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 detection 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 \( (R) \) 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 \(< 65\%\) )

1. INTRODUCTION

Cyanobacterial harmful algal blooms (Cyanobacteria 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 Cyanobacteria HAB monitoring and assessment is challenging. Furthermore, due to limited capabilities and resources available to agencies, there is no routine assessment and monitoring of Cyanobacteria 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...
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 (Chla) 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), Chla 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 Chla, cyanobacteria blooms, surface scum, and floating vegetation in coastal and inland waters. This baseline subtraction procedure calculates 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 (665), with bands located at 620, 665, and 681 nm) as an exclusion criterion.
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 Chla 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 Chla 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., Chlb, Chlc, 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 PRercurso 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 Chla from hyperspectral data in Long Bay, SC, USA [42]. Ryan and Ali [42] used PLSR to identify spectral bands that are more sensitive to Chla 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 Chla 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 Chla. Pahlevan et al. [45] introduced a mixture density network (MDN) to estimate Chla 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\textsubscript{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 Cyanobacterial 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\textsuperscript{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\textsubscript{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/F 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\textsubscript{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 Chla, 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 Chla 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]. Chla 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 illustrated in Fig. 1 (middle). The bottom plots in Fig. 1 show the log distribution of PC to Chla ratio (PC:Chla). This ratio indicates the presence and abundance of cyanobacteria relative to total phytoplankton biomass [22]. Cyanobacteria is dominant when PC:Chla ≥ 1. PC and Chla 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, 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

$$R_s = \frac{L_w}{E_d(0^+)}.$$

The in situ measurements of $R_s$ 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, $L_w(0^-)$, 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 $L_w = t L_w(0^-) n^{-2}$ [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 $R_s$ [52].

This closely resembled the measurements in the three central Indiana reservoirs where dual Ocean Optics USB4000 UV-NIR spectrometers (Ocean Insight) were employed to measure underwater remote sensing reflectance, $r_{s}$, 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 $L_w(0^-)$ 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, $r'_{s}$, using the CALMIT Data Acquisition Program software (CDAP; University of Nebraska at Lincoln)

$$r'_{s} = \frac{L_d(0^-)}{E_d(0^+)}.$$

Furthermore, $r_s$ is defined as $L_w(0^-)$ divided by the total downwelling irradiance just beneath the water surface, $E_d(0^-)$,
Fig. 1. Log distribution of (Top) PC, and (Middle) Chlα concentrations and (Bottom) PC: Chlα for each study site and all data combined. Statistics of PC (mg/m³), Chlα (mg/m³), and PC:Chlα 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:Chlα ≥ 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 $E_d(0^\circ) = 0.965E_d(0^\circ)$ [76] is made

$$r_n = \frac{r_{rs}^n}{0.965}. \quad (3)$$

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 $R_{rw}$ in Lake Erie during the weekly field sampling efforts, with a Satlantic Hypergun with radiance values at 137 channels (350–800 nm). The Hypgun measures upwelling radiance ($L_u$) at 150° relative to the solar azimuth [72], then at 40° from nadir at 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 $\pi$ [72], [73]. Water leaving radiance ($L_{rw}$) was corrected for diffuse sky contamination b: $L_{rw} = L_u - 0.028 \times L_{sky}$ [72], where 0.028 is taken to be the reflectivity of the water surface. The remote sensing reflectance ($R_{rs}$) was calculated...
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 ECC 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 $L_u(0^+)$ 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. Spyarakos 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
continuum of various OWTs encountered in aquatic environments and avoid under-representing OWTs in their matchup dataset.

To assign an HICO-like $R_s$ spectrum to one of the predefined OWTs, the $R_s$ 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 $R_s$ 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 $R_s$ bands in this study, with $\sim$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 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 and PC with in situ measured spectra [93], [94] leading to improved accuracies compared with three band models particularly for estimating Chl [95]. PLS was also coupled with ANN to model possible nonlinear relationships between cyanobacterial pigments Chl 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-$R_s$ 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 Chla from OLI in eight turbid lakes in eastern China.
ing and testing ML algorithms resulted in a total of 724 ran-
rithms consists solely of HICO-simulated hyperspectral bands.
that were used as input features in the best performing ML algorithm. The spectral sampling and 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 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

\[ W_i = \frac{StDev_i}{\sum_i StDev_i} \quad \text{where } i \text{ is } i\text{-th band.} \] (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.

\[ W_i = \frac{StDev_i}{\sum_i StDev_i} \quad \text{where } i \text{ is } i\text{-th band.} \] (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.

\[ W_i = \frac{StDev_i}{\sum_i StDev_i} \quad \text{where } i \text{ is } i\text{-th band.} \] (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.

\[ W_i = \frac{StDev_i}{\sum_i StDev_i} \quad \text{where } i \text{ is } i\text{-th band.} \] (4)

XGBoost: To determine the structure of the XGBoost model, six hyperparameters including, alpha, gamma, the number of trees (n_estimators), 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
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 (0.0001, 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

$$SSPB = 100 \text{sign}(z)(10^{\left| z \right|} - 1)[\%]$$

where $z = \text{Median}\left[\log_{10}\left(\frac{E}{M}\right)\right]$ (5)

$$\text{MdSA} = 100(10^y - 1)[\%]$$

where $y = \text{Median}\left[\log_{10}\left(\frac{E}{M}\right)\right]$ (6)

$$\text{MSA} = 100(10^y - 1)[\%]$$

where $y = \text{Mean}\left[\log_{10}\left(\frac{E}{M}\right)\right]$ (7)

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

(8)

$$\text{MAPE} = \frac{1}{n} \sum_{i=1}^{n} \frac{|M_i - E_i|}{M_i}$$

(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.

| Reference | Algorithm | Name |
|-----------|-----------|------|
| [118]     | $PC = 0.33 \times \frac{R_{rs}(650)}{R_{rs}(625)} + 0.26$ | S00  |
| [2]       | $PC = 0.30 \times \frac{R_{rs}(700)}{R_{rs}(660)} + 0.29$ | M09  |
| [23]      | $PC = 0.16 \times [R_{rs}^{2}(615) - R_{rs}^{2}(660)] \times R_{rs}(725) + 0.24$ | H10  |
| [119]     | $PC = 0.29 \times \frac{R_{rs}(709)}{R_{rs}(660)} + 0.28$ | M12  |

![Fig. 4. Average of spectra in each OWT.](image)
Fig. 5. Distribution of $R_s$ in each OWT. Mean and all $R_s$ are, respectively, shown in black solid line and dashed lines (colored by the study site) in each OWT. The nonstandardized HICO $R_s$ values are shown in this figure.

![Fig. 5](image)

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 |

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 $R_s$ 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. 5](image)

Fig. 6. Ranges of PC and Chl$\alpha$ 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.

Fig. 6 shows the average and range of PC and Chl$\alpha$ values in our dataset per OWT. OWT1 ($N = 36$) has the lowest PC and Chl$\alpha$ values of 0.51 ± 0.69 and 4.1 ± 2.02 mg/m$^3$, respectively. The highest PC and Chl$\alpha$ 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 $R_s$ 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$\alpha$ 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 |
| **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 | 0.63* |
|        | **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 $R_s$ 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 OWT 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 $R_s$ 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.

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 $R_s$ 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...
Concentrations of Chla 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 Chla 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 $R_\text{s}$ 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 $R_\text{s}$ 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 OWTS, 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 Chla (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-PRISMA produced the lowest SSPB and MdSA in OWT3, but MLP-HICO outperformed the rest with the lowest MdSA. 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 Chla values (100.2 ± 150.87 and 67.59±55.14 mg/m$^3$, respectively). In OWTS, 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:Chla. 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:Chla values are less than one. The performance of MLP-OLI was consistently lower than that of other sensors in this range of PC:Chla values. In the presence of cyanobacteria (PC:Chla ≥ 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
relatively high concentrations. Metsamaa et al. [15] show that the cyanobacteria double spectral feature can be detectable when Chl in the Baltic Sea is at least 8–10 mg/m³. 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 ≥ 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 Rrs 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 Rrs 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 Rrs at a specific wavelength (or equally high correlations in a range of the spectrum) that is not necessarily covered by the multispectral
bands. For example, the highest correlation between $R$ 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 $\sim 2\%$ 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 $R$ 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 $R$.

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 ($\sim 20$-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 $\sim 400$.

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 $\sim 600$, 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 $R$. 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...
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 $\sim$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 Chla (i.e., OWTs 1 and 2) and also highest in PC and Chla (i.e., OWT4), and when cyanobacteria were not dominant (PC:Chla < 1). MLP applied to LNext performed best between other multispectral scenarios in OWT1. When cyanobacteria were dominant (PC:Chla $\geq$ 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 ($\sim$620 nm) for OWTs 3, 4, and 5. The longer wavelengths in the spectrum ($\sim$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. 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.
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. Curnichiel 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. Fahnentiel, “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.
This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.
