Spectralon reflectance target (Labsphere, Inc) and the *in situ* spectra were processed in the laboratory to a pseudo underwater remote sensing reflectance, r'_{rs}, using the CALMIT Data Acquisition Program software (CDAP; University of Nebraska at Lincoln)

$$r'_{\mathrm{rs}} = \frac{L_u(0^-)}{E_d(0^+)}. \tag{2}$$

Furthermore, r_{rs}, is defined as $L_u(0^-)$ divided by the total downwelling irradiance just beneath the water surface, $E_d(0^-)$,

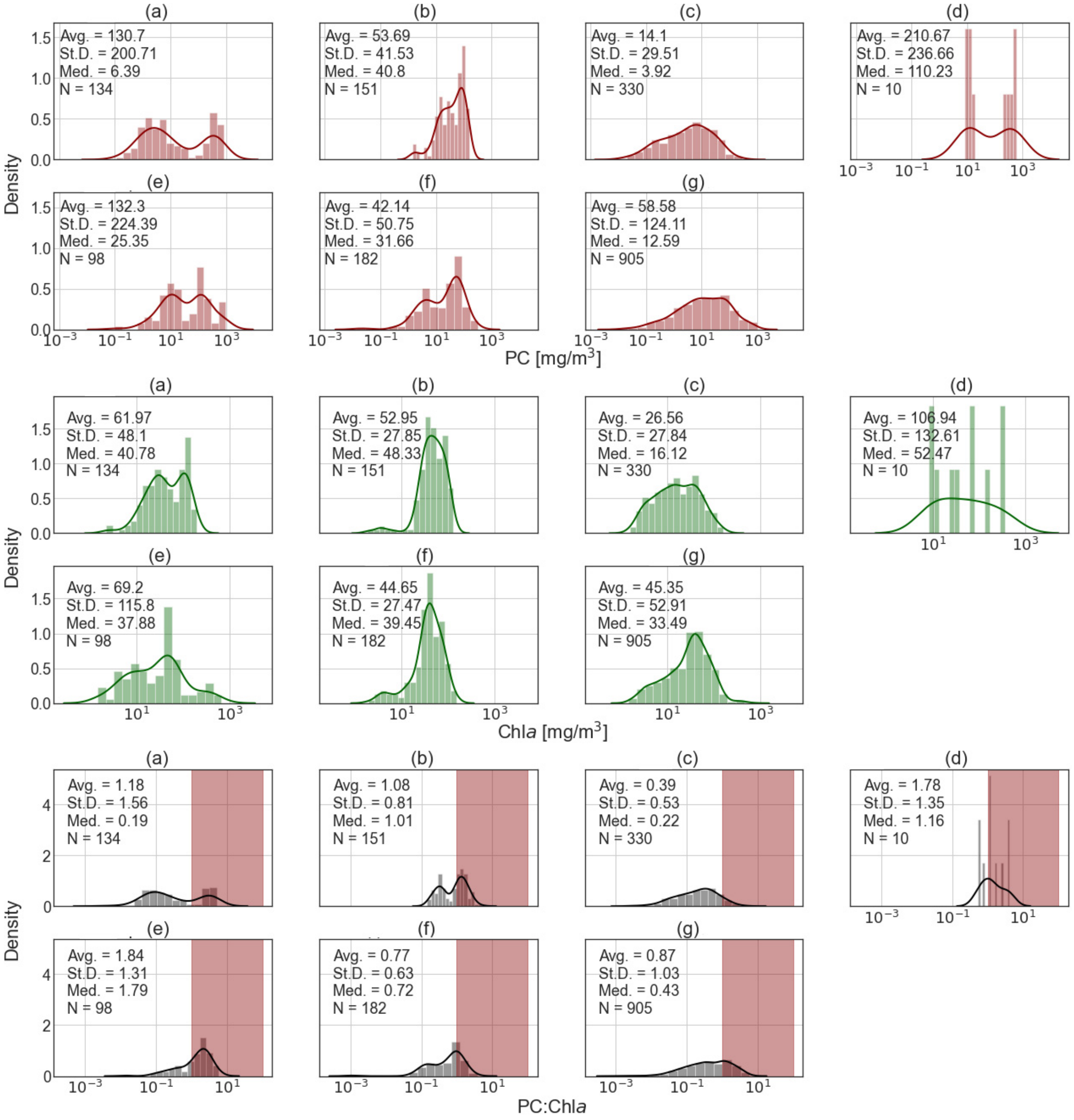

Fig. 1. Log distribution of (Top) PC, and (Middle) Chl*a* concentrations and (Bottom) PC: Chl*a* for each study site and all data combined. Statistics of PC (mg/m3), Chl*a*, (mg/m3), and PC:Chl*a* in each dataset, including average (Avg.), standard deviation (St.D.), and median (Med.), are also shown. Red bars in bottom plots show cyanobacteria-dominated samples (PC:Chl*a* $\geq$ 1). (a) Fermont lakes. (b) Indiana reservoir. (c) Lake Erie. (d) South African reservoir. (e) Spanish lakes. (f) Dutch lakes. (g) All data.

and can be computed from as below, if the assumption of $E_d(0^-) = 0.965E_d(0^+)$ [76] is made

$$r_{\text{rs}} = \frac{r'_{\text{rs}}}{0.965}. \tag{3}$$

Applying this transmittance factor is expected to introduce uncertainty as it only valid for the conditions as encountered on a number of oceanic cruises [76] and any corrective factor would have to account for the measurement geometry and light conditions encountered at the time of field data collection.

NOAA-GLERL measures surface water R_{rs} in Lake Erie during the weekly field sampling efforts, with a Satlantic Hypergun with radiance values at 137 channels (350–800 nm). The Hypergun measures upwelling radiance (L_u) at 150° relative to the solar azimuth [72], then at 40° from nadir at the water surface for 15 s and shifted 90° upward (~40° from zenith) to record sky radiance, L_{sky}, for 15 s. It is then positioned at the 18% reflective panel at 40° from nadir for 15 s. The radiance data are radiometrically calibrated and dark-offset corrected using factory calibration files, with irradiance (E_d) calculated as the radiance of the panel divided by the known reflectance of the panel (0.18) and multiplied by π [72], [73]. Water leaving radiance (L_w) was corrected for diffuse sky contamination b: $L_w = L_u - 0.028 * L_{\text{sky}}$ [72], where 0.028 is taken to be the reflectivity of the water surface. The remote sensing reflectance (R_{rs}) was calculated

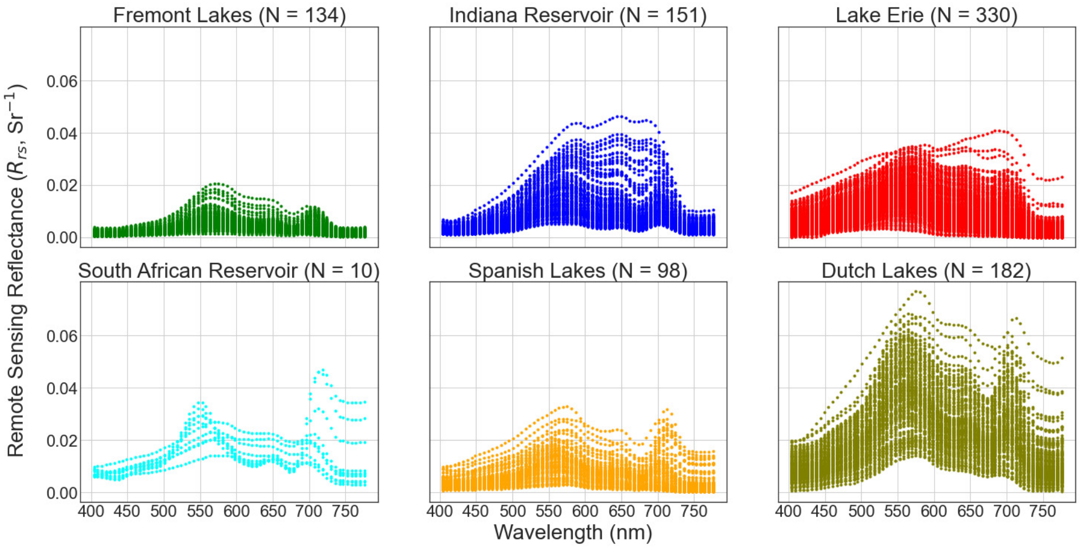

Fig. 2. Remote sensing reflectance (R_{rs}) measured at each study sites (colored by the dataset).

as L_w divided by E_d. In the NOAA-GLERL field operations, the same azimuth angle was always used which was greater than 90° between the sun and the sensor. Using a fixed bidirectional reflectance distribution function (BRDF) at this angle introduces only a 1%–2% potential error [77].

The *in situ* radiometric measurements made in Lake Erie by ECCC followed the method of [78]. Briefly, a Hyperspectral Profiler II (Seabird Scientific) was deployed to measure water column profiles of upwelling radiance, L_u(z), and downwelling irradiance, E_d(z), providing full spectrum observations from 398.9 to 803.5 nm at ~3.3-nm intervals. R_{rs} was calculated after extrapolating L_u(z) and E_d(z) to the surface and correcting for interactions at the air–water interface (z denotes the depth dependency in the acronyms). The cross-surface radiance transmittance $[L_w/L_u(0^-)]$ was assumed constant at 0.54 following [79].

The R_{rs} spectra in South African Reservoirs followed the protocols outlined in [73]. Briefly, an ASD FieldSpec spectroradiometer (ASD Inc., Boulder, CO) was used to measure radiance spectra ten times in sequence for a Spectralon target, sky, and water, from 350 to 999 nm at 1-nm intervals. The mean of radiance spectra was then calculated for each water target and R_{rs} was calculated based on equations provided in [72].

For the Spanish and Dutch lakes, an ASD-FR and a PR-650 were used to measure $L_w(0^+)$ and $E_d(0^+)$, respectively. For these lakes, R_{rs} spectra were calculated three times and the final spectra were retrieved from averaging all measurements, after removing any invalid observations. Measurements in Spanish and Dutch lakes were from 400 to 905 and 380 to 780 nm with 1- and 3–4-nm intervals, respectively. Details of the way measurements were done in each region, including the optical configurations and instrument characteristics, which are summarized in [32].

A data quality screening was carried out through visual inspections of R_{rs} data from all study regions. Outliers exhibiting abnormal spectral features, inconsistent with known spectral properties of water constituents, were excluded. The spectral resolution (and range) of R_{rs} data were different, at 1 nm (400–900 nm), 1 nm (350–900 nm), and 3 nm (348.42–802.54 nm) for data collected from the Fremont Lakes, Indiana Reservoirs, and Lake Erie, respectively. The R_{rs} data were convolved with the relative spectral responses of HICO and PRISMA to simulate their band-equivalent R_{rs} for hyperspectral analysis in ML models (Section III-C). Because of the finer spectral sampling of HICO across the VNIR, we refer to HICO-simulated R_{rs} as hyperspectral R_{rs} throughout unless otherwise noted (Sections III-A and III-B). Although HICO and PRISMA bands are within the range of 400–900 nm [48] and 400–2500 nm [49], respectively, there were no radiometric measurements beyond 802.54 nm in Lake Erie. Therefore, we considered a spectral range from 400 to 800 nm, common to all R_{rs} datasets (Fig. 2). Also, the original radiometric data were convolved with the relative spectral response of OLCI, MSI, OLI, and LNext to simulate band-equivalent R_{rs} for algorithm training and testing pertaining to multispectral data. The OLCI band at 400 nm was excluded due to inadequate radiometric coverage <400 nm in the Fremont Lakes data.

D. OWTs

In order to analyze algorithm performance over a range of OWTs, the typology developed in [50] and modified in [51] was used. Spyrakos *et al.* [50] collected a comprehensive dataset from more than 250 aquatic ecosystems, including inland waters and coastal areas, representing a wide range of optical conditions. The authors applied a functional data analysis smoothing method and k-means clustering approach on this data ($N = 4045$) to develop a typology of OWTs for natural waters. They identified 21 distinct OWTs when applying the k-means classification algorithm on inland and coastal waters. Pahlevan *et al.* [51] reduced the identified OWTs in [50] into seven types, to cover both the

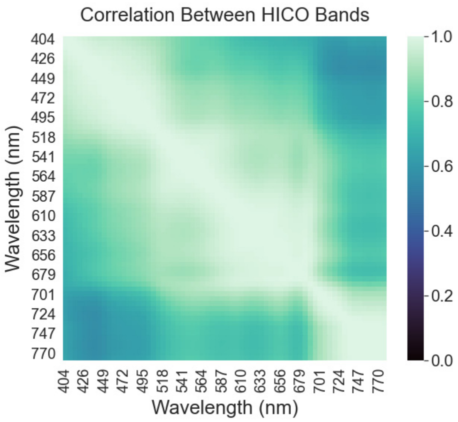

Fig. 3. Correlation matrix in hyperspectral data (not all HICO bands are labeled, due to limitation in figure size).

continuum of various OWTs encountered in aquatic environments and avoid under-representing OWTs in their matchup dataset.

To assign an HICO-like R_{rs} spectrum to one of the predefined OWTs, the R_{rs} values were first standardized by dividing it to the area under the curve. The area was calculated using numerical integration from 400 to 780 nm. This standardization approach can preserve the shape of the spectral curve across the different parts of the spectrum [80] as used in [50]. After standardization, applying the L2 norm (Euclidean) distance, the similarity of each spectrum to the associated spectrum for each OWT in [50] was calculated. Each spectrum was assigned to OWT with the closest distance.

E. ML Algorithms

Vandermeulen *et al.* [81] show that a continuous spectrum with 5–7-nm spectral sampling frequency is optimal to resolve the shape of peaks and valleys in R_{rs} for ocean color applications. The hyperspectral data in their study were collected from multiple sources to represent different optical features ranging from turbid freshwater and coastal waters to blue and oligotrophic waters. However, there is a strong correlation between observations at neighboring wavelengths in hyperspectral data (e.g., Fig. 3; correlation among HICO-resampled R_{rs} bands in this study, with ∼5.7-nm spectral resolution). Lee *et al.* [82] and Wolanin *et al.* [83] discussed the correlation in hyperspectral data and their first- and second-order derivatives using extensive and inclusive data measured and synthesized, respectively, to represent various aquatic environments. Therefore, employment of hyperspectral data for extracting OACs including PC requires techniques that can address their collinearity and high dimensionality.

Four ML regression approaches were tested in this study to retrieve PC from HICO bands: partial least squares (PLS), SVR, XGBoost, and multilayer perceptron (MLP). All the ML algorithms were adopted from Python package scikit-learn [84].

1) Overview of Selected Algorithms: PLSR: PLSR is an iterative statistical technique developed by Wold [85]. It was included in this study as a parametric regression algorithm due to its increasing popularity in remote sensing studies. Similar to principal component regression (PCR), PLSR models a response variable using new predictor variables (known as components), when there are a large number of predictors that are highly correlated or collinear. A terminating rule is employed to identify the optimal number of components. But, unlike PCR, PLSR considers the response variable when creating the components to explain the observed variability in the predictor variables [86]. This will often lead to the development of models that are able to fit the response variable with a fewer number of components [43], [87]. For the reasons summarized in [87], there are a few recognized advantages for PLSR over PCR. PLSR is considered as one of the staples of modern chemometrics [88]. This algorithm is a generalization of a multiple linear regression (MLR) approach that, unlike MLR, enables the analysis of data with numerous strongly collinear and noisy predictor variables. In situations where the input features have large dimensionality and are collinear, MLR can often overfit, which commonly occurs when applying regressing techniques to hyperspectral data. PLSR offers feature selection procedures to overcome the overfitting problem [89]–[91].

Robertson *et al.* [92] tested the performance of PLS for estimating cyanobacterial pigments in eutrophic inland waters, and PLS was later jointly used with genetic algorithm (GA) to quantify Chl*a* and PC with *in situ* measured spectra [93], [94] leading to improved accuracies compared with three band models particularly for estimating Chl*a* [95]. PLS was also coupled with ANN to model possible nonlinear relationships between cyanobacterial pigments Chl*a* and PC and spectral reflectance [96], [97].

SVR: Cortes and Vapnik [98] first identified support vector machines (SVMs). In the context of SVM, SVR was presented by Drucker *et al.* [99]. This ML algorithm is popular in remote sensing studies to robustly capture the nonlinear trends in data. SVR relies on kernel functions and thus is considered as a nonparametric approach. The kernel functions, such as the linear, polynomial, sigmoid, and radial basis functions (RBFs), are used to transform the nonlinear regression in the original feature space into a linear regression. Kwiatkowska and Fargion [100] used SVR to cross-calibrate two satellite ocean color sensors (MODIS and SeaWiFS). The objective of the research was to eliminate the inconsistencies between the corresponding data products and produce merged daily global ocean color coverage. Ruescas *et al.* [101] tested five different ML algorithms including SVR for the retrieval of colored dissolved organic matter from simulated MSI- and OLCI-R_{rs} data.

XGBoost: The XGBoost algorithm was proposed by Chen and He [102]. This algorithm is nonparametric and is a tree-based ensemble algorithm. It originates from the idea of “boosting” by integrating predictions from “weak” learners to develop a “strong” learner via an additive training process [103]. The XGBoost algorithm aims to reduce computational time and avoid the overfitting issue by introducing regularization parameters. The collinearity of input features does not affect the accuracy and prediction performance of the model. Cao *et al.* [104] employed XGBoost to retrieve Chl*a* from OLI in eight turbid lakes in eastern China.

TABLE II
AVERAGE SPECTRAL SAMPLING AND NUMBER OF BANDS, WITHIN 400–780-nm RANGE, FOR THE SIMULATED SPECTRA IN DIFFERENT SENSORS

Sensor	HICO	PRISMA	OLCI	MSI	OLI	LNext
Spectral sampling (nm)	<6	<9	<28	<60	<38	<42
Number of bands	66	44	14	6	5	9

MLP: MLP is one of the widely used ANN architectures which is investigated in this study as a parametric algorithm. It is a feed-forward ANN for approximating nonlinear regressions in a supervised learning technique, called back-propagation, for training the model parameters. The MLP model consists of at least three layers including the input, hidden, and output layers. The training process of the ANN models depends on the reduction of a loss function that is calculated based on the error between predicted and true values. The decline in the loss function follows an optimization algorithm with a learning rate. The learning rate controls the rate of change in the model (updating model parameters, weights and biases, in each layer) in response to the estimated error in the loss function. For further details on this algorithm and its parameters, hyperparameters refer to [105]. Schiller and Doerffer [106] developed an ANN as an approach to parameterize the inversion of a radiative transfer model. The study objective was to derive the concentrations of phytoplankton pigments, suspended matter and CDOM, as well as the aerosol path radiance from MERIS Rayleigh-corrected top-of-atmosphere reflectance spectra over turbid coastal waters. ANN has been extensively used in different studies for estimating water quality parameters from remote sensing observations [45], [107]–[110].

2) Input and Output Features: Input to all four ML algorithms consists solely of HICO-simulated hyperspectral bands. The best performing ML algorithm was then selected for testing on PRISMA-, OLCI-, MSI-, OLI-, and LNext-simulated R_{rs}. Table II summarizes the spectral sampling and the number of bands in each sensor that were used as input features in the best performing ML algorithm. The spectral sampling and number of bands are calculated for the spectral region captured in all study sites (400–780 nm). The panchromatic band in OLI was also included [111]. All input features are normalized using median centering and interquartile range (IQR) scaling which is robust to outliers. The output variable, PC, is first log-transformed and then scaled to the range of (0, 1).

3) Hyperparameter Tuning: A data split of 80/20 for training and testing ML algorithms resulted in a total of 724 randomly selected pairs of colocated R_{rs} and PC measurements for training and validating the algorithms and tuning hyperparameters in a fivefold cross validation, leaving the rest of the data for testing ($N = 181$). Other data splits (70/30, 60/40, and 50/50) were also investigated for training and testing the ML algorithms using HICO dataset. The cross validation approach was used to avoid overfitting and to ensure that the training dataset is randomly distributed in different segments and the model performance was not significantly influenced by the size and distribution of training datasets [90]. The tuning procedure for each ML model is described below.

PLSR: The parameter tuning procedure in the PLS model aims to optimize the number of components (n_components) and find a subset of input bands that can produce the lowest median symmetric accuracy (MdSA; see (4)) in a fivefold cross validation. A stepwise feature selection method was applied in this study to simultaneously find the optimal n_components and the subset of bands. The maximum value of n_components (n_max) for PLSR applied on hyperspectral data was set to 40, and for the multispectral missions (in case this model was the best performer), it was set to the number of bands in each (44, 14, 6, 5, and 9 for PRISMA, OLCI, MSI, OLI, and LNext, respectively). The principle of selection was to first develop a PLSR model with a selected n_components smaller than n_max value. Then, the input bands were sorted based on the importance metric derived from the developed PLSR model. The PLSR importance metric for each band was calculated as its weighted absolute value of the PLSR coefficient, where the weight [W; (2)] corresponds to the fraction of the standard deviation of the respective band to the total standard deviation of all bands. In the next step, PLSR models with the previously selected n_components were fit to different subsets of HICO bands, where in each run, a band with the lowest importance metric was discarded until the number of bands remaining was equal to the n_components. This approach was repeated for all different n_components values lower than n_max. The n_components and subset of bands that produced the lowest MdSA were selected as the optimal combination of parameters to develop the final PLSR model using all training data

$$W_i = \frac{StDev_i}{\Sigma_i StDev_i}, \quad \text{where } i \text{ is i-th band.} \tag{4}$$

SVR: A grid-search approach was utilized to find the kernel function and optimize the values for the penalty coefficient (C) and kernel parameter (gamma). RBF was selected as the kernel type. The C value minimizes the regularization error, and gamma defines the curvature in the RBF kernel. Values of C and gamma can affect the prediction skill of an SVR model [112]–[114]. These values for C and gamma were selected between (1, 10, 100) and (0.01, 0.1, 1), respectively. The combination of hyperparameters that produced the lowest MdSA in a fivefold cross validation was employed to develop the final SVR model using all training data.

XGBoost: To determine the structure of the XGBoost model, six hyperparameters including, alpha, gamma, the number of trees (n_estimator), maximum tree depth (max_depth), fraction of samples to randomly subsample at each step of training (subsamples), and fraction of features to be used randomly for each training step (colsample_bytree) were tuned in a grid-search strategy. The regularization parameters (alpha and gamma) were used to help reduce the model complexity and improve the performance. These two hyperparameters were selected between (1e-3, 0.01, 0.1, 1, 10, 100, 1000). The number of trees was within the range of (1, 20) with a

TABLE III

SELECTED STATE-OF-THE-ART ALGORITHMS FOR PREDICTING PC FROM R_{rs}, WITH COEFFICIENTS TUNED TO THE FREMONT, INDIANA, AND ERIE COMBINED DATASETS

Reference	Algorithm	Name
[118]	$PC = 0.33 \times \frac{R_{rs}(650)}{R_{rs}(625)} + 0.26$	S00
[2]	$PC = 0.30 \times \frac{R_{rs}(700)}{R_{rs}(600)} + 0.29$	M09
[23]	$PC = 0.16 \times [R_{rs}^{-1}(615) - R_{rs}^{-1}(600)] \times R_{rs}(725) + 0.24$	H10
[119]	$PC = 0.29 \times \frac{R_{rs}(709)}{R_{rs}(600)} + 0.28$	M12

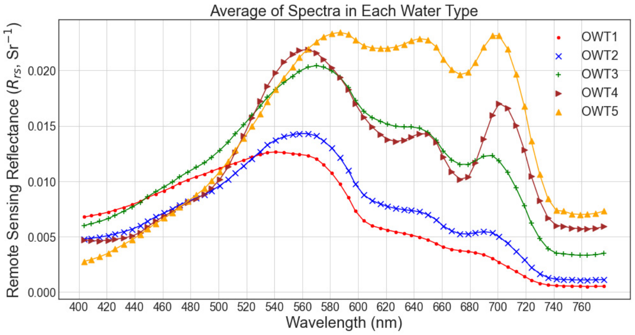

Fig. 4. Average of spectra in each OWT.

step size of 1. Other model parameters including max_depth, subsamples, and colsample_bytree were selected from 1 to 10, (0.7, 0.8, 0.9), and (0.7, 0.8), respectively. A fivefold cross validation approach was applied to the training dataset to determine the MdSA values. The combination of hyperparameters that produced the lowest MdSA was selected to develop the final XGBoost model using all training data.

MLP: The hyperparameter tuning approach in the MLP model was performed using a grid-search strategy to find the optimal values for the activation function, the number of hidden layers and nodes in each layer, the optimization algorithm, the learning rate, and the regularization term. The activation functions tested were rectified linear units (ReLUs), hyperbolic tangent (tanh), logistic, and identity functions. The optimal number of hidden layers tested was a maximum of three with no more than ten (even numbers in this range were tested) nodes in each. The optimization algorithm was selected among the limited Broyden–Fletcher–Goldfarb–Shanno (LBFGS) algorithm [115], stochastic gradient descent (SGD) [116], and Adam [117] optimization solver. The learning rate was adjusted according to three strategies of constant, inverse scaling (invscaling), and adaptive. The regularization term (alpha) was selected between (1e-3, 0.01, 0.1, 1, 10, 100, 1000). Similar to the other methods, a fivefold cross validation approach was applied to the training dataset in a grid-search strategy to calculate the MdSA values. The combination of hyperparameters that produced the lowest MdSA was used to develop the final MLP model using all training dataset.

F. State-of-the-Art PC Algorithms

The precision and accuracy of PC retrievals using the ML models applied to hyperspectral data were compared with those from well-validated state-of-the-art R_{rs} centered algorithms reviewed in [17], such as [2], [23], [118], and [119] (Table III). Simis *et al.* [22] also developed a band-ratio algorithm. However, since inherent optical properties (IOPs) such as absorption data were required to optimize the algorithm parameters, this approach was not included in the list of benchmark algorithms in this study. The band-ratio regressions target the PC absorption feature in the R_{rs} spectra between 600 and 625 nm. For the implementation of these algorithms, the closest HICO bands in the *in situ* R_{rs} spectra to the algorithm index were supplied, i.e., no attempt was made to recalibrate the algorithms' spectral indices. However, the algorithm coefficient and intercept were locally retuned using the training dataset in this study.

G. Performance Indicators

The performance of different approaches in estimating PC from hyperspectral and multispectral datasets was examined using both linear mean absolute percentage error (MAPE) and log-transformed metrics. The performance assessment is also reported based on the OWTs found in the dataset (in Section II-D). The evaluation metrics were calculated using the field-based testing dataset ($N = 181$, in the 80/20 split), which is independent of the training set ($N = 724$, in the 80/20 split). Calculations of the metrics are carried out using the estimated PC (E) against the field-measured data (M). These metrics include

$$\mathrm{SSPB} = 100\,\mathrm{sign}(z)(10^{|z|} - 1)[\%] \quad \text{where } z = \mathrm{Median}\left(\log_{10}\left(\frac{E}{M}\right)\right) \tag{5}$$

$$\mathrm{MdSA} = 100(10^{y} - 1)[\%] \quad \text{where } y = \mathrm{Median}\left|\log_{10}\left(\frac{E}{M}\right)\right| \tag{6}$$

$$\mathrm{MSA} = 100(10^{y} - 1)[\%] \quad \text{where } y = \mathrm{Mean}\left|\log_{10}\left(\frac{E}{M}\right)\right| \tag{7}$$

$$\mathrm{RMSLE} = \left[\frac{\sum_{i=1}^{N}(\log_{10}(E_i) - \log_{10}(M_i))^2}{n}\right]^{1/2} \tag{8}$$

$$\mathrm{MAPE} = \frac{1}{n}\sum_{i=1}^{N}\left|\frac{M_i - E_i}{M_i}\right| \tag{9}$$

where SSPB is the symmetric signed percentage bias, MdSA is the median symmetric accuracy (which was used to tune ML hyperparameters), and MSA is the mean symmetric accuracy. These metrics are symmetric and resistant to outliers [120]. SSPB, MdSA, and MSA are the key metrics to compare the results from different band configurations and algorithms. RMSLE is the root mean square log error. The slope associates with the linear regression fit between estimated and measured PC. Slope and MAPE are included to facilitate comparisons with the previously published results.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

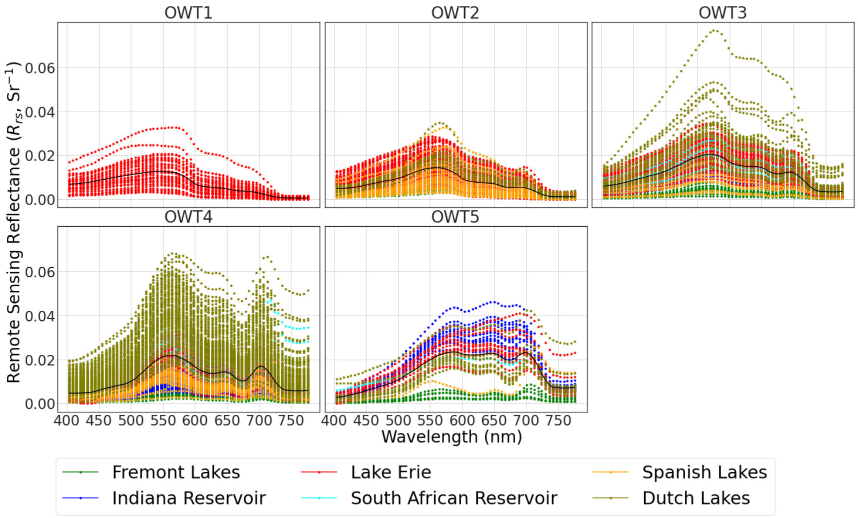

Fig. 5. Distribution of R_{rs} in each OWT. Mean and all R_{rs} are, respectively, shown in black solid line and dashed lines (colored by the study site) in each OWT. The nonstandardized HICO R_{rs} values are shown in this figure.

TABLE IV
PERCENTAGES OF EACH DATASET REPRESENTATIVE OF EACH OWT

	OWT1	OWT2	OWT3	OWT4	OWT5
Fremont Lakes	0	11.19	16.42	67.91	4.48
Indiana Reservoir	0	0	1.99	88.74	9.27
Lake Erie	10.91	43.33	23.33	18.48	3.94
South African Reservoir	0	0	40	50	10
Spanish lakes	0	41.84	8.16	48.98	1.02
Dutch lakes	0	4.95	13.74	76.37	4.95
Total Spectra	3.98	22.98	15.36	52.82	4.86

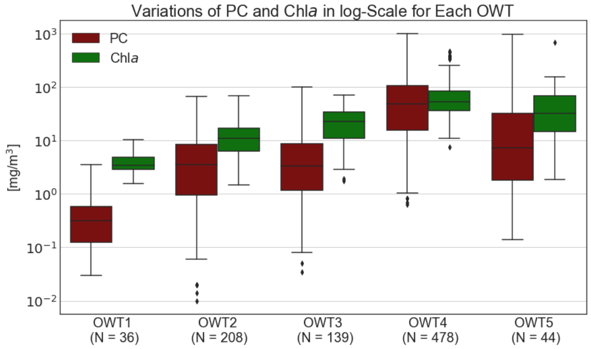

Fig. 6. Ranges of PC and Chl*a* are displayed in log scale for each OWT in a box and whisker plot. The boxes display the median, and the 25% and 75% quartiles of all data in each OWT. The whiskers are a representation of 1.5 multiplication of an IQR. Points are values outside this range.

III. RESULTS

A. OWTs

The assignment of the spectra to one of the 21 OWTs in [50] led to only 13 clusters in our dataset, where two of them had less than four members. Therefore, as suggested in [51], only a subset of the original OWTs in [50] were considered to provide a near-uniform distribution of spectra in each OWT and still cover the continuum of optical conditions in the dataset. This subset in our study has five clusters. OWTs 1 and 2 delineate the common spectra found in oligotrophic and/or coastal waters. OWTs 3 and 4 are found in lakes and coastal estuaries with increasing phytoplankton bloom densities and turbidity associated with detrital matter. OWT5 represents waters high in sediment. Fig. 4 compares the shape and magnitude of the calculated average of spectra in each OWT.

Fig. 5 and Table IV summarize the distributions of R_{rs} assembled from all study sites in each OWT ($N = 905$ for paired field-based radiometric and PC data in all sites). A large portion of spectra in this study were assigned to OWT4 with 478 spectral curves. Most of the spectra in the Fremont Lakes and Indiana Reservoirs represent OWT4 (∼68% and ∼89%, respectively). Only Lake Erie spectra represent OWT1 and are distributed in all OWTs, mainly in OWTs 2 and 3 (∼43% and ∼23%, respectively). There are no spectra from Indiana and South African Reservoirs in OWT2.

Fig. 6 shows the average and range of PC and Chl*a* values in our dataset per OWT. OWT1 ($N = 36$) has the lowest PC and Chl*a* values of 0.51 ± 0.69 and 4.1 ± 2.02 mg/m^3, respectively. The highest PC and Chl*a* were in OWT4 ($N = 478$) with values of 100.2 ± 150.87 and 67.59 ± 55.14 mg/m^3, respectively.

B. Correlation Analysis

To better understand the effects of the optical conditions on the information that each band may carry with respect to PC, Fig. 7 shows the correlation of PC with HICO R_{rs} measured in each individual band. The correlation analysis was performed for each OWT separately.

Each OWT shows different individual (or ranges of) HICO bands selected to have the highest correlation, emphasizing the impact of other OACs in masking the dual spectral features of PC around 620 and 650 nm. However, in all OWTs, there is maximal correlation around the red edge, marking this region as important in HAB detection (either through cross correlation of PC and Chl*a* or unique scattering features of cyanobacteria). In OWT1 (typical in oligotrophic and/or coastal waters), all spectral regions are almost equally important in PC retrieval. In OWT 2, the spectral range > 700 nm

TABLE V

Overall Performance Analyses of ML and Benchmark Models Applied to Hyperspectral and Multispectral Datasets. The Performance Indicators Are Calculated for the Testing Dataset, to Enable the Performance Comparison in Different Spectral Resolutions. Where ∗, +, and # Mark the Best Performance Within Each of the HICO ML, Multispectral MLP, and Benchmark Algorithm Assessments, Respectively

		Models	Scenarios	Performance Indicators					
				SSPB (%)	MdSA (%)	MSA (%)	RMSLE	MAPE	Slope
Resampled Hyperspectral R_{rs} Spectra	HICO	PLSR	PLSR-HICO	14.57	125.11	180.9	0.60	3.64	0.65
		SVR	SVR-HICO	9.46	71.28	132.54	0.58	2.2	0.72
		XGBoost	XGB-HICO	-7.82	102.37	141.98	0.53	3.0	0.71
		MLP	MLP-HICO	1.1^*	64.51^*	102.03^*	0.43^*	1.87^*	0.81^*
	PRISMA	MLP	MLP-PRISMA	6.12	63.85	104.5	0.44	2.14	0.81
	OLCI	MLP	MLP-OLCI	6	72.8^+	125.09^+	0.49	2.59	0.79^+
	MSI	MLP	MLP-MSI	-3.53^+	92.55	131.94	0.49	2.77	0.77
	OLI	MLP	MLP-OLI	4.8	125.61	176.33	0.59	4.6	0.65
	LNext	MLP	MLP- LNext	8.18	82.58	125.51	0.48^+	2.25^+	0.79^+
	Benchmark algorithms	S00	S00	$7.76^\#$	$132.33^\#$	$193.33^\#$	$0.61^\#$	$3.85^\#$	0.61
		M09	M09	-29.34	136.55	203.11	0.63	5.22	$0.67^\#$
		H10	H10	-30.36	187.24	292.70	0.78	14.69	0.46
		M12	M12	-20.66	153.17	217.39	0.67	8.32	0.64

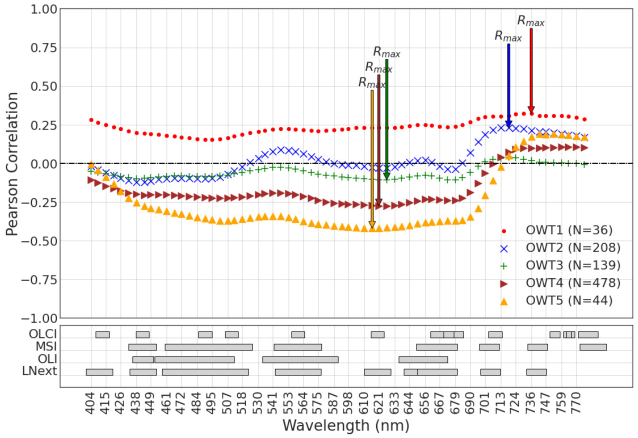

Fig. 7. Pearson correlation of PC with HICO R_{rs} measured in each individual band. OLCI, MSI, OLI, and LNext spectral coverages are plotted as a reference. OLI panchromatic band (503–676 nm) is not shown on the plot.

is the most important in PC retrieval. The highest correlations in OWTs 3, 4, and 5 were around the PC absorption peak around 620 nm. The blue region in OWTs 1 (typical in oligotrophic and/or coastal waters), 4 (typical in eutrophic waters), and 5 (sediment-rich waters) is contributing more information for PC retrieval compared with the other two OWTs. However, in OWTs 1, 3, 4, and 5, the correlations between PC and R_{rs} show small variations throughout most of the visible spectrum. This means that there is possibility for ambiguity when PC is resolved only through cross-correlations with the dominant optical features of the spectrum. The spectral coverage available with each multispectral dataset from OLCI, MSI, OLI, and LNext is plotted in Fig. 7 for comparison.

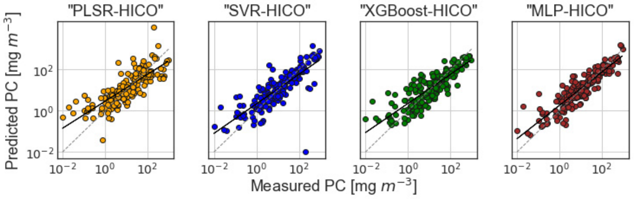

Fig. 8. Scatter plots derived from applying different ML models on HICO bands to retrieve PC ($N = 181$). Dashed line shows 1:1 relationship. Solid line shows fit regression for testing data ($N = 181$).

C. Performance Evaluation

The parameters of the four ML models were tuned for HICO, and the parameters of the best performing one were retuned for PRISMA and the multispectral datasets to assess the role of spectral sampling in the ML model performance. Further explanation of model development is provided in the Appendix. Table V summarizes all the model setups (scenarios; different models applied to different predictors). The performance of each model was assessed using the metrics listed in Section II-G for the test dataset ($N = 181$).

Fig. 8 illustrates the predicted PC derived from different ML models applied to HICO-resampled R_{rs} plotted against the measured values. As shown in Table V, the MLP model outperformed others with the lowest SSPB, MdSA, MSA, RMSLE, and MAPE and the highest slope. SVR performed

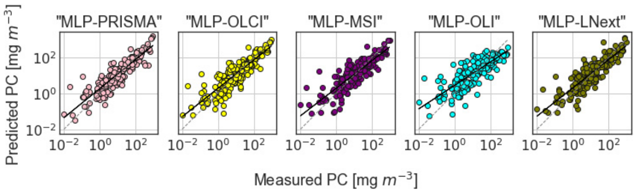

Fig. 9. Scatter plots derived showing the performance of MLPs on other spectral band settings as labeled. Dashed line shows 1:1 relationship. Solid line shows fit linear regression to the test data ($N = 181$).

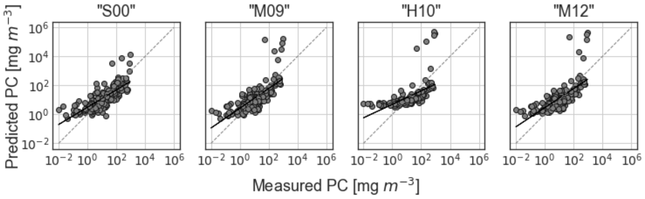

Fig. 10. Scatter plots illustrating performance of existing band-ratio models as employed on respective R_{rs} (λ). Dashed line shows 1:1 relationship. Solid lines show fit regression for testing data ($N = 181$).

better than PLSR. XGBoost outperformed PLSR with lower MdSA, MSA, RMSLE, and MAPE and higher slope. Therefore, MLP was selected as the best performing ML model to produce PC from HICO bands.

Note that experiments with other training/validation split sizes were also performed to find the best performing ML model when applied to HICO bands. MLP with 80/20 split size performed the best among all ML models with different split sizes. Therefore, all scenarios presented here (different ML models and hyperspectral and multispectral input features) are the results of employing this split in ML model development.

The performance of MLPs retuned for other band configurations (Table II) is shown in Fig. 9. MLP hyperparameter tuning results are summarized in the Appendix. Table V compares the performance of these models in retrieving PC from PRISMA- and multispectral-simulated reflectance against the ones derived from HICO bands. The table shows that between hyperspectral datasets, HICO outperformed PRISMA with lower SSPB, MSA, RMSLE, and MAPE. MLP-PRISMA produced a slightly lower MdSA and the slopes were equal in these two hyperspectral scenarios. The performance of MLP in PC retrieval degraded from hyperspectral datasets to OLCI, LNext, MSI, and OLI, respectively, in terms of MdSA, MSA, and slope. The analysis shows that OLI spectral bands were the least suitable to calculate PC from applying an MLP model to this dataset with the highest MdSA, MSA, RMSLE, and MAPE and the lowest slope.

Using the same dataset, we demonstrate that the ML models offer major improvements compared with previously published band-ratio algorithms (Section II-F). The PC values retrieved from these algorithms are plotted against the field-based PC measurements in Fig. 10. Table V summarizes their performance using the statistical indicators in Section II-G. Results show that S00 performs best in terms of SSPB, MSA, RMSLE, and slope. However, all band-ratio models performed poorly in comparison with the MLP models applied to either HICO, PRISMA, or multispectral spectra (Table V). Higher values of PC were overestimated by all benchmark models. Concentrations of Chl*a* and other pigments can modify the PC absorption and reflectance features. Also, these features can occur at different wavelengths depending on the variations in PC and Chl*a* concentrations [2]. However, these factors are not considered in the development of band-ratio algorithms [2], which might contribute to the poor performance of these algorithms compared with the ML models in the current study. Previous studies in [45], [104], [121], and [122] show the value of ML models and clearly demonstrate their advantages over empirical algorithms with hard-coded coefficients. The ML models tend to be flexible and learn the nonlinear association between R_{rs} and IOPs.

1) Performance Evaluation Based on OWTs: The performance evaluation in different scenarios was further categorized based on OWTs.

Fig. 11 shows which water type will benefit the most from each ML model applied to the HICO-simulated R_{rs} dataset. The MLP produced the lowest SSPB, MdSA, and MSA in OWTs 1, 2, and 4. In OWT3, all ML models performed almost equally in terms of MdSA (SVR produced marginally lower MdSA than other models). However, PLSR and MLP produced the lowest SSPB and MSA in this OWT, respectively. In OWT5, XGBoost (the lowest SSPB) and MLP (the lowest MSA) performed closely with equal MdSA.

Fig. 12 compares the performance of MLP models tuned for hyperspectral against the ones tuned for multispectral datasets, in different OWTs. Scenario MLP-PRISMA outperformed others in OWT1. MLP-HICO and MLP-LNext performed closely in this OWT. OWT1 includes waters with the lowest values for PC (0.51 ± 0.69 mg/m^3) and Chl*a* (4.1 ± 2.02 mg/m^3) in our matchup dataset (Fig. 6). MLP-HICO produced the lowest SSPB, MdSA, and MSA in OWT2. MLP-MSI produced the lowest SSPB and MdSA in OWT3, but MLP-HICO outperformed the rest with the lowest MSA. MLP applied to hyperspectral data produced the lowest SSPB, MdSA, and MSA in OWT4. This OWT is assigned to waters with the highest range of PC and Chl*a* values (100.2 ± 150.87 and 67.59±55.14 mg/m^3, respectively). In OWT5, with sediment-rich waters, MLP-OLI performed best with the lowest SSPB, MdSA, and MSA. MLP-LNext performed the best in OWT1 between multispectral datasets.

2) Performance Evaluation Based on PC:Chla: The performances of MLP models applied to hyperspectral and multispectral datasets were compared for different ranges of values for PC:Chl*a*. Results in Fig. 13 show that the performances of MLP models applied to hyperspectral data of HICO and PRISMA were comparable in terms of the lowest SSPB, MdSA, and MSA, when PC:Chl*a* values are less than one. The performance of MLP-OLI was consistently lower than that of other sensors in this range of PC:Chl*a* values. In the presence of cyanobacteria (PC:Chl*a* $\geq$ 1), scenario MLP-OLCI outperformed the rest in terms of SSPB, MdSA, and MSA. MLP-HICO and MLP-PRISMA produced comparable results to those of MLP-OLCI when cyanobacteria were dominant. MLP-OLI performed poorly in comparison with other sensors in the presence of cyanobacteria.

Kutser *et al.* [30] declared that MERIS band configuration (bands 6 and 7) allows detection of PC when it is present in

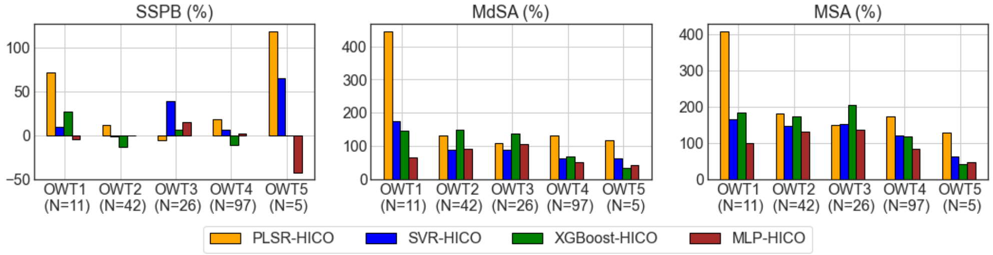

Fig. 11. Performance metrics for the four ML models implemented for HICO-resampled Rrs in retrieving PC values in different OWTs.

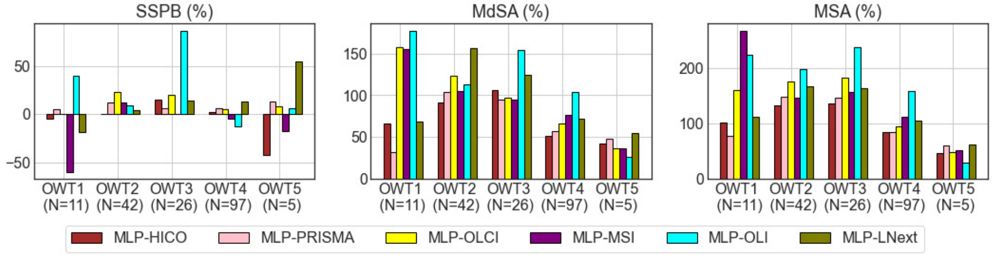

Fig. 12. Performance metrics for the MLP models implemented for the five datasets of hyperspectral and multispectral R_{rs} in retrieving PC values in different OWTs.

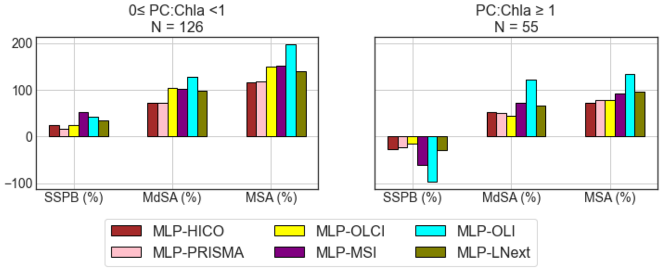

Fig. 13. Performance metrics for the MLP models implemented for hyperspectral and multispectral resampled-R_{rs} in retrieving PC values in different range of PC:Chl*a* values.

relatively high concentrations. Metsamaa *et al.* [15] show that the cyanobacteria double spectral feature can be detectable when Chl*a* in the Baltic Sea is at least 8–10 mg/m^3. Our results in Fig. 13 demonstrate that the PC determination using multispectral data of OLCI can perform best when cyanobacteria is dominant (PC:Chl*a* $\geq$ 1), which can happen even at low concentrations.

IV. Discussion

Remote sensing studies are shifting toward globally applicable models for retrieving and monitoring spatiotemporal distribution of cyanobacteria. However, diversity in optical conditions, both temporally and spatially, makes this task challenging. This study confirms that hyperspectral data, through application of ML algorithms, can be used to estimate PC when relevant information is extracted from hundreds of bands. Spectral resolutions of HICO and PRISMA outperformed the ones of multispectral sensors to retrieve PC from an MLP model, when all OWTs were combined. However, results demonstrated that the best performing spectral resolution and ML algorithm is different in each OWT.

A. Modeling Algorithm

Between ML models applied to HICO-resampled R_{rs} data, MLP outperformed others in retrieving PC for almost all OWTs. MLP leverages the spectral information in all bands to be able to recognize the pattern of optical complexity in each OWT. SVR also takes advantage of the full spectral information by applying nonlinear kernels and mapping the features into a higher dimensional space to create linear (or approximately linear) problems. As Table V demonstrates, this model performed poorly overall, compared with MLP when applied to the HICO dataset. Unlike MLP and SVR, PLSR and XGBoost are developed based on a reduced number of input features and a subset of spectral bands with the highest information. As Figs. 16 and 17 show, the three most important bands selected in PLSR were 461, 467, and 730 nm and in XGBoost were 501, 713, and 719 nm, respectively, while other bands contributed less to the model development. PLSR discarded 50 bands and XGBoost associated an importance metric less than 0.005–29 bands. However, the reduced feature space is not necessarily able to capture the optical complexity of all OWTs. Therefore, the MLP model was selected as the ML model to capture the nonlinear and complex interaction between HICO-resampled R_{rs} and PC. Employing all HICO spectral bands, the MLP model estimated PC across a broad spectrum of OWTs. Pyo *et al.* [123] also utilized an NN as the regression model to estimate PC from airborne hyperspectral data for Baekje weir located at Geum River in South Korea.

B. Hyperspectral Versus Multispectral Data

MLP was further tested for estimating PC from PRISMA and multispectral datasets. As Fig. 7 shows, each OWT has the highest correlation with HICO-resampled R_{rs} at a specific wavelength (or equally high correlations in a range of the spectrum) that is not necessarily covered by the multispectral

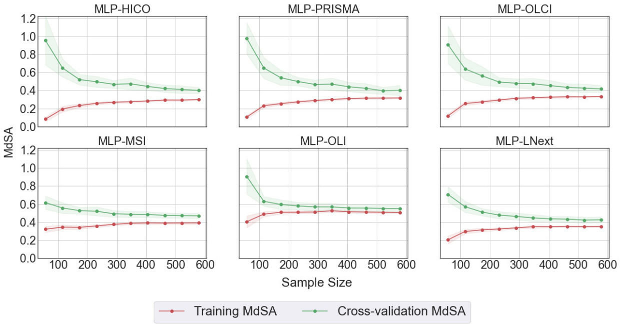

Fig. 14. Learning curves for MLP models applied to hyperspectral and multispectral datasets.

bands. For example, the highest correlation between R_{rs} and PC occurs at wavelengths >700 nm in OWTs 1 and 2, and OLI spectral bands do not cover this range (Fig. 7). Adding the OLI panchromatic band (503–676 nm) improved the performance of MLP with a ~2% decrease in MdSA compared with the scenario of removing it from the input features (results of the latter are not presented here). But MLP-OLI had the worst performance among all hyperspectral and multispectral scenarios. Also, a cross correlation between R_{rs} and other OACs can occur at selected wavelengths where employing HICO bands can potentially untangle this complexity. For example, in OWT4, 673 nm is in the range of wavelengths contributing the most information in PC retrieval; however, it could be a cross correlation of PC and Chl*a* with R_{rs}.

Metsamaa *et al.* [15] show that hyperspectral reflectance, with 10-nm spectral resolution and high signal-to-noise ratio (SNR > 1000:1), is required to capture the cyanobacteria characteristic double feature seen in relatively clear, cyanobacteria-dominated waters (PC absorption feature at 630 nm and reflectance peak at 650 nm for the Baltic Sea). On the other hand, the drawback in most of the current multispectral satellite sensors is their lack of specific spectral bands to capture this specific spectral feature [18]. Between past and current satellite sensors, OLCI and its heritage MERIS are the most appropriate options that meet the minimum spectral requirement to detect optical characteristics of cyanobacteria, the absorption peak near 620 nm, for accurate PC detection and monitoring. Landsat Next will provide continuity with instruments onboard Landsat-8 and -9 and compatibility with Sentinel-2 data. The proposed additional narrow spectral bands (~20-nm FWHM) include the orange (620 nm) and red (650 nm) part of the spectrum to retrieve Chl*a*, PC, and turbidity. Thus, the LNext multispectral measurement concept will also cover the double spectral feature of PC with bands centered at 620 and 650 nm.

C. Sample Size

In estimation problems, the available training sample size must be large enough to span the complexity of optical conditions so that the model is able to accommodate the available training samples reasonably well and generalize to new data [124]. Therefore, a fivefold cross validation approach was used in the training process to assess how well each scenario can approximate PC. The learning curves in Fig. 14 show the training and cross validation MdSA errors when MLP models are trained to hyperspectral and multispectral datasets using differently sized training datasets. In the MLP-OLI model, the training and cross validation errors did not increase and decrease, respectively, when the size of training data increased. Also, the gap between calculated MdSA errors in training and cross validation was small for different sizes of training data. Therefore, increasing the training data size did not improve its performance and this model produced the highest MdSA errors compared with other models when it was trained using all training data. The MLP-OLI model was unable to capture the hidden underlying patterns between input spectral features (five OLI bands including the panchromatic band) and PC. Increasing the size of training data increased the training MdSA error and decreased the cross validation MdSA error for MLP models applied to HICO, PRISMA, OLCI, MSI, and LNext. The MdSA curves for training and cross validation in MLP-MSI and MLP-LNext did not change significantly after using a training dataset with a size of ~400.

When all training data were used, the training MdSA error for MLP-MSI was more than the ones for MLP-HICO, MLP-PRISMA, MLP-OLCI, and MLP-LNext. Also, the gap between cross validation and training MdSA errors in MLP-MSI was shorter compared with MLP-HICO, MLP-PRISMA, and MLP-OLCI. The shorter gap, as well as the larger MdSA errors, implies that although the MLP-MSI model produced low variance, this model was less successful in capturing the data complexity compared with MLP-HICO, MLP-PRISMA, MLP-OLCI, and MLP-LNext. The training and cross validation MdSA errors in MLP-HICO and MLP-PRISMA were marginally lower than the ones for MLP-OLCI and MLP-LNext when all data were used for training. MLP-LNext performed closely to MLP-OLCI. This demonstrates that, with the same sample size of ~600, MLP-HICO and MLP-PRISMA were better in modeling the patterns in the spectral input features and their complex nonlinear relationships with PC. That said, the MdSA value of these models is still larger than (or around) the uncertainties in field-based measurements. These uncertainties arise from random and systematic errors and are propagated to the model predictions [122]. Even though the test dataset ($N = 181$) was independent of the data used in the training of each ML model, the data were still originating from the same study sites with similar optical characteristics as the training dataset which brings uncertainties in the generalizability of the conclusions of this study to other sites.

D. Uncertainty in In Situ Radiometry Data

The assumption of an angular distribution of upwelling radiance just beneath the surface (BRDF), that is independent of the viewing direction (i.e., a diffuse BRDF), is expected to introduce uncertainty in R_{rs}. Although BRDF correction algorithms are developed in the literature [125], creating a correction factor applicable to all OWTs is challenging and is likely to introduce more uncertainty and error, due to assumptions on the relationship of upwelling radiance with

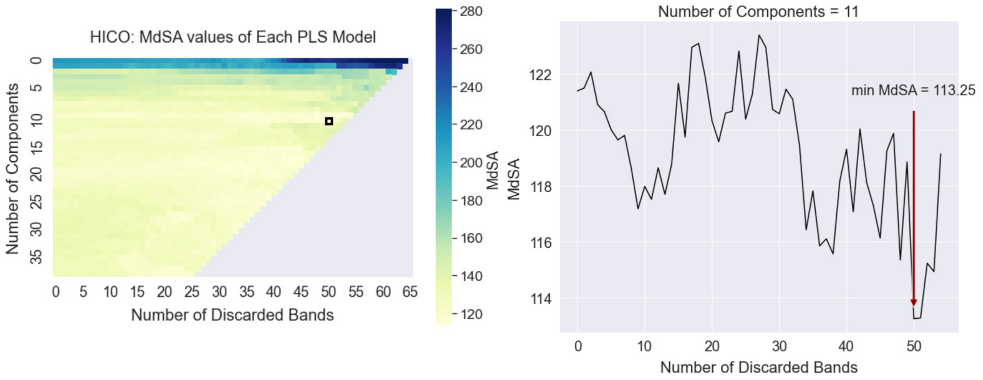

Fig. 15. (Left) MdSA of each PLS model developed using the selected number of components (y-axis) and discarding a subset of hyperspectral data with the lowest importance metric values (x-axis). MdSA values are calculated in a fivefold cross validation approach. The gray area shows the number of bands that were not excluded due to the required minimum number of bands to achieve a PLSR model with the specified n_components. The black box shows the number of components and number of discarded bands that produced the lowest MdSA value. (Right) MdSA of each PLS model developed with 11 components and discarding different numbers of bands (x-axis) with the lowest importance metric values.

absorption and backscatter in different OWTs. The dataset in this study is a combination of different instruments and methodological approaches (with different illumination and viewing geometries), and data analysis methods (with different air–water interface effect correction approaches), in waters with varying optical conditions. This will introduce by nature a level of unavoidable variability and uncertainty in the R_{rs} dataset.

V. Conclusion

Results of this study show that when developing algorithms applicable to different optical water conditions is considered, the performance of MLP models applied to hyperspectral data (including HICO and PRISMA) surpasses that of those applied to multispectral datasets with median biases of ~73%, 93%, 126%, and 83% for OLCI, MSI, OLI, and LNext, respectively. Therefore, this study quantifies the MLP performance loss when datasets with lower spectral resolutions are used for PC mapping. Knowing the extent of performance loss, researchers can either employ hyperspectral data at the cost of computational complexity, or alternatively utilize datasets with reduced spectral capability in the absence of hyperspectral data. A few selected band-ratio algorithms, that target PC absorption at 620 nm, were also tested in this study. Results showed that these models performed poorly in comparison with ML models applied to hyperspectral and multispectral datasets.

The performance assessment of different scenarios was also conducted for the derived optical conditions in the matchup dataset. MLP applied to HICO and PRISMA outperformed other scenarios when the optical water type includes waters lowest in PC and Chl*a* (i.e., OWTs 1 and 2) and also highest in PC and Chl*a* (i.e., OWT4), and when cyanobacteria were not dominant (PC:Chl*a* < 1). MLP applied to LNext performed best between other multispectral scenarios in OWT1. When cyanobacteria were dominant (PC:Chl*a* ≥ 1), MLP-OLCI outperformed other scenarios in estimating PC. A correlation analysis was also conducted on HICO-resampled R_{rs} data, to find the band with the highest correlation with PC in each OWT. The most relevant information for PC retrieval was around the PC absorption peak (~620 nm) for OWTs 3, 4, and 5. The longer wavelengths in the spectrum (>700 nm) carried the most information for PC retrieval in OWTs 1 and 2. Shorter wavelengths in the blue region seem to play the most important role in PC estimation for OWTs 1 (typical optical condition in coastal and oligotrophic waters), 4 (eutrophic water), and 5 (sediment-rich waters). These correlation analyses reinforced the value of employing hyperspectral data to extract PC in different water types. Band-ratio algorithms (albeit incorporating selected red edge bands) and multispectral datasets cannot investigate the full spectrum. We conclude that for a robust estimation of PC in optically complex waters where the characteristic spectral features are commonly masked, hyperspectral R_{rs} offers adequate spectral cues to retrieval ML algorithms adept at mining relevant information. MLP model advances PC estimation from *in situ* hyperspectral radiometric data [126] and/or highly accurate atmospherically corrected remote sensing data and enables categorical discrimination of PC-dominated Cyano HABs.

This study is particularly informative for future research and operations when merged PC products from multispectral and hyperspectral instruments are desired. Important pathways are the future Landsat Next mission that can also make headway in PC retrieval.

Appendix

Tuning ML parameters for hyperspectral and multispectral datasets are summarized as below.

PLSR: Fig. 15 (left) illustrates the change in MdSA values with different numbers of n_components and discarded bands, calculated based on the training dataset in a fivefold cross validation. The lowest MdSA value of 113.25 was produced for a PLSR model with 11 components when discarding the 50 bands with the lowest importance. The sensitivity of the PLSR model with 11 components to the number of discarded bands with the lowest importance is shown in Fig. 15 (right). Fig. 16 illustrates the importance metrics calculated for each HICO band in the optimized PLSR model. The discarded bands are shown in red bars.

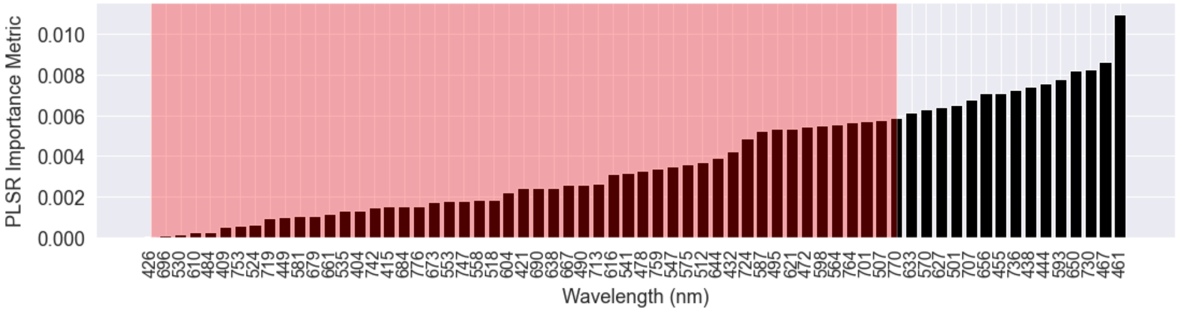

Fig. 16. HICO bands sorted based on the PLSR importance metric. The red bars are the discarded bands.

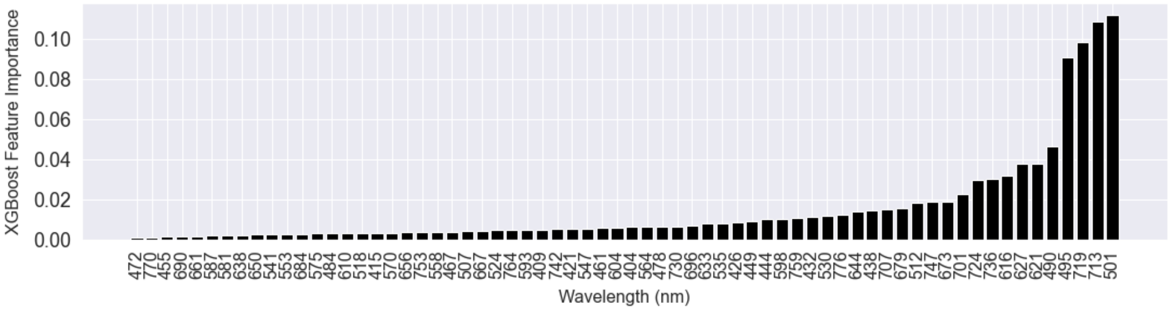

Fig. 17. HICO bands sorted based on the XGBoost feature importance metric.

SVR: The model hyperparameters including C and gamma were tuned in a grid-search approach applied to the HICO training dataset in a fivefold cross validation. The minimum value of 68.40 for MdSA was produced with values of 100 and 0.01 for C and gamma, respectively.

XGBoost: The grid-search approach, applied to the training dataset, examined MdSA values in a fivefold cross validation, to tune the XGBoost hyperparameters values for HICO bands as input features. Results showed that the minimum MdSA value of 74.85 was produced with values of 0.7, 7, 18, and 0.9 for colsample_bytree, max_depth, n_estimators, and subsample, respectively. Lambda and alpha were 0.001 and 0.01, respectively. The feature importance metrics calculated for each HICO band in the optimized XGBoost model are shown in Fig. 17.

MLP: The lowest MdSA value of 55.66 was produced in a grid-search approach applied to the HICO training dataset in a fivefold cross validation, with a tanh activation function, "constant" for the learning rate, and LBFGS solver. The values for alpha and the number of nodes in each of three hidden layers were tuned at 0.001 and (10, 8, 4), respectively.

The MLP models tuned for PRISMA and multispectral datasets had the same learning rate (constant) and solver (LBFGS) as the final MLP model applied to HICO. Activation function for MLP-PRISMA, MLP-OLCI, and MLP-OLI was tanh. ReLU was selected as the activation function in MLP-MSI, and MLP-LNext. The values for alpha were tuned at 0.001, 0.001, 0.1, 0.01, and 0.001 for PRISMA, OLCI, MSI, OLI, and LNext, respectively. The number of nodes in each hidden layer for MLP applied to PRISMA, OLCI, MSI, OLI, and LNext was (8, 4), (8, 10, 2), (10, 8, 6), (8, 8), and (10, 2, 4), respectively.

Acknowledgment

The authors would like to thank Steve Ruberg (NOAA-GLERL) for providing the field data. Financial assistance was provided through the Global Water Futures program, Transformative Sensor Technologies and Smart Watershed project (C.R. Duguay, PI). Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

References

[1] *Toxic Cyanobacteria in Water*, 2nd ed. World Health Organization, Geneva, Switzerland, 2021.

[2] S. Mishra, D. Mishra, and W. Schluchter, "A novel algorithm for predicting phycocyanin concentrations in cyanobacteria: A proximal hyperspectral remote sensing approach," *Remote Sens.*, vol. 1, no. 4, pp. 758–775, Oct. 2009, doi: 10.3390/rs1040758.

[3] W. W. Carmichael and G. L. Boyer, "Health impacts from cyanobacteria harmful algae Blooms: Implications for the North American Great Lakes," *Harmful Algae*, vol. 54, pp. 194–212, Apr. 2016, doi: 10.1016/j.hal.2016.02.002.

[4] R. Helmer, I. Hespanhol, and E. B. Welch. (1997). *Water Pollution Control-A Guide to the Use of Water Quality Management Principles*. [Online]. Available: http://www.earthprint.com

[5] D. Sun *et al.*, "A novel support vector regression model to estimate the phycocyanin concentration in turbid inland waters from hyperspectral reflectance," *Hydrobiologia*, vol. 680, no. 1, pp. 199–217, Jan. 2012, doi: 10.1007/s10750-011-0918-7.

[6] T. T. Wynne, R. P. Stumpf, M. C. Tomlinson, and J. Dyble, "Characterizing a cyanobacterial Bloom in Western Lake Erie using satellite imagery and meteorological data," *Limnol. Oceanogr.*, vol. 55, no. 5, pp. 2025–2036, 2010, doi: 10.4319/lo.2010.55.5.2025.

[7] R. P. Stumpf, T. T. Wynne, D. B. Baker, and G. L. Fahnenstiel, "Interannual variability of cyanobacterial Blooms in Lake Erie," *PLoS ONE*, vol. 7, no. 8, Jan. 2012, Art. no. e42444, doi: 10.1371/journal.pone.0042444.

[8] M. Kahru and R. Elmgren, "Multidecadal time series of satellite-detected accumulations of cyanobacteria in the Baltic Sea," *Biogeosciences*, vol. 11, no. 13, pp. 3619–3633, Jul. 2014, doi: 10.5194/bg-11-3619-2014.

[9] J. M. Clark *et al.*, "Satellite monitoring of cyanobacterial harmful algal Bloom frequency in recreational waters and drinking water sources," *Ecol. Indicators*, vol. 80, pp. 84–95, Sep. 2017, doi: 10.1016/j.ecolind.2017.04.046.

[10] E. A. Urquhart, B. A. Schaeffer, R. P. Stumpf, K. A. Loftin, and P. J. Werdell, "A method for examining temporal changes in cyanobacterial harmful algal Bloom spatial extent using satellite remote sensing," *Harmful Algae*, vol. 67, pp. 144–152, Jul. 2017, doi: 10.1016/j.hal.2017.06.001.

[11] S. Mishra, R. P. Stumpf, B. A. Schaeffer, P. J. Werdell, K. A. Loftin, and A. Meredith, "Measurement of cyanobacterial Bloom magnitude using satellite remote sensing," *Sci. Rep.*, vol. 9, no. 1, pp. 1–17, Dec. 2019, doi: 10.1038/s41598-019-54453-y.

[12] A. Vander Woude, S. Ruberg, T. Johengen, R. Miller, and D. Stuart, "Spatial and temporal scales of variability of cyanobacteria harmful algal Blooms from NOAA GLERL airborne hyperspectral imagery," *J. Great Lakes Res.*, vol. 45, no. 3, pp. 536–546, Jun. 2019, doi: 10.1016/j.jglr.2019.02.006.

[13] C. E. Binding, L. Pizzolato, and C. Zeng, "EOLakeWatch; delivering a comprehensive suite of remote sensing algal Bloom indices for enhanced monitoring of Canadian eutrophic lakes," *Ecol. Indicators*, vol. 121, Feb. 2021, Art. no. 106999, doi: 10.1016/j.ecolind.2020.106999.

[14] A. G. Dekker, "Detection of optical water quality parameters for eutrophic waters by high resolution remote sensing," Ph.D. dissertation, Vrije Univ. Amsterdam, Amsterdam, The Netherlands, 1993.

[15] L. Metsamaa, T. Kutser, and N. Strömbeck, "Recognising cyanobacterial Blooms based on their optical signature: A modelling study," *Boreal Environ. Res.*, vol. 11, no. 6, pp. 493–506, 2006.

[16] T. Kutser, "Passive optical remote sensing of cyanobacteria and other intense phytoplankton Blooms in coastal and inland waters," *Int. J. Remote Sens.*, vol. 30, no. 17, pp. 4401–4425, 2009, doi: 10.1080/01431160802562305.

[17] I. Ogashawara, D. Mishra, S. Mishra, M. Curtarelli, and J. Stech, "A performance review of reflectance based algorithms for predicting phycocyanin concentrations in inland waters," *Remote Sens.*, vol. 5, no. 10, pp. 4774–4798, Sep. 2013, doi: 10.3390/rs5104774.

[18] Y. Yan, Z. Bao, and J. Shao, "Phycocyanin concentration retrieval in inland waters: A comparative review of the remote sensing techniques and algorithms," *J. Great Lakes Res.*, vol. 44, no. 4, pp. 748–755, Aug. 2018, doi: 10.1016/j.jglr.2018.05.004.

[19] R. K. Vincent *et al.*, "Phycocyanin detection from LANDSAT TM data for mapping cyanobacterial Blooms in Lake Erie," *Remote Sens. Environ.*, vol. 89, no. 3, pp. 381–392, Feb. 2004, doi: 10.1016/j.rse.2003.10.014.

[20] R. K. Vincent, *Fundamentals of Geological and Environmental Remote Sensing*. Upper Saddle River, NJ, USA: Prentice-Hall, 1997.

[21] L. Li, R. E. Sengpiel, D. L. Pascual, L. P. Tedesco, J. S. Wilson, and A. Soyeux, "Using hyperspectral remote sensing to estimate chlorophyll-a and phycocyanin in a mesotrophic reservoir," *Int. J. Remote Sens.*, vol. 31, no. 15, pp. 4147–4162, 2010, doi: 10.1080/01431161003789549.

[22] S. G. H. Simis, S. W. M. Peters, and H. J. Gons, "Remote sensing of the cyanobacterial pigment phycocyanin in turbid inland water," *Limnol. Oceanogr.*, vol. 50, no. 1, pp. 237–245, Jan. 2005, doi: 10.4319/lo.2005.50.1.0237.

[23] P. D. Hunter, A. N. Tyler, L. Carvalho, G. A. Codd, and S. C. Maberly, "Hyperspectral remote sensing of cyanobacterial pigments as indicators for cell populations and toxins in eutrophic lakes," *Remote Sens. Environ.*, vol. 114, pp. 2705–2718, Nov. 2010, doi: 10.1016/j.rse.2010.06.006.

[24] C. Le, Y. Li, Y. Zha, Q. Wang, H. Zhang, and B. Yin, "Remote sensing of phycocyanin pigment in highly turbid inland waters in Lake Taihu, China," *Int. J. Remote Sens.*, vol. 32, no. 23, pp. 8253–8269, Dec. 2011, doi: 10.1080/01431161.2010.533210.

[25] M. W. Matthews, S. Bernard, and L. Robertson, "An algorithm for detecting trophic status (chlorophyll-a), cyanobacterial-dominance, surface scums and floating vegetation in inland and coastal waters," *Remote Sens. Environ.*, vol. 124, pp. 637–652, Sep. 2012, doi: 10.1016/j.rse.2012.05.032.

[26] M. W. Matthews and D. Odermatt, "Improved algorithm for routine monitoring of cyanobacteria and eutrophication in inland and near-coastal waters," *Remote Sens. Environ.*, vol. 156, pp. 374–382, Jan. 2015, doi: 10.1016/j.rse.2014.10.010.

[27] T. T. Wynne *et al.*, "Relating spectral shape to cyanobacterial Blooms in the Laurentian Great Lakes," *Int. J. Remote Sens.*, vol. 29, no. 12, pp. 3665–3672, 2008, doi: 10.1080/01431160802007640.

[28] R. P. Stumpf *et al.*, "Challenges for mapping cyanotoxin patterns from remote sensing of cyanobacteria," *Harmful Algae*, vol. 54, pp. 160–173, Apr. 2016, doi: 10.1016/j.hal.2016.01.005.

[29] R. H. Ma, W. Kong, H. Duan, and S. Zhang, "Quantitative estimation of phycocyanin concentration using MODIS imagery during the period of cyanobacterial Blooming in Taihu Lake," *China Environ. Sci.*, vol. 29, no. 3, pp. 254–260, 2009.

[30] T. Kutser, L. Metsamaa, N. Strömbeck, and E. Vahtmäe, "Monitoring cyanobacterial Blooms by satellite remote sensing," *Estuarine, Coastal Shelf Sci.*, vol. 67, nos. 1–2, pp. 303–312, Mar. 2006, doi: 10.1016/j.ecss.2005.11.024.

[31] A. Ruiz-Verdú, S. G. H. Simis, C. de Hoyos, H. J. Gons, and R. Peña-Martínez, "An evaluation of algorithms for the remote sensing of cyanobacterial biomass," *Remote Sens. Environ.*, vol. 112, no. 11, pp. 3996–4008, Nov. 2008, doi: 10.1016/j.rse.2007.11.019.

[32] S. G. H. Simis, A. Ruiz-Verdú, J. A. Domínguez-Gómez, R. Peña-Martinez, S. W. M. Peters, and H. J. Gons, "Influence of phytoplankton pigment composition on remote sensing of cyanobacterial biomass," *Remote Sens. Environ.*, vol. 106, no. 4, pp. 414–427, Feb. 2007, doi: 10.1016/j.rse.2006.09.008.

[33] A. Morel and L. Prieur, "Analysis of variations in ocean color," *Limnol. Oceanogr.*, vol. 22, no. 4, pp. 709–722, 1977, doi: 10.4319/lo.1977.22.4.0709.

[34] E. Vangi *et al.*, "The new hyperspectral satellite PRISMA: Imagery for forest types discrimination," *Sensors*, vol. 21, no. 4, p. 1182, 2021, doi: 10.3390/s21041182.

[35] J. Transon, R. d'Andrimont, A. Maugnard, and P. Defourny, "Survey of hyperspectral earth observation applications from space in the sentinel-2 context," *Remote Sens.*, vol. 10, no. 2, p. 157, 2018, doi: 10.3390/rs10020157.

[36] M. Drusch *et al.*, "The fluorescence explorer mission concept—ESA's Earth explorer 8," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 3, pp. 1273–1284, Mar. 2017.

[37] K. Cawse-Nicholson *et al.*, "NASA's surface biology and geology designated observable: A perspective on surface imaging algorithms," *Remote Sens. Environ.*, vol. 257, May 2021, Art. no. 112349, doi: 10.1016/j.rse.2021.112349.

[38] H. R. Gordon *et al.*, "A semianalytic radiance model of ocean color," *J. Geophys. Res., Atmos.*, vol. 93, no. D9, pp. 10909–10924, Sep. 1988.

[39] H. R. Gordon and B. A. Franz, "Remote sensing of ocean color: Assessment of the water-leaving radiance bidirectional effects on the atmospheric diffuse transmittance for SeaWiFS and MODIS intercomparisons," *Remote Sens. Environ.*, vol. 112, no. 5, pp. 2677–2685, May 2008, doi: 10.1016/j.rse.2007.12.010.

[40] G. Kim *et al.*, "Hyperspectral imaging from a multipurpose floating platform to estimate chlorophyll-a concentrations in irrigation pond water," *Remote Sens.*, vol. 12, no. 13, p. 2070, Jun. 2020, doi: 10.3390/rs12132070.

[41] L. Wei, H. Pu, Z. Wang, Z. Yuan, X. Yan, and L. Cao, "Estimation of soil arsenic content with hyperspectral remote sensing," *Sensors*, vol. 20, no. 14, p. 4056, 2020, doi: 10.3390/s20144056.

[42] K. Ryan and K. Ali, "Application of a partial least-squares regression model to retrieve chlorophyll-a concentrations in coastal waters using hyper-spectral data," *Ocean Sci. J.*, vol. 51, no. 2, pp. 209–221, Jun. 2016, doi: 10.1007/s12601-016-0018-8.

[43] Z. Wang, Y. Sakuno, K. Koike, and S. Ohara, "Evaluation of Chlorophyll-a estimation approaches using iterative stepwise elimination partial least squares (ISE-PLS) regression and several traditional algorithms from field hyperspectral measurements in the Seto inland Sea, Japan," *Sensors*, vol. 18, no. 8, p. 2656, Aug. 2018, doi: 10.3390/s18082656.

[44] J. Pyo *et al.*, "An integrative remote sensing application of stacked autoencoder for atmospheric correction and cyanobacteria estimation using hyperspectral imagery," *Remote Sens.*, vol. 12, no. 7, p. 1073, Mar. 2020, doi: 10.3390/rs12071073.

[45] N. Pahlevan *et al.*, "Seamless retrievals of chlorophyll-a from sentinel-2 (MSI) and sentinel-3 (OLCI) in inland and coastal waters: A machine-learning approach," *Remote Sens. Environ.*, vol. 240, Apr. 2020, Art. no. 111604, doi: 10.1016/j.rse.2019.111604.

[46] N. Pahlevan *et al.*, "Hyperspectral retrievals of phytoplankton absorption and chlorophyll-a in inland and nearshore coastal waters," *Remote Sens. Environ.*, vol. 253, Feb. 2021, Art. no. 112200, doi: 10.1016/j.rse.2020.112200.

[47] J. G. Ghatkar, R. K. Singh, and P. Shanmugam, "Classification of algal Bloom species from remote sensing data using an extreme gradient boosted decision tree model," *Int. J. Remote Sens.*, vol. 40, no. 24, pp. 9412–9438, Dec. 2019, doi: 10.1080/01431161.2019.1633696.

[48] R. L. Lucke *et al.*, "Hyperspectral imager for the coastal ocean: Instrument description and first images," *Appl. Opt.*, vol. 50, no. 11, pp. 1501–1516, Apr. 2011, doi: 10.1364/AO.50.001501.

[49] D. Labate *et al.*, "The PRISMA payload optomechanical design, a high performance instrument for a new hyperspectral mission," *Acta Astronautica*, vol. 65, nos. 9–10, pp. 1429–1436, Nov. 2009, doi: 10.1016/j.actaastro.2009.03.077.

[50] E. Spyrakos *et al.*, "Optical types of inland and coastal waters," *Limnol. Oceanogr.*, vol. 63, no. 2, pp. 846–870, 2018, doi: 10.1002/lno.10674.

[51] N. Pahlevan *et al.*, "ACIX-Aqua: A global assessment of atmospheric correction methods for landsat-8 and sentinel-2 over lakes, rivers, and coastal waters," *Remote Sens. Environ.*, vol. 258, Jun. 2021, Art. no. 112366, doi: 10.1016/j.rse.2021.112366.

[52] D. Gurlin, A. A. Gitelson, and W. J. Moses, "Remote estimation of chl-α concentration in turbid productive waters—Return to a simple two-band NIR-red model?" *Remote Sens. Environ.*, vol. 115, no. 12, pp. 3479–3490, Dec. 2011, doi: 10.1016/j.rse.2011.08.011.

[53] W. J. Moses *et al.*, "Estimation of chlorophyll-α concentration in turbid productive waters using airborne hyperspectral data," *Water Res.*, vol. 46, no. 4, pp. 993–1004, Mar. 2012, doi: 10.1016/j.watres.2011.11.068.

[54] L. Li *et al.*, "An inversion model for deriving inherent optical properties of inland waters: Establishment, validation and application," *Remote Sens. Environ.*, vol. 135, pp. 150–166, Aug. 2013, doi: 10.1016/j.rse.2013.03.031.

[55] L. Li, L. Li, and K. Song, "Remote sensing of freshwater cyanobacteria: An extended IOP inversion model of inland waters (IIMIW) for partitioning absorption coefficient and estimating phycocyanin," *Remote Sens. Environ.*, vol. 157, pp. 9–23, Feb. 2015, doi: 10.1016/j.rse.2014.06.009.

[56] M. W. Matthews, S. Bernard, H. Evers-King, and L. Robertson Lain, "Distinguishing cyanobacteria from algae in optically complex inland waters using a hyperspectral radiative transfer inversion algorithm," *Remote Sens. Environ.*, vol. 248, Oct. 2020, Art. no. 111981, doi: 10.1016/j.rse.2020.111981.

[57] M. Matthews and S. Bernard, "Characterizing the absorption properties for remote sensing of three small optically-diverse South African reservoirs," *Remote Sens.*, vol. 5, no. 9, pp. 4370–4404, Sep. 2013, doi: 10.3390/rs5094370.

[58] L. Vanliere and R. D. Gualti, *Restoration and Recoveryof Shallow Eutrophic Lake Ecosystems in The Netherlands, Developments in Hydrobiology*. Norwell, MA, USA: Kluwer, 1992.

[59] R. Sarada, M. G. Pillai, and G. A. Ravishankar, "Phycocyanin from *Spirulina* sp: Influence of processing of biomass on phycocyanin yield, analysis of efficacy of extraction methods and stability studies on phycocyanin," *Process Biochem.*, vol. 34, no. 8, pp. 795–801, 1999.

[60] K. Randolph, J. Wilson, L. Tedesco, L. Li, D. L. Pascual, and E. Soyeux, "Hyperspectral remote sensing of cyanobacteria in turbid productive water using optically active pigments, chlorophyll α and phycocyanin," *Remote Sens. Environ.*, vol. 112, no. 11, pp. 4009–4019, Nov. 2008, doi: 10.1016/j.rse.2008.06.002.

[61] D. E. Stewart and F. H. Farmer, "Extraction, identification, and quantitation of phycobiliprotein pigments from phototrophic plankton," *Limnol. Oceanogr.*, vol. 29, no. 2, pp. 392–397, Mar. 1984, doi: 10.4319/lo.1984.29.2.0392.

[62] Y. Z. Yacobi, J. Köhler, F. Leunert, and A. Gitelson, "Phycocyanin-specific absorption coefficient: Eliminating the effect of chlorophylls absorption," *Limnol. Oceanogr.: Methods*, vol. 13, no. 4, pp. 157–168, 2015, doi: 10.1002/lom3.10015.

[63] E.-A. Nusch, "Comparison of different methods for chlorophyll and phaeopigment determination," *Arch. für Hydrobiol. Ergebnisse der Limnol.*, vol. 14, pp. 14–36, Mar. 1980.

[64] N.-A. Welschmeyer, "Fluorometric analysis of chlorophyll α in the presence of chlorophyll b and pheopigments," *Limnol. Oceanogr.*, vol. 39, pp. 1985–1992, Dec. 1994.

[65] B. J. Speziale, S. P. Schreiner, P. A. Giammatteo, and J. E. Schindler, "Comparison of N,*N*-dimethylformamide, dimethyl sulfoxide, and acetone for extraction of phytoplankton chlorophyll," *Can. J. Fisheries Aquatic Sci.*, vol. 41, no. 10, pp. 1519–1522, Oct. 1984, doi: 10.1139/f84-187.

[66] H. Horváth, A. W. Kovács, C. Riddick, and M. Présing, "Extraction methods for phycocyanin determination in freshwater filamentous cyanobacteria and their application in a shallow lake," *Eur. J. Phycol.*, vol. 48, no. 3, pp. 278–286, Aug. 2013, doi: 10.1080/09670262.2013.821525.

[67] *Manual of Analytical Methods, Major Ions and Nutrients*, ECCC, Burlington, ON, Canada, 1997.

[68] D. P. Sartory and J. U. Grobbelaar, "Extraction of chlorophyll α from freshwater phytoplankton for spectrophotometric analysis," *Hydrobiologia*, vol. 114, no. 3, pp. 177–187, Jul. 1984, doi: 10.1007/BF00031869.

[69] A. Bennett and L. Bogorad, "Complementary chromatic adaptation in a filamentous blue-green alga," *J. Cell Biol.*, vol. 58, no. 2, pp. 419–435, Aug. 1973, doi: 10.1083/jcb.58.2.419.

[70] S. W. Wright, S. W. Jeffrey, and R. F. C. Mantoura, *Phytoplankton Pigments in Oceanography: Guidelines to Modern Methods*. Paris, France: Unesco, 1997.

[71] V. W. F. A. Quesada, "Adaptation of cyanobacteria to the light regime within Antarctic microbial mats," *Int. Vereinigung für Theoretische Und Angew. Limnol.*, vol. 25, no. 2, pp. 960–965, 1003.

[72] C. D. Mobley, "Estimation of the remote-sensing reflectance from above-surface measurements," *Appl. Opt.*, vol. 38, no. 36, p. 7442, Dec. 1999, doi: 10.1364/ao.38.007442.

[73] J. L. Mueller *et al.*, "Radiometric measurements and data analysis protocols," in *Ocean Optics Protocols for Satellite Ocean Color Sensor Validation, Revision 4*, vol. 3, J. L. Mueller, G. S. Fargion, and C. R. Mcclain, Eds. Greenbelt, MD, USA: Goddard Space Flight Space Centre, 2003, pp. 1–78.

[74] A. A. Gitelson *et al.*, "A simple semi-analytical model for remote estimation of chlorophyll-α in turbid waters: Validation," *Remote Sens. Environ.*, vol. 112, no. 9, pp. 3582–3593, Sep. 2008, doi: 10.1016/j.rse.2008.04.015.

[75] X. Quan and E. S. Fry, "Empirical equation for the index of refraction of seawater," *Appl. Opt.*, vol. 34, no. 18, p. 3477, Jun. 1995, doi: 10.1364/ao.34.003477.

[76] A. Morel and S. Maritorena, "Bio-optical properties of oceanic waters: A reappraisal," *J. Geophys. Res.: Oceans*, vol. 106, no. C4, pp. 7163–7180, Apr. 2001, doi: 10.1029/2000jc000319.

[77] S. Hlaing *et al.*, "Assessment of a bidirectional reflectance distribution correction of above-water and satellite water-leaving radiance in coastal waters," *Appl. Opt.*, vol. 51, no. 2, pp. 220–237, 2012, doi: 10.1364/AO.51.000220.

[78] C. E. Binding, A. Zastepa, and C. Zeng, "The impact of phytoplankton community composition on optical properties and satellite observations of the 2017 Western Lake Erie algal Bloom," *J. Great Lakes Res.*, vol. 45, no. 3, pp. 573–586, Jun. 2019, doi: 10.1016/j.jglr.2018.11.015.

[79] J. Wei, Z. Lee, M. Lewis, N. Pahlevan, M. Ondrusek, and R. Armstrong, "Radiance transmittance measured at the ocean surface," *Opt. Exp.*, vol. 23, no. 9, p. 11826, May 2015, doi: 10.1364/oe.23.011826.

[80] V. Vantrepotte, H. Loisel, D. Dessailly, and X. Mériaux, "Optical classification of contrasted coastal waters," *Remote Sens. Environ.*, vol. 123, pp. 306–323, Aug. 2012, doi: 10.1016/j.rse.2012.03.004.

[81] R. A. Vandermeulen, A. Mannino, A. Neeley, J. Werdell, and R. Arnone, "Determining the optimal spectral sampling frequency and uncertainty thresholds for hyperspectral remote sensing of ocean color," *Opt. Exp.*, vol. 25, no. 16, p. A785, Aug. 2017, doi: 10.1364/oe.25.00a785.

[82] Z. Lee, K. Carder, R. Arnone, and M. He, "Determination of primary spectral bands for remote sensing of aquatic environments," *Sensors*, vol. 7, no. 12, pp. 3428–3441, Dec. 2007, doi: 10.3390/s7123428.

[83] A. Wolanin, M. A. Soppa, and A. Bracher, "Investigation of spectral band requirements for improving retrievals of phytoplankton functional types," *Remote Sens.*, vol. 8, no. 10, p. 871, 2016, doi: 10.3390/rs8100871.

[84] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Oct. 2011. [Online]. Available: http://jmlr.org/papers/v12/pedregosa11a.html

[85] H. Wold, "Estimation of principal components and related models by iterative least squares," in *Multivariate Analysis*, P. R. Krishnajah, Ed. New York, NY, USA: Academic, 1966, pp. 391–420.

[86] G. Hanrahan, F. Udeh, and D. G. Patil, "Chemometrics and statistics–multivariate calibration techniques," in *Encyclopedia of Analytical Science*, 2nd ed. Amsterdam, The Netherlands: Elsevier, 2004, pp. 27–32.

[87] P. D. Wentzell and L. V. Montoto, "Comparison of principal components regression and partial least squares regression through generic simulations of complex mixtures," *Chemometrics Intell. Lab. Syst.*, vol. 65, no. 2, pp. 257–279, 2003, doi: 10.1016/S0169-7439(02)00138-7.

[88] S. Wold, M. Sjöström, and L. Eriksson, "PLS-regression: A basic tool of chemometrics," *Chemometrics Intell. Lab. Syst.*, vol. 58, no. 2, pp. 109–130, 2001, doi: 10.1016/S0169-7439(01)00155-1.

[89] K. Kawamura, N. Watanabe, S. Sakanoue, and Y. Inoue, "Estimating forage biomass and quality in a mixed sown pasture based on partial least squares regression with waveband selection," *Grassland Sci.*, vol. 54, no. 3, pp. 131–145, Sep. 2008, doi: 10.1111/j.1744-697x.2008.00116.x.

[90] K. C. Flynn, A. E. Frazier, and S. Admas, "Nutrient prediction for tef (Eragrostis tef) plant and grain with hyperspectral data and partial least squares regression: Replicating methods and results across environments," *Remote Sens.*, vol. 12, no. 18, p. 2867, Sep. 2020, doi: 10.3390/rs12182867.

[91] P. Sinha *et al.*, "The potential of *in-situ* hyperspectral remote sensing for differentiating 12 banana genotypes grown in Uganda," *ISPRS J. Photogramm. Remote Sens.*, vol. 167, pp. 85–103, Sep. 2020, doi: 10.1016/j.isprsjprs.2020.06.023.

[92] A. L. Robertson, L. Li, L. Tedesco, J. Wilson, and E. Soyeux, "Using a partial least squares (PLS) method for estimating cyanobacterial pigments in eutrophic inland waters," *Proc. SPIE*, vol. 7454, Aug. 2009, Art. no. 745408, doi: 10.1117/12.824632.

[93] K. Song, L. Li, S. Li, L. Tedesco, B. Hall, and Z. Li, "Hyperspectral retrieval of phycocyanin in potable water sources using genetic algorithm–partial least squares (GA–PLS) modeling," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 18, no. 1, pp. 368–385, Aug. 2012, doi: 10.1016/j.jag.2012.03.013.

[94] K. Song, D. Lu, L. Li, S. Li, Z. Wang, and J. Du, "Remote sensing of chlorophyll-α concentration for drinking water source using genetic algorithms (GA)-partial least square (PLS) modeling," *Ecol. Informat.*, vol. 10, pp. 25–36, Jul. 2012, doi: 10.1016/j.ecoinf.2011.08.006.

[95] K. Song *et al.*, "Remote estimation of chlorophyll-α in turbid inland waters: Three-band model versus GA-PLS model," *Remote Sens. Environ.*, vol. 136, pp. 342–357, Sep. 2013, doi: 10.1016/j.rse.2013.05.017.

[96] K. Song *et al.*, "Using partial least squares-artificial neural network for inversion of inland water chlorophyll-α," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 2, pp. 1502–1517, Feb. 2014, doi: 10.1109/TGRS.2013.2251888.

[97] K. Song, L. Li, L. P. Tedesco, S. Li, B. E. Hall, and J. Du, "Remote quantification of phycocyanin in potable water sources through an adaptive model," *ISPRS J. Photogramm. Remote Sens.*, vol. 95, pp. 68–80, Sep. 2014, doi: 10.1016/j.isprsjprs.2014.06.008.

[98] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995, doi: 10.1007/BF00994018.

[99] H. Drucker, C. J. C. Burges, L. Kaufman, A. Smola, and V. Vapnik, "Support vector regression machines," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 9, 1997, pp. 155–161. [Online]. Available: https://proceedings.neurips.cc/paper/1996/file/d38901788c533e8286cb6400b40b386d-Paper.pdf

[100] E. J. Kwiatkowska and G. S. Fargion, "Application of machine-learning techniques toward the creation of a consistent and calibrated global chlorophyll concentration baseline dataset using remotely sensed ocean color data," *IEEE Trans. Geosci. Remote Sens.*, vol. 41, no. 12, pp. 2844–2860, Dec. 2003, doi: 10.1109/TGRS.2003.818016.

[101] A. B. Ruescas, M. Hieronymi, G. Mateo-Garcia, S. Koponen, K. Kallio, and G. Camps-Valls, "Machine learning regression approaches for colored dissolved organic matter (CDOM) retrieval with S2-MSI and S3-OLCI simulated data," *Remote Sens.*, vol. 10, no. 5, p. 786, 2018, doi: 10.3390/rs10050786.

[102] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, vols. 13–17, Aug. 2016, pp. 785–794.

[103] L. Wu and J. Fan, "Comparison of neuron-based, kernel-based, tree-based and curve-based machine learning models for predicting daily reference evapotranspiration," *PLoS ONE*, vol. 14, no. 5, May 2019, Art. no. e0217520, doi: 10.1371/journal.pone.0217520.

[104] Z. Cao *et al.*, "A machine learning approach to estimate chlorophyll-α from Landsat-8 measurements in inland lakes," *Remote Sens. Environ.*, vol. 248, Oct. 2020, Art. no. 111974, doi: 10.1016/j.rse.2020.111974.

[105] F. Rosenblatt, *Principles of Neurodynamics: Perceptions and the Theory of Brain Mechanisms*. Washington, DC, USA: Spartan, 1962.

[106] H. Schiller and R. Doerffer, "Neural network for emulation of an inverse model operational derivation of case II water properties from MERIS data," *Int. J. Remote Sens.*, vol. 20, no. 9, pp. 1735–1746, Jan. 1999, doi: 10.1080/014311699212443.

[107] R. Doerffer and H. Schiller, "The MERIS case 2 water algorithm," *Int. J. Remote Sens.*, vol. 28, nos. 3–4, pp. 517–535, Feb. 2007, doi: 10.1080/01431160600821127.

[108] L. González Vilas, E. Spyrakos, and J. M. Torres Palenzuela, "Neural network estimation of chlorophyll a from MERIS full resolution data for the coastal waters of Galician rias (NW Spain)," *Remote Sens. Environ.*, vol. 115, no. 2, pp. 524–535, Feb. 2011, doi: 10.1016/j.rse.2010.09.021.

[109] D. Odermatt *et al.*, "MERIS observations of phytoplankton Blooms in a stratified eutrophic lake," *Remote Sens. Environ.*, vol. 126, pp. 232–239, Nov. 2012, doi: 10.1016/j.rse.2012.08.031.

[110] S. C. J. Palmer *et al.*, "Validation of ENVISAT MERIS algorithms for chlorophyll retrieval in a large, turbid and optically-complex shallow lake," *Remote Sens. Environ.*, vol. 157, pp. 158–169, Feb. 2015, doi: 10.1016/j.rse.2014.07.024.

[111] A. Castagna, S. Simis, H. Dierssen, Q. Vanhellemont, K. Sabbe, and W. Vyverman, "Extending landsat 8: Retrieval of an orange *contra*-band for inland water quality applications," *Remote Sens.*, vol. 12, no. 4, p. 637, 2020, doi: 10.3390/rs12040637.

[112] C. Liu, S. Q. Yin, M. Zhang, Y. Zeng, and J. Y. Liu, "An improved grid search algorithm for parameters optimization on SVM," *Appl. Mech. Mater.*, vols. 644–650, pp. 2216–2219, Sep. 2014, doi: 10.4028/www.scientific.net/AMM.644-650.2216.

[113] I. Syarif, A. Prugel-Bennett, and G. Wills, "SVM parameter optimization using grid search and genetic algorithm to improve classification performance," *Telkomnika*, vol. 14, no. 4, pp. 1502–1509, 2016, doi: 10.12928/TELKOMNIKA.v14i4.3956.

[114] Y. Sun, S. Ding, Z. Zhang, and W. Jia, "An improved grid search algorithm to optimize SVR for prediction," *Soft Comput.*, vol. 25, no. 7, pp. 5633–5644, Apr. 2021, doi: 10.1007/s00500-020-05560-w.

[115] D. C. Liu and J. Nocedal, "On the limited memory BFGS method for large scale optimization," *Math. Program.*, vol. 45, no. 3, pp. 503–528, 1989.

[116] L. Bottou, "Stochastic gradient descent tricks," in *Neural Networks: Tricks of the Trade*, 2nd ed., G. Montavon, G. B. Orr, and K.-R. Müller, Eds. Berlin, Germany: Springer, 2012, pp. 421–436.

[117] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Represent. (ICLR)*, 2015, pp. 1–15.

[118] J. F. Schalles and Y. Z. Yacobi, "Remote detection and seasonal patterns of phycocyanin, carotenoid and chlorophyll pigments in eutrophic waters," *Ergebnisse Der Limnol.*, vol. 55, pp. 153–168, Feb. 2000.

[119] S. Mishra, "Remote sensing of harmful algal bloom," Ph.D. dissertation, Mississippi State Univ., Starkville, MS, USA, 2012.

[120] S. K. Morley, T. V. Brito, and D. T. Welling, "Measures of model performance based on the log accuracy ratio," *Space Weather*, vol. 16, no. 1, pp. 69–88, Jan. 2018, doi: 10.1002/2017SW001669.

[121] V. Sagan *et al.*, "Monitoring inland water quality using remote sensing: Potential and limitations of spectral indices, bio-optical simulations, machine learning, and cloud computing," *Earth-Sci. Rev.*, vol. 205, Jun. 2020, Art. no. 103187, doi: 10.1016/j.earscirev.2020.103187.

[122] B. Smith *et al.*, "A chlorophyll-α algorithm for Landsat-8 based on mixture density networks," *Frontiers Remote Sens.*, vol. 1, p. 5, Feb. 2021, doi: 10.3389/frsen.2020.623678.

[123] J. Pyo *et al.*, "A convolutional neural network regression for quantifying cyanobacteria using hyperspectral imagery," *Remote Sens. Environ.*, vol. 233, Nov. 2019, Art. no. 111350, doi: 10.1016/j.rse.2019.111350.

[124] A. F. Al-Anazi and I. D. Gates, "Support vector regression to predict porosity and permeability: Effect of sample size," *Comput. Geosci.*, vol. 39, pp. 64–76, Feb. 2012, doi: 10.1016/j.cageo.2011.06.011.

[125] A. Gilerson *et al.*, "Bidirectional reflectance function in coastal waters: Modeling and validation," *Proc. SPIE*, vol. 8175, Oct. 2011, Art. no. 81750O, doi: 10.1117/12.898449.

[126] D. Vansteenwegen, K. Ruddick, A. Cattrijsse, Q. Vanhellemont, and M. Beck, "The pan-and-tilt hyperspectral radiometer system (PANTHYR) for autonomous satellite validation measurements—Prototype design and testing," *Remote Sens.*, vol. 11, no. 11, p. 1360, 2019, doi: 10.3390/rs11111360.