Remote Sensing of Environment 248 (2020) 111974

Contents lists available at ScienceDirect

Remote Sensing of Environment

journal homepage: www.elsevier.com/locate/rse

A machine learning approach to estimate chlorophyll-*a* from Landsat-8 measurements in inland lakes

Zhigang Cao[a,b], Ronghua Ma[a,*], Hongtao Duan[a], Nima Pahlevan[c], John Melack[d], Ming Shen[a,b], Kun Xue[a]

[a] Key Laboratory of Watershed Geographic Sciences, Nanjing Institute of Geography and Limnology, Chinese Academy of Sciences, Nanjing 210008, China
[b] University of Chinese Academy of Sciences, Beijing 100049, China
[c] Science Systems and Applications Inc. 10210 Greenbelt Rd. Suite 600, Lanham, MD 20706, USA
[d] Bren School of Environmental Science and Management, University of California, Santa Barbara, CA 93106, USA

ARTICLE INFO

Keywords:
Landsat
Machine learning
Eutrophication
Lakes

ABSTRACT

Landsat-8 Operational Land Imager (OLI) provides an opportunity to map chlorophyll-*a* (Chl*a*) in lake waters at spatial scales not feasible with ocean color missions. Although state-of-the-art algorithms to estimate Chl*a* in lakes from satellite-borne sensors have improved, there are no robust and reliable algorithms to generate Chl*a* time series from OLI imageries in turbid lakes due to the absence of a red-edge band and issues with atmospheric correction. Here, a machine learning approach termed the extreme gradient boosting tree (BST) was employed to develop an algorithm for Chl*a* estimation from OLI in turbid lakes. This model was developed and validated by linking Rayleigh-corrected reflectance to near-synchronous *in situ* Chl*a* data available from eight lakes in eastern China (N = 225) and three coastal and inland waters in SeaWiFS Bio-optical Archive and Storage System (N = 97). The BST model performed well on a subset of data (N = 102, R^2 = 0.79, root mean squared difference = 7.1 μg L^{-1}, mean absolute percentage error = 24%, mean absolute error = 1.4, Bias = 0.9), and had better Chl*a* retrievals than several band-ratio algorithms and a random forest approach. The performance of BST model was judged as appropriate only for the range of conditions in the training data. Given these limitations, spatial and temporal variations of Chl*a* in hundreds of lakes larger than 1 km^2 in eastern China for the period of 2013–2018 were mapped using the BST model. OLI-derived Chl*a* indicated that small lakes (< 50 km^2) had greater Chl*a* than the larger lakes. This research suggests that machine-learning models provide practical approaches to estimate Chl*a* in turbid lakes *via* broadband instruments like OLI and that extending to other regions requires training with a representative dataset.

1. Introduction

Lake ecosystems vary widely in their trophic status and have experienced significant degradation largely as a result of intensified human activities (Adrian et al., 2009). For example, a recent satellite-derived analysis of a trophic state index for large lakes (> 10 km^2) worldwide suggested that 63% were eutrophic, 26% mesotrophic, and 11% oligotrophic in the summer of 2012 (Wang et al., 2018). Phytoplankton biomass serves as an important water quality parameter because abundance of algae can reflect the state of eutrophication and aid the evaluation of health risks of aquatic ecosystems (Hunter et al., 2009). In practice, phytoplankton biomass is typically quantified as the concentration of chlorophyll-a (Chl*a*), a photosynthetic pigment presented in all algal species.

Chl*a* in lakes has been measured from space using ocean color instruments, including the Moderate Resolution Imaging Spectroradiometer (MODIS, Terra:1999~, Aqua:2002~) (Dall'Olmo et al., 2005; Qi et al., 2014), Medium Resolution Imaging Spectrometer (MERIS, 2003–2012) (Gilerson et al., 2010; Sayers et al., 2015), and Ocean and Land Color Instrument (OLCI, 2016~) (Kravitz et al., 2020; Smith et al., 2018). However, the coarse spatial resolution of ocean color instruments precludes their applications to small and medium sized lakes and reservoirs. In fact, lakes less than 100 km^2 account for 63% of the global lake area (Downing et al., 2006); hence, medium and small lakes constitute the important component of global inland waters. Remote sensing of medium to small lakes requires moderate (~300 m) to high (10– 30 m) spatial resolution sensors, such as observations available through the Landsat series, Sentinel-2 A/B, and the Satellite

* corresponding author.
E-mail address: rhma@niglas.ac.cn (R. Ma).

https://doi.org/10.1016/j.rse.2020.111974
Received 21 October 2019; Received in revised form 29 June 2020; Accepted 30 June 2020

Pour l'Observation de la Terre (SPOT).

The Operational Land Imager (OLI) onboard Landsat-8 launched in 2013 is a 12 bit push-broom sensor with moderate spatial resolution (30 m), with a band near 443 nm for coastal and inland applications, and significantly improved signal-to-noise ratio (SNR) (Pahlevan et al., 2014). Although OLI has limited spectral resolution compared to ocean color instruments, the improved spatial resolution and radiometric performance make OLI appropriate for application to inland waters (Pahlevan et al., 2017b). Furthermore, the OLI could be utilized to generate a harmonized dataset together with the Multi-Spectral Instrument (MSI) on Sentinel-2A and 2B (Pahlevan et al., 2019). At present, several atmospheric corrections for OLI over coastal and inland waters have been implemented and demonstrated acceptable accuracy in some cases (Franz et al., 2015; Vanhellemont and Ruddick, 2018). However, the uncertainties in remote sensing reflectance (R_{rs}) derived through these approaches remain large in eutrophic and turbid lakes, primarily due to difficulties estimating aerosol signals across the visible spectrum (Pahlevan et al., 2017a; Wang and Jiang, 2018). Accuracy of atmospheric correction largely limits the development and application of existing algorithms for Chl*a* estimation in lakes.

Most of existing algorithms for Chl*a* retrieval were developed for *in situ*, airborne hyperspectral, and MERIS/OLCI data (Table 1), since the red-edge band near 700–710 nm is critical for Chl*a* estimation in turbid waters (Gitelson, 1992). However, these algorithms cannot be applied to the OLI data owing to the absence of observations in this critical channel. Several empirical algorithms have been developed for Chl*a* estimation with OLI, and include the band ratio, spectral matching technique (SMT), and spectral index (Concha and Schott, 2016; Page et al., 2018). However, empirical algorithms developed for fairly clear lakes (Concha and Schott, 2016) and eutrophic lakes (Ha et al., 2017; Kuhn et al., 2019; Watanabe et al., 2018) were only applied to field R_{rs} data or tested on a few OLI images. Considering the value of a long-term measurements of Chl*a* to examine eutrophication in lakes, a reliable algorithm to obtain the time series of Chl*a* in turbid lakes from OLI images would be beneficial.

In recent years, machine learning has been used to investigate oceanic, coastal, and inland water environments (Reichstein et al., 2019). For example, several approaches, including the deep neural network (DNN) (Cao et al., 2019a), convolutional neural network (Pyo et al., 2019), mixture density networks (Pahlevan et al., 2020), and random forest (RF) (Chen et al., 2019), have been used to derive absorption coefficients, Chl*a*, and cyanobacteria concentration. Machine learning can utilize complex networks and structure to capture data-rich features of the input data and to obtain explicit relationships with the output variable (Pyo et al., 2019). Hence, this approach provides a way to estimate Chl*a* in lakes from the OLI imagery by compensating for the limited spectral bands. However, scalability of machine learning approaches to regions with bio-optical characteristics different from those used in the training phase has been limited due to their empirical nature (Palmer et al., 2015).

Recognizing the lack of an algorithm to generate time series of Chl*a* in lakes from OLI images, this research develops a machine-learning algorithm for Chl*a* estimation using OLI-derived Rayleigh-corrected reflectance and synchronous Chl*a*, with improved model performances over those in the literature. Through this algorithm, we aim to produce regional, long-term, and high-spatial resolution Chl*a* products in hundreds of lakes. First, an overview of the study area is presented, followed by a description of the methods of acquisition and processing of *in situ* measurements and satellite data. Second, the principles of the machine-learning method, the training dataset preparation, and the model development and evaluation are provided. Then, the model is applied to retrieve Chl*a* concentrations in lakes larger than 1 km^2 in eastern China to assess their spatial and temporal variability. Finally, the physical mechanisms, strengths and limitations, and broader implications of the model are discussed.

2. Material and methods

2.1. Study area

Lakes in the middle and lower reach of Yangtze River (MLRY) and a reach of Huai River (RHR) in eastern China were selected (27.78°-36.10°N, 111.70°-121.67°E, Fig. 1). MLRY and RHR reaches include 605 natural lakes larger than 1 km^2 (not including reservoirs) with 473 lakes less than 10 km^2, and a total lake area of 20,762 km^2 (Ma et al., 2011). The lakes distributed throughout the region include river-connected lakes (such as Poyang, Dongting, and Hongze) and lakes without direct interaction with the main rivers (such as Taihu and Chaohu) (Fig. 1). The lakes are located in a region with a subtropical monsoon climate and are shallow with average depths from 1.1 to 8.4 m. Overall, suspended particulate matter (SPM) concentrations in these lakes ranges between 1 and 300 mg L^{-1}, and resuspension events induced by high winds further increase turbidity (Hou et al., 2017). The area has one of the fastest growing economies in China, and the aquatic

Table 1
Examples of Chl*a* retrieval algorithms and their application areas.

Algorithm form	Reference	Application area	Data source
Semi-analytical method	Gons (1999)	8 lakes and estuaries	*In situ*
$R_{rs}(709)/R_{rs}(665) = > a_{ph}(665)$	Gons et al. (2008)	Great Lakes, USA	MERIS
$= >$ Chl*a*	Duan et al. (2012)	3 eutrophic lakes, China	*In situ*
Simple band ratio	Dekker and Peters (1993)	10 lakes in Netherlands	*In situ*
Chl*a* $= f(R_{rs}(709)/R_{rs}(665))$	Ruddick et al. (2001)	Netherland and Belgian waters	*In situ*
	Duan et al. (2012)	3 eutrophic lakes, China	*In situ*
Multi band algorithm	Dall'Olmo et al. (2005)	4 lakes and reservoirs, USA	*In situ*
$[R_{rs}(671)^{-1}\text{-}R_{rs}(710)^{-1}] \times R_{rs}(740)$	Gitelson et al. (2008)	Lakes and reservoirs, USA	*In situ*
$[R_{rs}(662)^{-1}\text{-}R_{rs}(693)^{-1}]\times$ $[R_{rs}(740)^{-1}\text{-}R_{rs}(705)^{-1}]$	Le et al. (2009)	Lake Taihu, China	*In situ*
Spectral index	Mishra and Mishra (2012)	4 estuaries and bays, USA	MERIS,
NDCI, NGRDI	Feng et al. (2014)	Lake Poyang, China	MERIS
Normalized spectral index	Shi et al. (2015)	Lake Taihu, China	MODIS
Other methods	Craig et al. (2012)	Compass Buoy station, Canada	*In situ*
Empirical Orthogonal Function	Qi et al. (2014)	Lake Taihu, China	MODIS
Machine learning	Pahlevan et al., (2020)	Global areas	Sentinel-2 and 3
Landsat algorithms	Giardino et al. (2001)	Lake Iseo, Italy	Landsat 5
Band ratios,	Duan et al. (2007)	Lake Chagan, China	Landsat 5
Band difference	Watanabe et al. (2018)	Barra Bonita Reservoir, Brazil	Landsat 8
	Ha et al. (2017)	Lake West, Vietnam	Landsat 8
Spectral matching method	Concha and Schott (2016)	Rochester Embayment, USA	Landsat 8
Neural network	Prasad et al. (2020)	Ganga River, India	Landsat 8

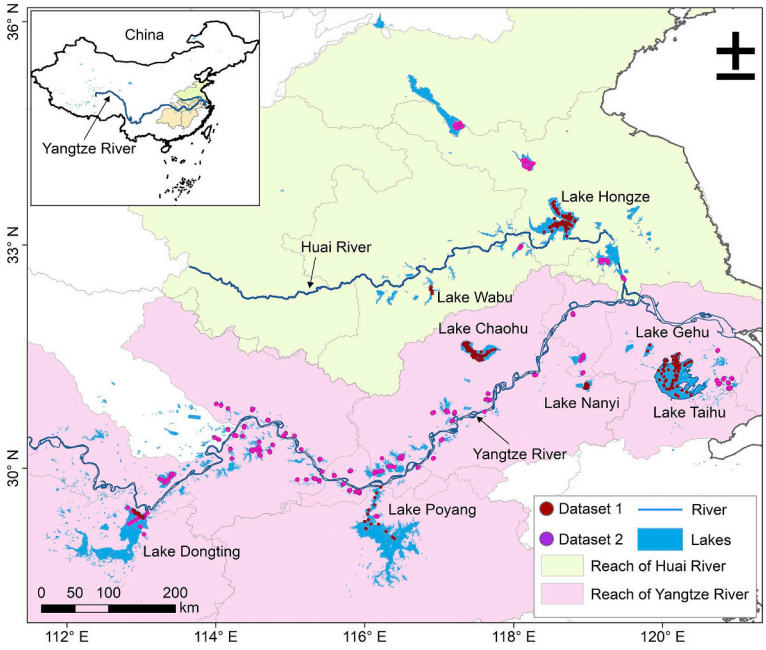

Fig. 1. Spatial distributions of the 605 lakes in the middle and lower reach of Yangtze River (MLYR) and the reach of Huai River (RHR). The colored circles represent the sampled sites, the red points are Dataset 1, and the purple points are Dataset 2. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

ecosystems have deteriorated due to human activities (Feng et al., 2019). All lakes in the region are mesotrophic or eutrophic, and the frequent occurrence of cyanobacteria blooms in Taihu and Chaohu has seriously threatened the drinking water obtained from the lakes (Duan et al., 2009).

2.2. Field data

We used three types of field data: field surveys (N = 522), SeaBASS data (N = 97), and monthly measurements in Lake Taihu (N = 180) by the Taihu Laboratory for Lake Ecosystem Research (TLLER, http://taihu.niglas.cas.cn). Our sampling includes remote sensing reflectance, absorption coefficients, Chl*a*, and SPM concentrations; other datasets only have Chl*a* concentrations. Some of data from our field surveys (N = 225) and SeaBASS data (N = 97) concurrent with the Landsat-8 measurements were used to develop and validate the algorithm for Chl*a* estimations, and monthly time series of Chl*a* in Lake Taihu were used to compare to the OLI-derived Chl*a* as an independent dataset.

2.2.1. In situ data

A total of 522 *in situ* samples were collected from 67 lakes in the MLYR and RHR during the 2013–2018 period (Fig. 1). Two comprehensive surveys in August 2012 and October 2017 were done. We divided our field data into two categories. Dataset 1 (DS1) (225 samples) is the near-coincident data obtained under Landsat-8 OLI overpasses at eight lakes (Fig. 1; Table 2) used to develop the algorithm for Chl*a* estimation. The details for the generation of Dataset-1 are explained in section 2.4.1. Dataset 2 (DS2) is the remaining dataset (297 samples) used to analyze the bio-optical properties of the lakes together with DS1.

At each field site, remote sensing reflectance (R_{rs}) ranging from 350 nm to 1050 nm with an interval of 1 nm was measured with a field spectrometer (FieldSpec 4 Hi-Res spectroradiometer, Analytical Spectra Devices, Inc.) (Mueller et al., 2003). We measured the spectrum at a 135° azimuth with respect to the sun and with a nadir viewing angle of 45°. The total water leaving radiance (L_{sw}), the radiance of reference gray panel (L_p), and the sky radiance (L_{sky}) were measured to estimate the R_{rs} following: $R_{rs} = (L_{sw}-\rho^{*}L_{sky})^{*}\rho_p/\pi L_p$, where ρ is the Fresnel reflectance and assumed to be 0.028 based on the wind speed and sky conditions from field measurements (Mobley, 1999). ρ_p is the reflectance of the gray panel (30%). Surface water samples (depth: ~0.50 m) were collected using a 2-L polyethylene water-sampler and stored in the dark and kept cool before the concentrations of constituents and the absorption coefficients were measured in the laboratory.

A glass fiber filter (pore size: 0.70 μm, diameter 47 mm, Whatman GF/F) was used to filter the water and was subsequently soaked in 90% acetone to extract the pigments (Jeffrey and Humphrey, 1975). The light absorbance of the extracted solution was measured at 630(A_{630}), 645(A_{645}), 663(A_{663}), and 750 nm(A_{750}) using a UV2600

Table 2
Description of lake area, samples number, and sampled dates for Dateset-1.

Lake	Area (km^2)	Number of samples	Date
Lake Taihu	2425	64	2013/03/25, 2013/12/19, 2015/12/16, 2017/03/08, 2017/05/11, 2017/05/27, 2017/12/05, 2017/12/21
Lake Chaohu	769.6	47	2013/05/14, 2013/06/15, 2014/09/03, 2014/09/22, 2015/10/11, 2016/01/15, 2017/11/01, 2018/06/13
Lake Hongze	1576.9	39	2014/10/24, 2016/03/28, 2016/12/16, 2018/08/25
Lake Nanyi	148.4	12	2018/10/28
Lake Gehu	146.5	3	2018/07/17
Lake Wabu	163.0	5	2018/07/15
Lake Dongting	2432.5	23	2016/07/23
Lake Poyang	2933.0	32	2013/10/12, 2014/01/09, 2014/07/20, 2014/10/15 2015/07/04, 2015/10/11, 2016/07/09
Total	/	225	/

spectrophotometer (Lorenzen, 1967). Chl*a* concentration was calculated using Eq. (1) (Jeffrey and Humphrey, 1975).

$$Chla = (C_1 \times (A_{663} - A_{750}) + C_2 \times (A_{645} - A_{750}) + C_3 \times (A_{630} - A_{750})) \times V_e/V_f \times l_c \quad (1)$$

where, C_1-C_3 are constants, V_e is the volume of extract in milliliters (10 mL), V_f is the volume of lake water filtered in liters (0.05–0.20 L), and l_c is the path length of the cuvette in centimeters (1 cm). Here, C_1 = 11.64, C_2 = -2.16, C_3 = 0.1, which were determined by the Chinese Environmental Protection Administration using water samples from Chinese lakes.

Water samples (50– 200 mL) were filtered through pre-combusted and pre-weighed Whatman GF/F filters, subsequently dried at 105 °C for 4 h to determine SPM concentrations gravimetrically (Cao et al., 2017). The SPM was further differentiated into suspended inorganic matter (SPIM) and suspended organic matter (SPOM) by combusting the organic matter from the filters at 450 °C for 4 h and reweighing the filters. The absorption coefficient of CDOM [$a_g(\lambda)$] at 280–700 nm was measured in filtered water (Millipore filter, pore size: 0.22 μm) in a quartz cuvette with a light path of 1 cm in a spectrophotometer (Shimadzu UV 2700) with Milli-Q water as the reference (Ma et al., 2006).

2.2.2. SeaBASS dataset

To expand the dataset for the development of the Chl*a* algorithm, 97 samples (Dataset 3, DS3, in Fig. A1), including Chl*a* concentrations in Lake Erie, San Francisco Bay, and Chesapeake Bay corresponding to the OLI overpass were extracted from the SeaWiFS Bio-optical Archive and Storage System (SeaBASS) (https://seabass.gsfc.nasa.gov/search). Phytoplankton species vary in different waters, yet the optical property of Chl*a*, a proxy for phytoplankton abundance, is rarely influenced, so that most of studies use Chl*a* to develop models to estimate Chl*a* (Liu et al., 2020; Neil et al., 2019; Pahlevan et al., 2020). Some of the data in SeaBASS are based on high-performance liquid chromatography (HPLC) (Wright et al., 1991). Although Chl*a* determined spectrophotometrically may be overestimated when compared to the Chl*a* from the HPLC method (*e.g.*, Pinckney et al. (1994), the effect is assumed to be minor in the eutrophic waters examined as the difference between the measured approaches is relatively small with high Chl*a* concentrations (Liu et al., 2019).

2.2.3. Independent dataset in Lake Taihu

We also used *in situ* Chl*a* from Lake Taihu for the 2013–2015 period to validate the OLI derived Chl*a* time series in Lake Taihu. Samples were collected monthly at 14 pre-defined stations in the northern basin of Lake Taihu, and Chl*a* was determined using the method described in section 2.2.1. Overall, we selected 180 samples at five stations distributed across Lake Taihu (see section 3.2.2).

2.3. Satellite data acquisition and processing

1262 OLI scenes (radiance data) in MLRY and RHR were obtained from the United States Geological Survey (USGS) portal (https://earthexplorer.usgs.gov/) (Fig. 2). Radiance data were used instead of reflectance because reflectance was processed using a model developed for land applications, and radiance is not prone to error over water (Ilori et al., 2019). Only images with < 70% cloud cover, based on visual examination, were downloaded. A full atmospheric correction test (NIR-SWIR and MUMM) through the SeaWiFS Data Analysis System (SeaDAS, version 7.5) often failed in most of pixels in the lakes because the turbid waters caused the assumptions to fail. The dark-spectrum-function method in ACOLITE (Vanhellemont and Ruddick, 2018) may result in large uncertainties in MLRY lakes (*e.g.*, Lake Taihu) due to excessive algae particles and strongly absorptive aerosols (Wang et al., 2019). Therefore, the following procedure was used to create pseudo-reflectance products (Cao et al., 2017; Feng et al., 2012). First, Rayleigh-corrected reflectance (R_{rc}) was derived after correction for Rayleigh scattering and gaseous absorption effects using SeaDAS 7.5 with ancillary data (such as meteorological data) (Franz et al., 2015). To remove, at least partially, the aerosol signal, the following algorithm was used: $R'_{rc}(\lambda) = R_{rc}(\lambda) - R_{rc}(2201)$. This method may retain residual aerosol signals in other bands; however it partially removes the bulk aerosol, haze, or glint signal (Cao et al., 2017; Feng et al., 2012).

Cloud-contaminated pixels were removed *via* a threshold set on the SWIR reflectance ($R_{rc}(2201) > 0.018$) (Aurin et al., 2013). Waterbody boundaries were extracted using a scheme of normalized difference water index (NDWI) threshold segmentation (Li and Sheng, 2012), which is an automated mapping algorithm based on hierarchical image segmentation and delineates each waterbody using a local segmentation threshold. Subsequently, the segmentation-based water boundary of OLI was screened using the Chinese lake boundaries (Ma et al., 2011). Reservoirs, rivers, and ponds were excluded, and only lakes larger than 1 km^2 were considered. OLI may not observe narrow sections in some thin lakes (< 30 m) and regions covered by macrophytes. Because algal blooms common to our study lakes are usually caused by cyanobacteria, surface scums can be present; hence, a threshold −0.004 on the floating-algae-index (FAI) was used to exclude pixels with algal blooms (Hu et al., 2010).

2.4. Machine learning approach

2.4.1. Calibration dataset

To ensure the high quality of *in situ* and concurrent OLI data used to develop the algorithm, the field data were selected according to several criteria. First, the satellite RGB images and survey records were visually examined to exclude stations covered by cloud, heavy haze, fog, and algae scums. Because low Chl*a* concentrations were rare and reduced the accuracy of BST model in a preliminary examination, samples with Chl*a* concentration less than 1 μg L^{-1} were excluded. Second, the time lag between field data collection and OLI data acquisition was constrained to < 6 h. Third, a coefficient of variation (CV) test for each band (R_{rc} in a 3 × 3 window centered at the sampling station with CV < 10%) was employed to assure the uniformity of the water surrounding the sampled location (Qi et al., 2014). After applying these criteria, 322 matches between OLI data and field- measured data were identified (detailed information for these samples shown in Table 2). From this dataset, 225 samples were obtained in our study area (eastern China) and 97 were extracted from SeaBASS. These data were randomly divided into a training dataset (N = 220) and a validation dataset (N = 102).

2.4.2. BST model

One of the most widely used and efficient machine learning models, the extreme gradient boosting method, *XGBoost* (BST), is used to train the Chl*a* model (Chen and Guestrin, 2016; Ghatkar et al., 2019). BST is one of boosting-tree models integrating many regression trees (*i.e.*, decision trees where the values of target variable are continuous) to construct a strong classifier (Chen and Guestrin, 2016), and has good prediction accuracy and is widely used in machine learning. The algorithm continuously adds trees to achieve feature splitting, each tree has several leaf nodes, each leaf node corresponds to a score, and finally the sum of the corresponding score in each tree is the prediction (Fig. 3). From a mathematical perspective, the addition of a tree generates a novel function to fit the residual error of the last tree. The object function of BST is expressed as follows (Eq. (2)):

$$Obj = \sum_{i=1}^{N} L(M_i, E_i) + \sum_{k=1}^{K} \Omega(f_k) \quad (2)$$

$$where\ \Omega(f_k) = \gamma T + \frac{1}{2}\lambda \|\omega\|^2$$

where $L(M_i, E_i)$ is the difference between the estimated (E_i) and measured value (M_i), and N is the number of data points in the training

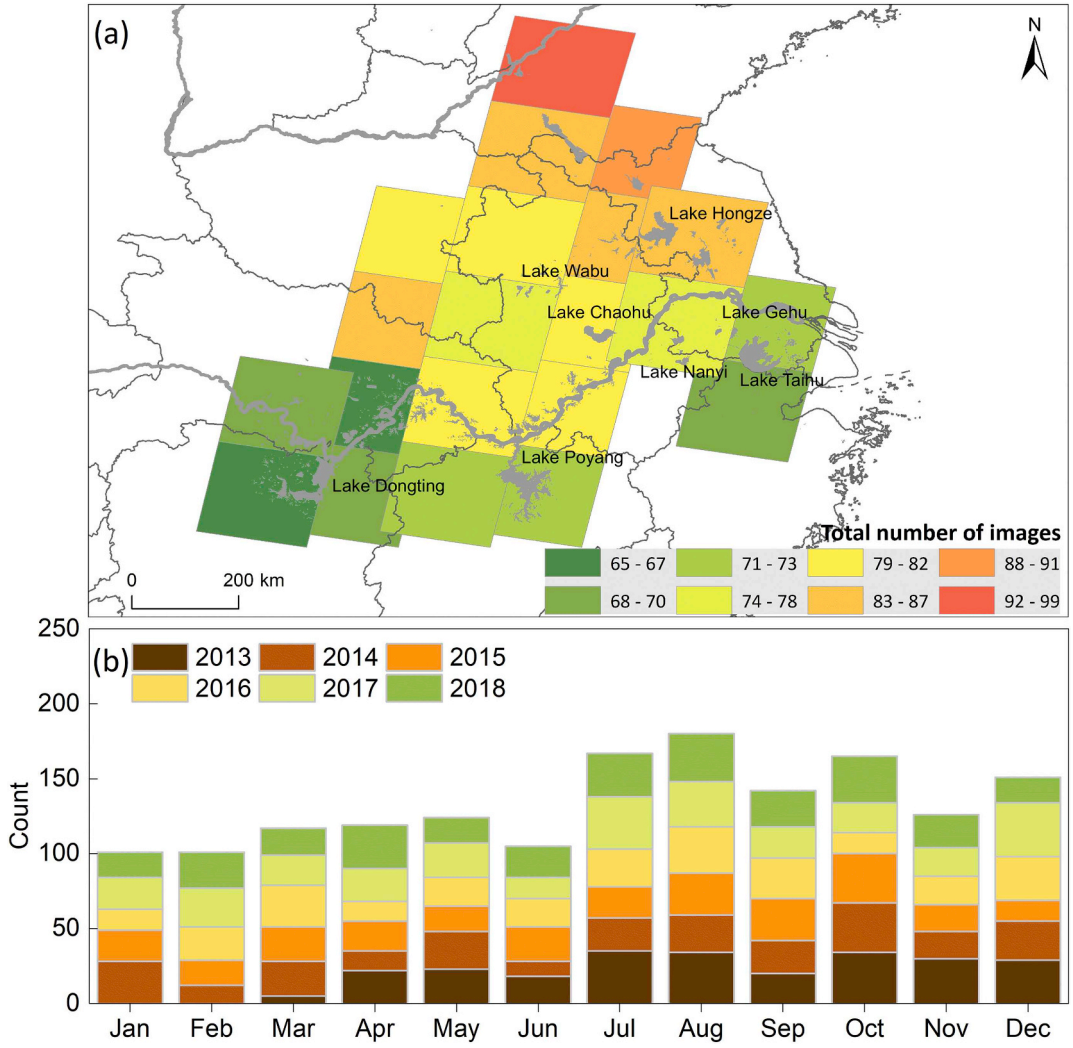

Fig. 2. Geographic and temporal coverage of Landsat-8 archive from March 2013 to December 2018 used in this study.

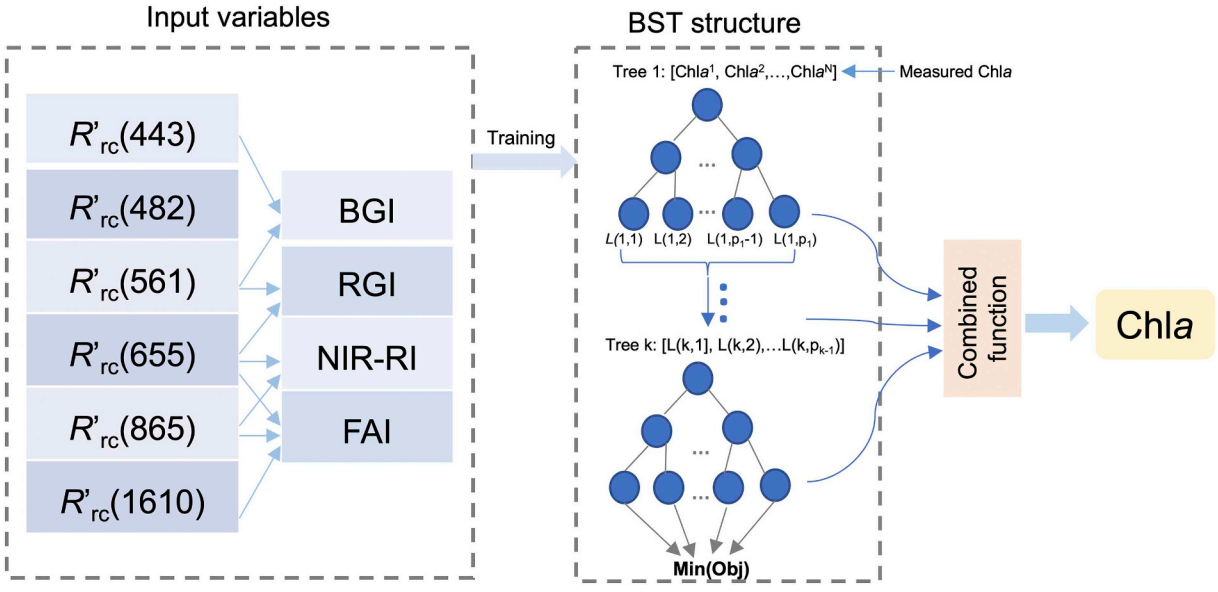

Fig. 3. Schematic block diagram illustrating the main process of a BST algorithm, an optimized distributed gradient boosting method that solve the object function (Obj, Eq. (2)) with decision tree ensembles and additive training to obtain an optimal Chl*a* retrieval from OLI R_{rc}. Initial input for the training is field Chl*a* and subsequent inputs are the residual errors fitted by the previous tree [*i.e.*, L(k, p_k) at the *p*-th leaf node in the *k*-th tree], and an optimal model structure with the lowest Obj value is used to estimate Chl*a* by summing up the score in the corresponding leaves.

Table 3
Bio-optical and water quality properties measured in the lakes from two surveys in August 2012 and October 2017. S.D. denotes the standard deviation.

	August 2012				October 2017			
	N	Range	Mean ± S.D.	Median	N	Range	Mean ± S.D.	Median
Chl*a* (μg L^{-1})	187	7.1–258.7	77.4 ± 54.9	63.8	110	0.4–149.7	24.3 ± 31.3	12.78
SPM (mg L^{-1})	169	3.0–218.9	22.7 ± 21.8	19.4	109	10.4–245.0	46.6 ± 40.7	30.7
SPIM (mg L^{-1})	169	0.5–203.0	14.9 ± 20.6	9.3	109	1.3–232.0	38.1 ± 38.8	22.0
$a_{ph}(443)$ (m^{-1})	169	0.04–8.7	2.6 ± 1.8	2.2	109	0.2–5.9	1.1 ± 1.0	0.6
$a_d(443)$ (m^{-1})	169	0.2–15.1	2.1 ± 1.6	1.8	109	0.4–10.3	2.5 ± 1.9	1.7
$a_g(443)$ (m^{-1})	163	0.3–4.9	0.9 ± 0.6	0.8	109	0.3–2.5	1.1 ± 0.4	1.0

dataset. $\Omega(f_k)$ is complexity of the *k*-th tree, where K is the total number of trees in this model. T is the number of leaves in the tree, and ω is the leaf weights, and λ and γ are the regularization coefficients (given as constants). *Obj* can be divided into two terms: the first is the training loss, which measures how well the model fits to training data, and the second one is the regularization, which controls the complexity of model and alleviates overfitting.

2.4.3. Model structure

The input of our model has 10 spectral variables (Fig. 3): R'_{rc} for the first six bands of OLI (443, 482, 561, 655, 865, 1610 nm), the blue-green ratio index (BGI) ($R'_{rc}(443)/R'_{rc}(561)$) (Ha et al., 2017), red-green ratio index (RGI) ($R'_{rc}(655)/R'_{rc}(561)$) (Watanabe et al., 2018), near-infrared-red ratio index (NIRRI) ($R'_{rc}(865)/ R'_{rc}(655)$) (Duan et al., 2007), and FAI (Page et al., 2018). Band combinations of NIR, red, and green on Landsat sensors have been demonstrated to retrieve Chl*a* in some coastal and inland waters (Dekker and Peters, 1993; Duan et al., 2007; Giardino et al., 2001; Ha et al., 2017). The output is Chl*a*. We examined different input parameters using different combinations of bands and band ratios, and the input variables with these 10 variables produced the best performance. To determine the structure of the BST model, several hyperparameters, including the learning rate, maximum tree depth, subsample rate and regularization, need to be tuned. A step-by-step tuning with grid search strategy was used to determine the hyperparameters of the BST model. A 5-fold cross validation was employed to ensure that the training dataset was randomly distributed in different segments, so that the model performance was not significantly influenced by the training dataset (Fan et al., 2017).

In addition, we developed another algorithm for Chl*a* estimation from OLI measurement with the random forest (RF) model to compare to the performance of BST model. RF models are a class of ensemble learning techniques for regression *via* establishing multiple decision trees during training, and outputs the average predicted value for each tree. RF models have been widely used to obtain the water quality parameters from satellite measurements (Chen et al., 2019). For the development of the RF model, we used the same input variables and training dataset, and the hyperparameters were determined by the strategy of grid search.

2.5. Accuracy assessment

The coefficient of determination (R^2), slope (in linear regression), root mean square error (RMSE), mean absolute percentage error (MAPE), bias (system error) and mean absolute error (MAE) were used to perform statistical analyses. R^2, RMSE, and MAPE are common metrics to assess the performance of models based on the original data distribution, MAE is the mean absolute error computed in log-space, and the bias represents log-transformed residuals. Bias and MAE computed in log-transformed space are believed to provide a good assessment of the algorithms for the log-normal distributions of Chl*a* (Pahlevan et al., 2020; Seegers et al., 2018). For example, a MAE of 1.5 indicates relative measurement error of 50%, while a bias of 1.5 indicates that the model is 1.5× (50%) greater on average than the estimated variable. These metrics are more robust and straightforward quantities for evaluating remote sensing algorithms with the log-distributions (Seegers et al., 2018); they are defined as follows:

$$RMSE = \frac{1}{N}\sqrt{\sum_{i=1}^{N}(M_i - E_i)^2} \quad (3)$$

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\frac{|M_i - E_i|}{M_i} \quad (4)$$

$$bias = 10\text{^}\left(\frac{\sum_{i=1}^{N} log_{10}(M_i) - log\,10(E_i)}{N}\right) \quad (5)$$

$$MAE = 10\text{^}\left(\frac{\sum_{i=1}^{N} |log_{10}(M_i) - log\,10(E_i)|}{N}\right) \quad (6)$$

where N is the number of data pairs, the subscript *i* denotes individual data points, and M and E represent the measured values and estimated values, respectively.

3. Results

3.1. Optical properties of the field dataset

Two comprehensive surveys including 67 lakes in the MLRY and RHR had average Chl*a* in summer (mean ± standard deviation: 77.4 ± 54.9 μg L^{-1}) higher than in autumn (24.3 ± 31.3 μg L^{-1}), whereas SPM in autumn (46.6 ± 40.7 mg L^{-1}) was higher than that in summer (22.7 ± 21.8 mg L^{-1}) (Table 3). The proportion of SPIM in SPM was more than that in SPOM in both seasons, suggesting that the suspended particulates were mainly SPIM. Absorption coefficients, a_{ph} (443) (mean: 2.6 m^{-1}) in summer were higher than $a_d(443)$ (2.1 m^{-1}), whereas $a_d(443)$ in autumn (2.5 m^{-1}) was higher than $a_{ph}(443)$ (1.1 m^{-1}). The lakes generally had high $a_d(443)$ associated with high SPIM, and high $a_{ph}(443)$ in summer could be related to the growth of phytoplankton. Overall, $a_g(443)$ (mean: 0.9 m^{-1}), was always below 5 m^{-1}, indicating that these lakes had low organic matter. A tenary plot of $a_{ph}(443)$, $a_d(443)$, and $a_g(443)$ shows a the relative proportions of different materials (Fig. A2). On average, $a_d(443)$ and $a_{ph}(443)$ contributed about 55% and 30% of absorption coefficients, which were higher than that for $a_g(443)$, suggesting that the optical properties in the MLRY and RHR lakes were dominated by inorganic matter and phytoplankton.

For studied lakes, $R_{rs}(550)$ ranged from 0.0150 to 0.065 sr^{-1} with a mean of 0.0383 ± 0.0084 sr^{-1}, which is a relatively broad spectral range. R_{rs} troughs occurred in the blue and red bands and peaks were located in the green band and ~ 700 nm (Fig. 4a), which are typical spectral features in turbid waters. The dataset used in the development of the BST model had average Chl*a*, SPM and $a_g(443)$ of 29.2 ± 66.2 μg L^{-1} (range: 1.0–99.0 μg L^{-1}), 55.1 ± 31.7 mg L^{-1} (5.0–239.0 mg L^{-1}) and 0.9 ± 0.4 m^{-1} (0.2–2.4 m^{-1}), respectively (Figs. 4b-d). SPM and $a_g(443)$ in DS1 were close to that of all *in situ* data (Table 3), but Chl*a* in DS1 was lower slightly than that of all data,

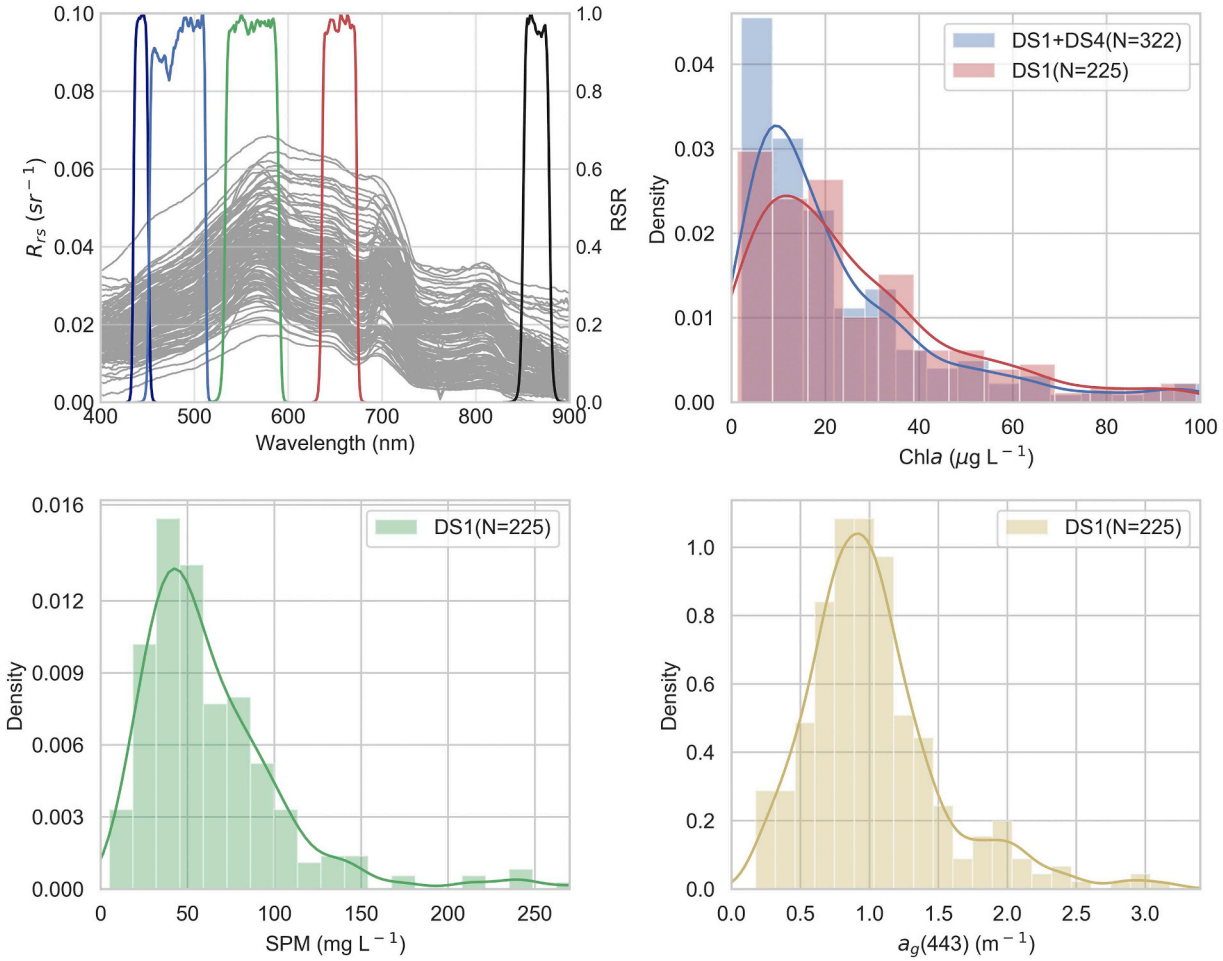

Fig. 4. (a) *In situ* R_{rs} (gray lines) with the relative spectral responses (RSR, color lines, (Barsi et al., 2014)) of Landsat-8 OLI. Panels (b)-(d) show the distributions of Chl*a*, SPM and $a_g(443)$ used in the development of algorithm, respectively. Chl*a* statistics include the dataset in SeaBASS and our data, whereas SPM and $a_g(443)$ were only from the dataset in study area. The curved lines are kernel density corresponding to the histograms.

because adding the Chl*a* in DS4 (*i.e.*, SeaBASS) shifted the distribution to lower Chl*a* (Fig. 4b). Overall, data distributions reflect principal optical properties of the study lakes in MLRY and RHR, suggesting that model developed by DS1 would be applicable to the Chl*a* retrievals of lakes in study area.

3.2. Performance of the algorithm

3.2.1. Development and validation of BST model

BST model performed well as indicated by the statistical metrics computed with the validation dataset (Fig. 5a; Table 4) (R^2 = 0.79, slope = 0.75, RMSE = 7.1 μg L^{-1}, MAPE = 24.2%, MAE = 1.4, bias = 0.9). The model did slightly underestimate especially high Chl*a* values (> 60 μg L^{-1}) which could be related to paucity of high Chl*a* concentration in training dataset (section 2.3). The BST model performed better than the RF model (Fig. 5b) and other algorithms tested (Fig. 5c). While the RF model did reasonably well estimating Chl*a* (R^2 = 0.73, RMSE = 7.8 μg L^{-1}, MAPE = 29.9%, MAE = 1.5, bias = 0.8), it had significant deviations at high and low Chl*a*. Likewise, an algorithm developed by Shi et al. (2015) had a large deviation from *in situ* Chl*a* (R^2 = 0.21, RMSE = 11.3 μg L^{-1}, MAPE = 64.1%, MAE = 1.8, bias = 0.9) (Table 1; Fig. 5c). The data in SeaBASS were not used to evaluate the performance of the Shi et al. (2015b) model. Three algorithms for Chl*a* estimation in other regions (Table 1) were implemented and tested with our *in situ* dataset: $R_{rs}(865)/R_{rs}(655)$ (Duan et al., 2007), $R_{rs}(655)/R_{rs}(561)$ (Watanabe et al., 2018), and $R_{rs}(561)/R_{rs}(482)$ (Ha et al., 2017). The *in situ* R_{rs} were resampled using the relative spectral responses (RSR) of OLI prior to assessment of algorithms. These indices had poor correlations with *in situ* Chl*a* in DS1 (R^2 < 0.01, RMSE > 20 μg L^{-1}).

The performance evaluations of the BST model trained with different input variables indicated the BST model with ten variables was the best model (Table 4). The BST model only with four spectral indices (SI) (MAPE = 39.1%, MAE = 1.7, Bias = 0.8) had better performance than that only with six R'_{rc} (MAPE = 42.9%, MAE = 1.9, Bias = 0.7), suggesting that additional SI can significantly improve the performance of Chl*a* retrievals *via* the BST model. The BST model trained with R'_{rc} estimated Chl*a* better than that with the uncorrected R_{rc}, indicating that removing the offset signal in the 2201-nm channel from R_{rc} (Section 2.3) could improve the model accuracy. The addition of SWIR bands could facilitate the ability of BST model to retrieve the Chl*a* in turbid lakes (*e.g.*, the performance of BST-7R_{rc} was better than that of the BST-5R_{rc}).

3.2.2. Further validation on satellite images

Data from two turbid and eutrophic shallow lakes, Chaohu and Taihu, were used to estimate Chl*a* and to determine the viability of the BST algorithm for producing spatial and temporal products. Fig. 6 illustrates the OLI-derived Chl*a* products for Lake Chaohu on October 11, 2015, as well as the synchronous *in situ* Chl*a* (Fig. 6b). We infer that the OLI is able to obtain a realistic Chl*a* map, comparable to the true-color composite image (Fig. 6a) and *in situ* Chl*a* data (Figs. 6b-c). However, the OLI-derived values were slightly less than the *in situ* data for those pixels contaminated by algae scums.

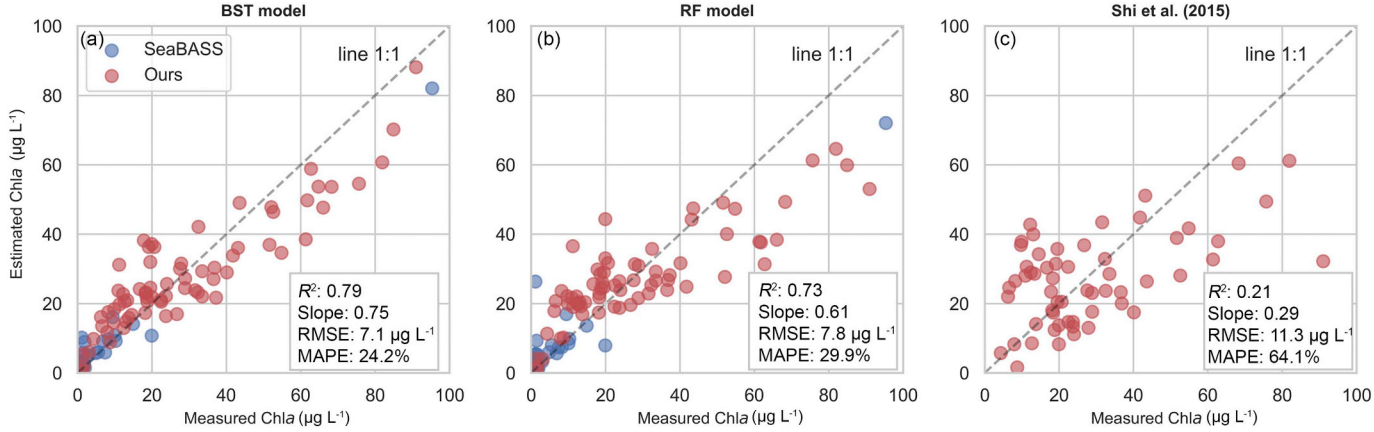

Fig. 5. Performance evaluation of Chl*a* retrievals using the BST (a), RF (b) and Shi et al. (2015) (c) implemented for OLI derived R_{rc}. BST and RF algorithms (N = 102) show the results of validated dataset, and Shi et al. (2015b) was evaluated using only the data in eastern China lakes (N = 71). Detailed performance metrics and broader evaluation for other models are included in Table 4.

Table 4
Performance metrics associated with Chl*a* retrievals using the BST, RF, Shi et al. (2015). BST models with different input variables. BST and RF algorithms were evaluated with validation dataset, and Shi et al. (2015) was evaluated using all dataset in our study lakes. BST-6$\dot{R}_{rc}$, BST-4SI, BST-7R_{rc}, and BST-5R_{rc} represent the BST models trained by the $\dot{R}_{rc}$ (*i.e.*, $R_{rc}(\lambda)$-$R_{rc}(2201)$) with six single bands, four spectral index, R_{rc} at seven single bands, and the R_{rc} at five visible bands, respectively.

Model	N	R^2	Slope	RMSE (μg L^{-1})	MAPE (%)	MAE	Bias
BST	102	0.79	0.75	7.1	24.2	1.4	0.9
RF	102	0.73	0.61	7.8	29.9	1.5	0.8
Shi et al. (2015)	71	0.21	0.29	11.3	64.1	1.8	0.9
BST-6$\dot{R}_{rc}$	102	0.53	0.40	8.5	42.9	1.9	0.7
BST-4SI	102	0.59	0.42	7.9	39.1	1.7	0.8
BST-7R_{rc}	102	0.49	0.39	9.1	43.2	1.8	0.8
BST-5R_{rc}	102	0.36	0.33	9.9	45.8	1.89	0.8

Monthly *in situ* (see section 2.2.3) and OLI-derived Chl*a* time series in Lake Taihu from 2013 to 2015 are illustrated in Fig. 7. Five stations situated in different sections of the lake span a wide range of optical properties. Chl*a* in three bays (Figs. 7 a-c) and western regions (Fig. 7d) were higher than that of the central lake (Fig. 7e) indicated by OLI-derived and *in situ* data. Chl*a* in summer and autumn were higher than other seasons, as observed by Shi et al. (2015). Furthermore, OLI-derived Chl*a* had a temporal trend consistent with the *in situ* Chl*a* at five stations. The *in situ* time series of Chl*a* had a larger standard deviation than that of OLI-derived Chl*a*, which may be related to more *in situ* data. OLI-derived Chl*a* did underestimate compared to *in situ* data where Chl*a* was more than 60 μg L^{-1} (*e.g.*, Fig. 7a).

3.3. Spatial and temporal variations of Chla from 2013 to 2018

3.3.1. Spatial differences of Chla among the lakes

BST algorithm was used to calculate Chl*a* in the lakes larger than 1 km^2 in the MLRY and RHR from 2013 to 2018. We calculated the time averaged Chl*a* for each lake, and then merged the annual and seasonal mean Chl*a* of each lake (Fig. 8). Pixels covered by algal scums (*i.e.*, FAI > -0.004) were excluded. Chl*a* during 2013–2018 ranged between 2.2 and 75.1 μg L^{-1}, the mean ± S.D. was 31.1 ± 7.9 μg L^{-1} (median:30.3 μg L^{-1}). Average Chl*a* had inter-lake differences, and each lake had spatial variability (Fig. 9a). For example, Chl*a* was low in Lake Poyang, Dongting and Hongze and high in Lake Taihu, Chaohu and small lakes along the middle reach of Yangtze River (Fig. 8). To further examine Chl*a* changes among various lake sizes, a stratified statistical analysis was done. Mean Chl*a* in the lakes with the areas < 10 km^2, 10–50 km^2, 50–100 km^2, 100–500 km^2 and > 500 km^2 were 31.9 ± 6.6 μg L^{-1}(N = 473), 30.1 ± 6.2 μg L^{-1} (N = 92), 29.2 ± 4.73 μg L^{-1}(N = 13), 28.2 ± 3.54 μg L^{-1}(N = 20) and 22.6 ± 4.8 μg L^{-1} (N = 7), respectively (Fig. 9b). The small lakes usually had higher Chl*a* than large lakes in the MLRY and RHR.

3.3.2. Seasonal variations in Chla

OLI-derived Chl*a* had distinct seasonal patterns across all of the hundreds of lakes in the MLRY and RHR. Chl*a* was highest in summer

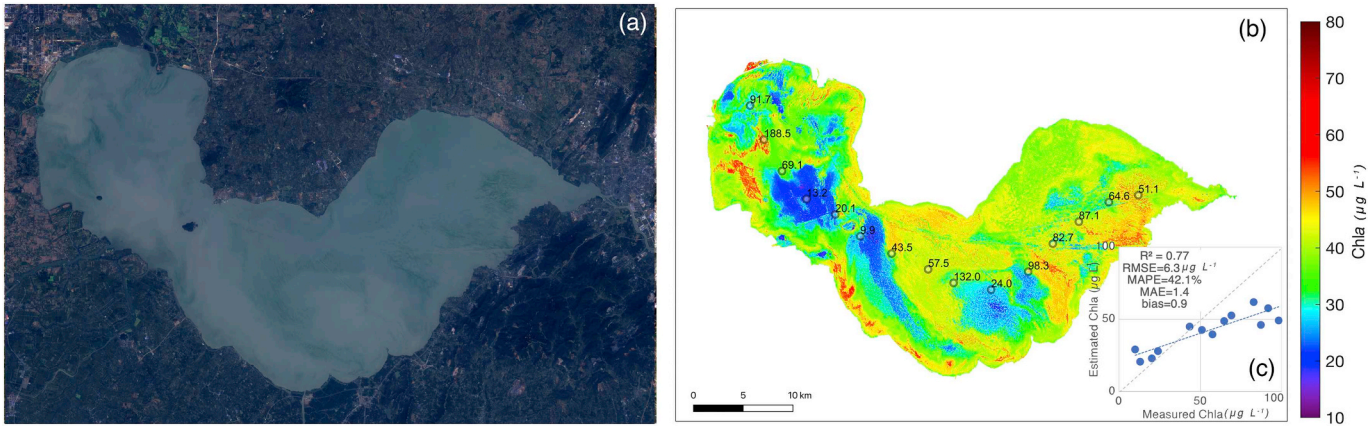

Fig. 6. (a) True-color composite image in Lake Chaohu (2015/10/11). (b) Chl*a* produced by the OLI using the BST model and the *in situ* Chl*a* on the data overlaid. (c) Comparison between the *in situ* and OLI-derived Chl*a*. The point pairs with measured Chl*a* > 100 μg L^{-1} were not included in the statistics.

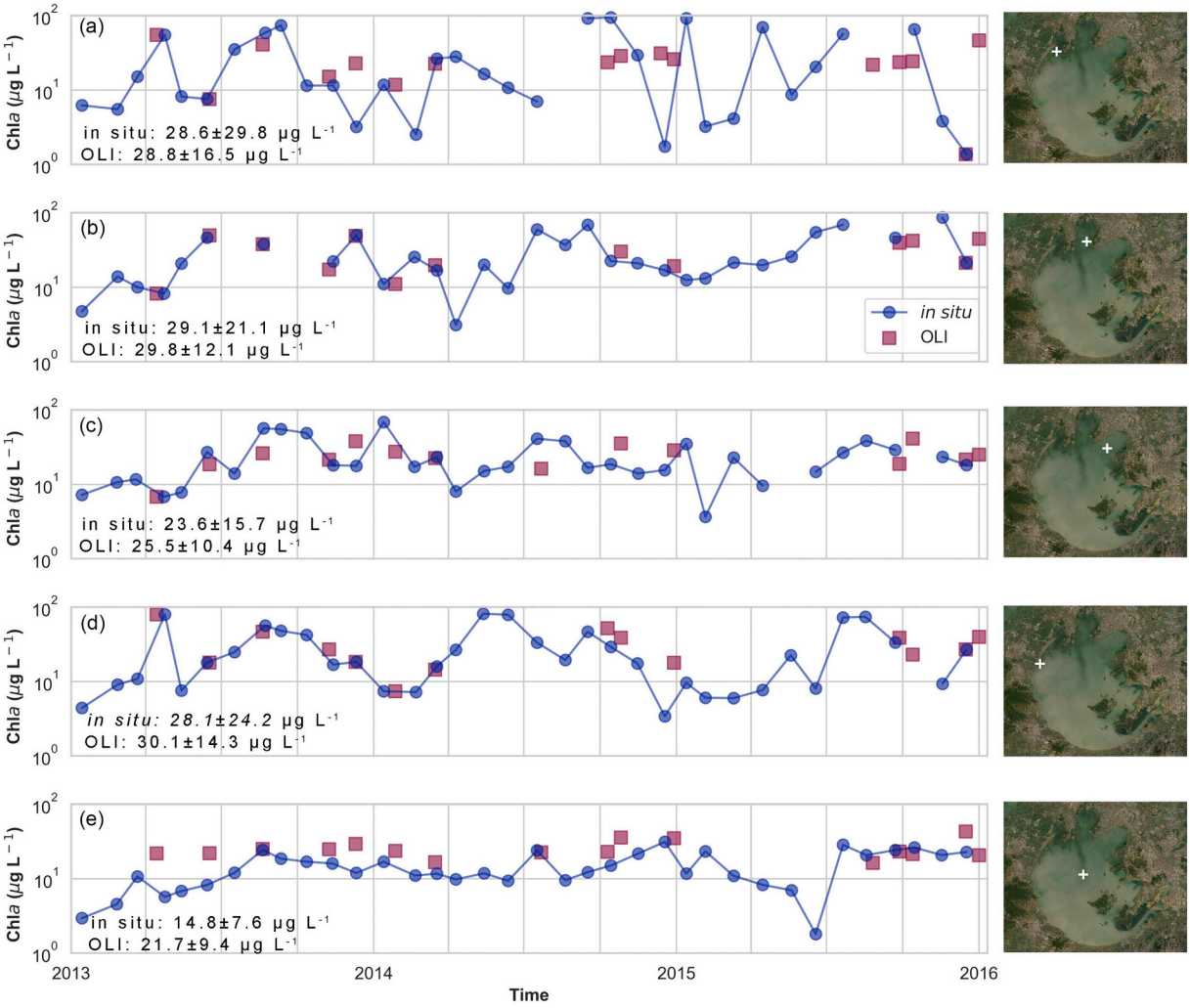

Fig. 7. Time series of *in situ* and OLI-derived Chl*a* for five stations from 2013 to 2015 in Lake Taihu. X-axis is the specific time of field measurement and satellite overpass for each point in Lake Taihu.

(33.4 ± 10.3 μg L^{-1}) and lowest in winter (26.9 ± 6.5 μg L^{-1}) (Fig. A3; Fig. 10a). From 2013 to 2018, annual seasonal mean Chl*a* did not have significant temporal variability (Fig. 10a). The fluctuation of Chl*a* among summers was large ranging between 32.2 and 44.0 μg L^{-1} and was less in winter (26.7–28.2 μg L^{-1}). Observations used in statistics for seasonal averages of Chl*a* every year included only 2–4 images due to cloud cover and long revisit time of Landat-8 (16 d) (Fig. 10a), indicating that annual seasonal average Chl*a* may not reflect the actual variations of Chl*a*. Mean Chl*a* concentration considered by season among different sizes of lakes had an inverse relation to the lake area: the small lakes had a higher Chl*a* than that in large lakes in each season from 2013 to 2018 (Fig. 10 a and b). Average Chl*a* of lakes less than 10 km^2 in summer reached as high as 49.7 μg L^{-1} (Fig. 10b).

4. Discussion

4.1. How does the BST model work?

The BST model had good performance for Chl*a* retrievals from OLI in turbid lakes, even though OLI is not equipped with a red-edge band and a robust atmospheric correction remains a challenge in such environments (Kravitz et al., 2020; Wang et al., 2019). Though BST model is of "black box" nature, it is based on optical properties of Chl*a* in turbid lakes. We carried out a feature-ranking assessment on the ten input features (six spectral bands and four indices; section 2.4.3) to assist in the interpretation of the model outcomes and how each feature contributes to Chl*a* predictions. The goal is to understand how the model captures non-linearities between these features and Chl*a*, in presence of fairly large NAP and CDOM (Table 3). In doing so, the ranking score for each attribute was calculated for a single decision tree by the amount that splitting nodes for each attribute improves the performance measure, weighed by the number of observations at the node. Then, the relative importance was computed as the ratios between the importance score of every feature and the sum of all of importance scores. Importance scores were determined by 225 samples in the process of training BST model. The analysis showed that the NIR band contributed about 50% to the BST model for Chl*a* retrievals (Fig. 11a), and the FAI, $R'_{rc}(865)/R'_{rc}(561)$ contributed ~10%. The enhanced signal in the NIR band or that manifested in FAI correlated well with high biomass, leading to a reasonable model performance in our study sites. While other input features had low relative importance scores, they may be helpful by reducing impacts of NAP and CDOM on the Chl*a* retrievals.

4.2. Strengths of approach

The BST approach performs well for a wide range of optical properties in turbid and eutrophic waters. First, the performance of BST model was not largely impacted by perturbations caused by high concentrations of suspended particulates or aerosol conditions. For

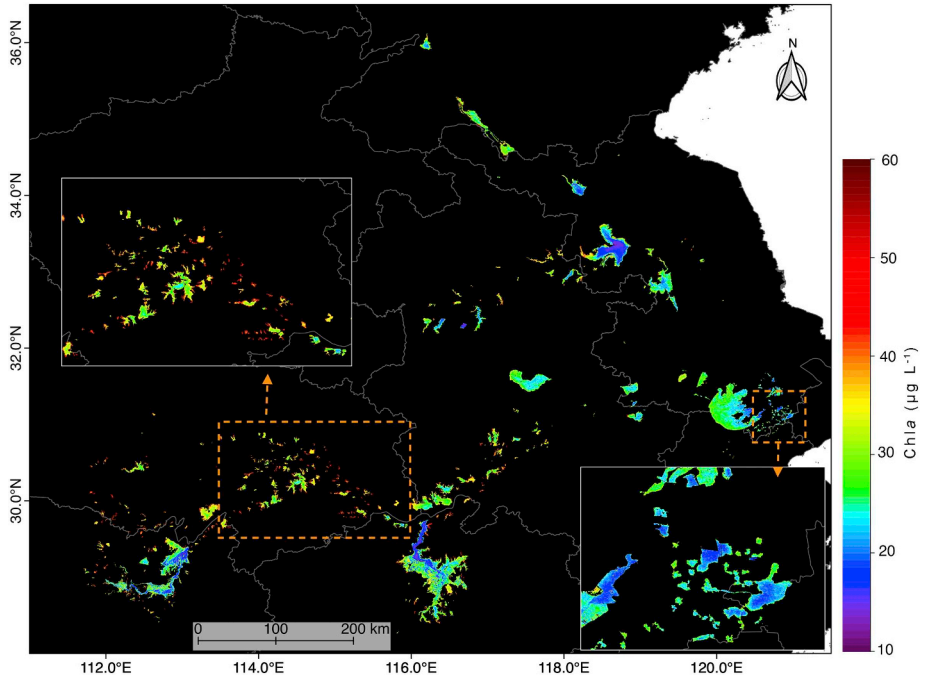

Fig. 8. Mean Chl*a* distributions derived from the BST algorithm with OLI images in the lakes in eastern China for the 2013–2018. Zoom a and b shows two regions covered by many small lakes.

example, some lakes in eastern China, such as Poyang, Dongting, and Hongze, have Chl*a* $< 20\ \mu g\ L^{-1}$ and SPM concentrations $> 50\ mg\ L^{-1}$ (Cao et al., 2017; Feng et al., 2014; Wu et al., 2013). While it has proven difficult to empirically retrieve Chl*a* in turbid waters (Feng et al., 2014), the BST model appears to offer reasonable estimates of Chl*a* with a decent accuracy ($N = 24$, $R^2 = 0.55$, RMSE = 3.2 μg L^{-1}, MAPE = 51%, MAE = 1.3, bias = 0.8) (Fig. 8). Generally, relatively low SNRs in the SWIR bands over turbid and/or algal-dominated waters along with strongly absorbing aerosols could lead to significant uncertainties in R_{rs} (Pahlevan et al., 2017a; Wang and Jiang, 2018). Therefore, many pixels may not be processed through the standard atmospheric correction processors (Wang et al., 2019). Although the simple correction by subtracting the $R_{rc}(2201)$ from R_{rc} in six bands was only a partial correction, it improved the performance of the BST model (Table 4).

The RF model is another efficient method to train models (Chen et al., 2019). However, our application of the RF approach did not produce the expected performance (Fig. 5b; Table 4). We hypothesize that the performance of the RF approach was reduced because it did not incorporate covariances of input features and did not allow for regularization parameters to control overfitting of the model. BST learns the covariances among the input features and improves the ability of generalization through the regularization design (Chen and Guestrin, 2016). Also, the BST model implements a gradient descent on the

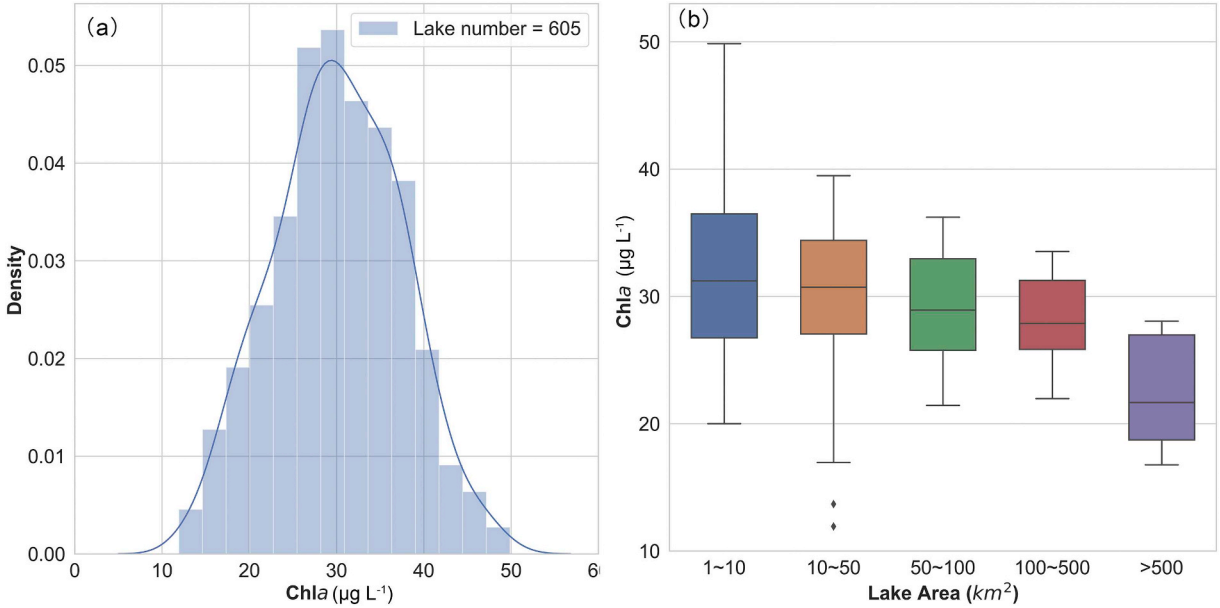

Fig. 9. (a) The distributions of mean Chl*a* for the lakes in eastern China during 2013–2018. The 1st quantile was 25.2 μg L^{-1}, 3rd quantile was 35.4 μg L^{-1}, and the median is 30.13 μg L^{-1}. (b) The variations of mean Chl*a* in different size classes.

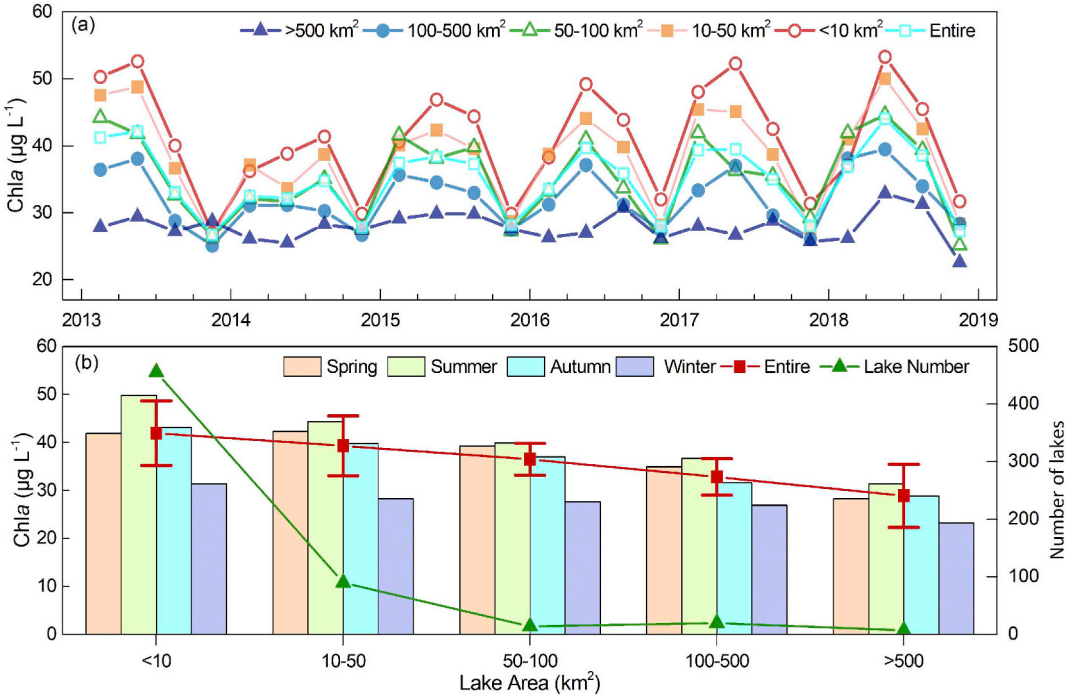

Fig. 10. (a) Annual seasonal variations of Chl*a* for different sizes of lakes: < 10 km^2, 10–50 km^2, 50–100 km^2, 100–500 km^2, > 500 km^2 and all lakes. About three OLI images were used for the Chl*a* statistics per season. (b) The seasonally variations of Chl*a* in different sizes of lakes. The green line is the number of lakes for each lake size class. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

residual errors, so that the output result is continuous and can obtain a fast and accurate convergence.

4.3. Limitations on applying to other areas

One of limitations of machine learning-based model is that its performance is limited by the representativeness of the training dataset. Comparing to the global dataset collected in approximately 250 inland waters (N = 4035 samples) encompassing 13 different optical water types (Spyrakos et al., 2017), our dataset did not include Chl*a* in clear or hypereutrophic waters, and did not cover waters with high CDOM (> 1 m^{-1}). Thus, the sensitivity of BST model to the training dataset should be determined for examining its universality. Here, we trained another BST model using 70% of the DS1 dataset, and the performance of this alternative model was evaluated with the remaining 30% of the DS1 and SeaBASS dataset, respectively (Fig. 11 b and c). This retrained model had a slightly lower performance on the validated dataset than that of BST model described in section 2.4, possibly due to a number of samples with low Chl*a* included in the model improving the accuracy for the estimation of low Chl*a* in the water. The retrained BST model had fair performance for eastern Chinese lakes (R^2 = 0.59, RMSE = 9.3 μg L^{-1}, MAPE = 41.7%) (Fig. 11b), indicating that the BST model was effective for the whole study area, including the small lakes. However, this retrained BST model did not produce satisfactory results when applied only to the SeaBASS dataset (R^2 = 0.25, RMSE = 2.6 μg L^{-1}, MAPE = 67.6%). This result suggests that the model developed by a local dataset may not have sufficient accuracy when applied to other areas. Nevertheless, the validation on separate datasets does provide another method to evaluate the performance and applicability of an algorithm. To achieve more reliable Chl*a* in our study region and to extend its potential application to other lakes, we developed the BST model using the DS1 and SeaBASS datasets (see section 2.4).

Another issue is that the BST model had larger uncertainties in the

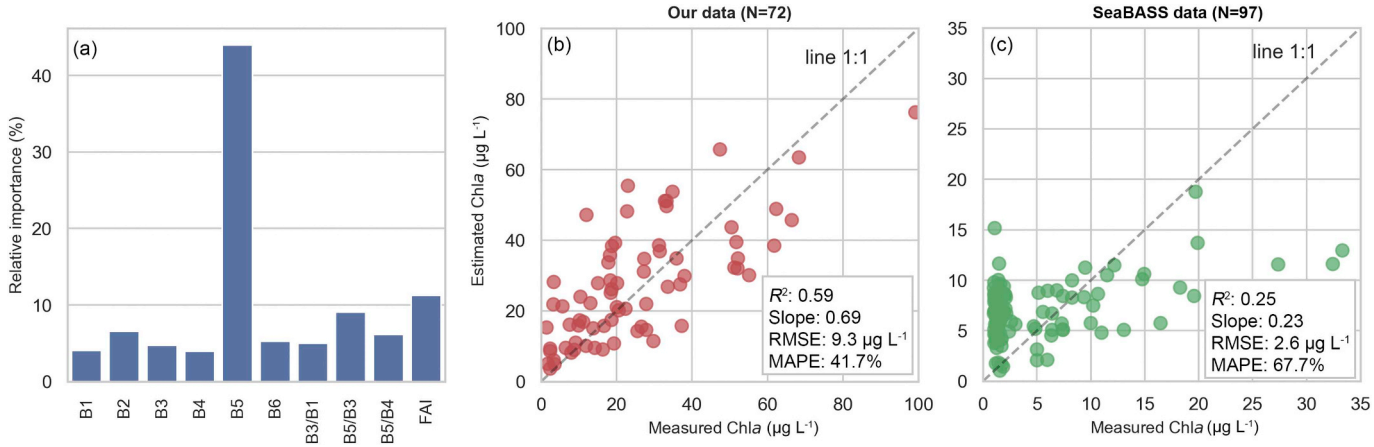

Fig. 11. (a) Relative importance of each input variable to the BST model. (b) and (c) show the performance of the model only trained by the 70% dataset (N = 153) in our study lakes on remaining 30% dataset (N = 72) in our lakes and the SeaBASS dataset (N = 97), respectively.

retrievals of high or low Chl*a* (< 5 μg L^{-1}or > 60 μg L^{-1}) (Fig. 5) potentially due to the paucity of high Chl*a* concentrations in training dataset. Also, BST model was not trained to estimate Chl*a* in bloom conditions, which constrained the application of the model. Likewise, the training dataset used for the development of BST model was mainly from medium and large lakes (Table 2). That small lakes have higher Chl*a* than that in large lakes was consistent with *in situ* Chl*a* in this area (Liu et al., 2019), the actual retrievals of Chl*a* in small lakes may have potential uncertainties. Adding data from small lakes would likely improve the model.

The BST model does not explicitly understand the mechanism and error propagation from the input and output variables (Reichstein et al. 2019), which is the nature of machine-learning techniques. BST combines all input variables to obtain the output variables (*i.e.*, Chl*a*). As a result, it is difficult to explain how each process affects the final output. Nevertheless, the use of NIR band in the model does make sense since the NIR band is important for atmospheric correction in turbid waters (Ruddick et al., 2000; Siegel et al., 2000), and is the featured band for phytoplankton-dominated waters and has been used to identify algae blooms and retrieve Chl*a* (Hu et al., 2010; Ruddick et al., 2001).

4.4. Considerations for generating reliable remote sensing products of water quality

The revisit time of Landsat-8 is 16 days and, when combined with periods of cloud cover, common in eastern China, results in few scenes per year. Thus, annual and monthly mean water products may deviate from the true conditions (Cao et al., 2019b). Therefore, only average Chl*a* for all images and seasonal mean Chl*a* in the lakes in the MLRY and RHR were reported.

Although the observation frequency of Landsat-8 may be similar to field surveys (often monthly or less), Landsat-8 data reduces time, effort, and costs to obtain water quality information for an entire lake. In recent years, several freely-available, high-quality, moderate-resolution satellites have been launched, *e.g.*, Sentinel-2 MSI and Landsat-8 OLI, providing an opportunity for remote sensing of lakes. The Landsat-Sentinel-2 virtual constellation can achieve global coverage with ~3 day revisit time (Li and Roy, 2017). If the BST model is extended to MSI instruments using additional synchronous datasets, this may generate reliable Chl*a* products *via* MSI and OLI, to enable limnologists, aquatic ecologists and water resource managers to enhance their monitoring efforts, especially in small lakes (Pahlevan et al., 2019).

5. Conclusion

To estimate Chl*a* concentration with OLI measurements in lake waters of the eastern China region, a practical and effective machine learning approach was developed. The BST algorithm trained with 70% of the *in situ* Chl*a* data and near-synchronous OLI-derived R_{rc} intermediate products demonstrated satisfactory performance. This algorithm provides an ability to retrieve Chl*a* in turbid lakes and identifies spatial distributions of Chl*a* in lakes. BST-derived Chl*a* in Lake Taihu had temporal variations consistent with an *in situ* time series. The BST algorithm was employed to retrieve spatial and temporal distributions of Chl*a* in 605 lakes in the MLRY and RHR from 2013 to 2018. Chl*a* in summer was highest and lowest in winter. Chl*a* concentrations were inversely related to lake size. The BST model provides a practical method to estimate Chl*a* from OLI measurements in turbid waters when accurate atmospheric correction is challenging and may preclude reliable Chl*a* retrievals. However, the performance of BST models trained by different training datasets showed that accuracy for Chl*a* retrievals was limited by the size and range of the dataset, demonstrating that the BST model developed here requires more data prior to application in other regions.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors would like to thank Robert Frouin (University of California, San Diego) for providing valuable suggestions, and Zhaoshi Wu (Nanjing Institute of Geography and Limnology, Chinese Academy of Sciences) for providing some of the field data. Financial support was provided by the National Natural Science Foundation of China (41771366, 41671358, 41431176, 41971309) and the program of China Scholarship Council (201904910725). We further appreciate reviews provided by three anonymous reviewers. We acknowledge data support from Lake-Watershed Science Data Center, National Earth System Science Data Sharing Infrastructure, National Science & Technology Infrastructure of China (http://lake.geodata.cn). The BST model used in this study can be accessed from a GitHub repository (https://github.com/zgcao/bst_oli).

Appendix

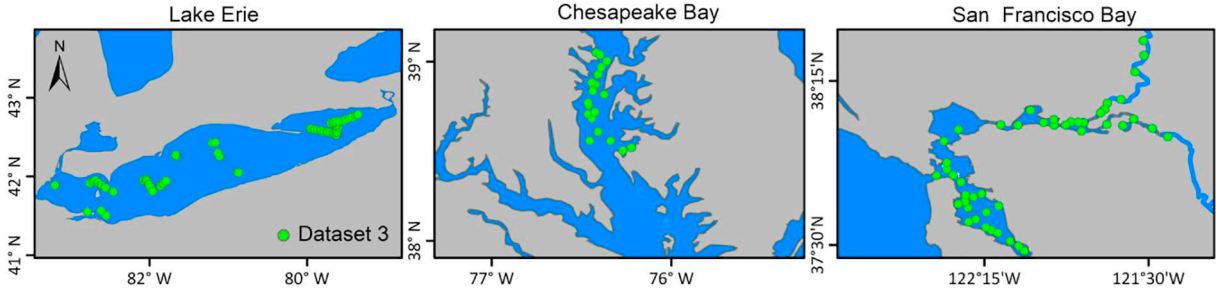

Fig. A1. Locations of Dataset-3 retrieved from SeaBASS in Lake Erie, Chesapeake Bay, and San Francisco Bay.

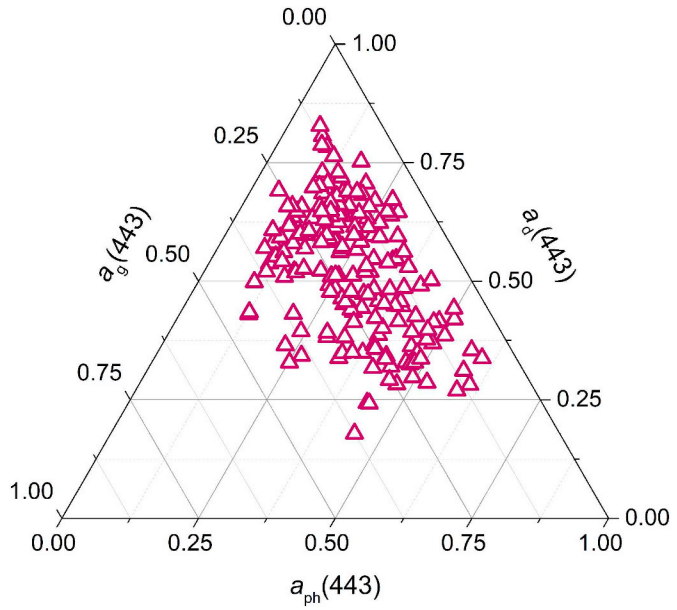

Fig. A2. Comparisons of absorption coefficients at 443 nm for phytoplankton [$a_{ph}(443)$], SPIM [$a_d(443)$], and CDOM [$a_g(443)$] in Dataset-1.

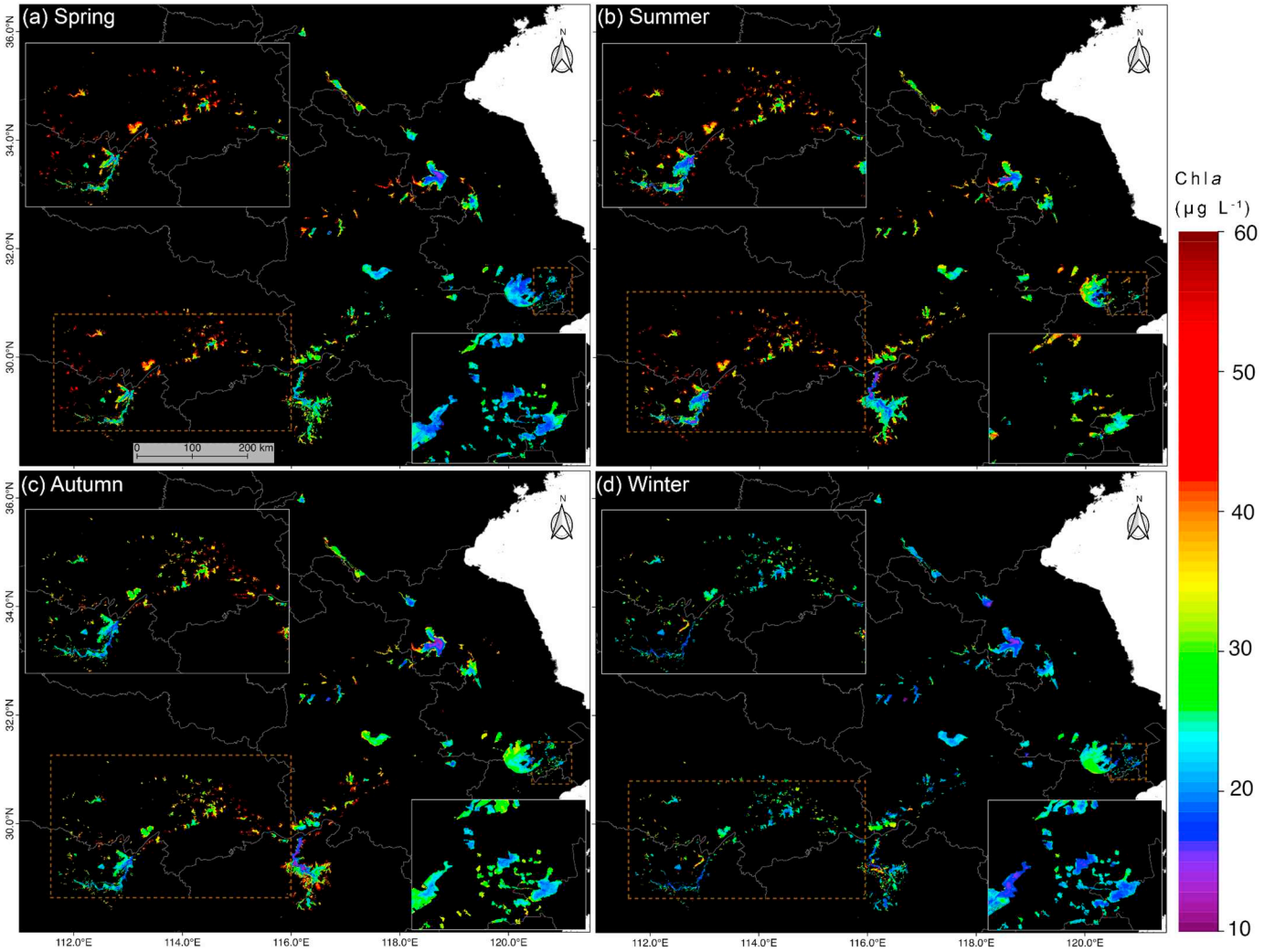

Fig. A3. Seasonal mean Chl*a* in the spring(a), summer(b), autumn(c) and winter(d) in the lakes in eastern China plain during 2013–2018.

References

Adrian, R., Reilly, C.M.O., Zagarese, H., Baines, S.B., Hessen, D.O., Keller, W., Livingstone, D.M., Sommaruga, R., Straile, D., Van Donk, E., 2009. Lakes as sentinels of climate change. Limnol. Oceanogr. 54, 2283–2297.

Aurin, D., Mannino, A., Franz, B., 2013. Spatially resolving ocean color and sediment dispersion in river plumes, coastal systems, and continental shelf waters. Remote Sens. Environ. 137, 212–225.

Barsi, J., Lee, K., Kvaran, G., Markham, B., Pedelty, J., 2014. The spectral response of the Landsat-8 operational land imager. Remote Sens. 6, 10232–10251.

Cao, Z., Duan, H., Feng, L., Ma, R., Xue, K., 2017. Climate- and human-induced changes in suspended particulate matter over Lake Hongze on short and long timescales. Remote Sens. Environ. 192, 98–113.

Cao, Z., Ma, R., Duan, H., Xue, K., 2019a. Effects of broad bandwidth on the remote sensing of inland waters: implications for high spatial resolution satellite data applications. ISPRS J. Photogramm. Remote Sens. 153, 110–122.

Cao, Z.G., Ma, R.H., Duan, H.T., Xue, K., Shen, M., 2019b. Effect of satellite temporal resolution on long-term suspended particulate matter in inland lakes. Remote Sens. 11, 1–18.

Chen, T., Guestrin, C., 2016. XGBoost: A scalable tree boosting system. In. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, San Francisco, California, USA, pp. 785–794.

Chen, S., Hu, C., Barnes, B.B., Wanninkhof, R., Cai, W.-J., Barbero, L., Pierrot, D., 2019. A machine learning approach to estimate surface ocean pCO2 from satellite measurements. Remote Sens. Environ. 228, 203–226.

Concha, J.A., Schott, J.R., 2016. Retrieval of color producing agents in Case 2 waters using Landsat 8. Remote Sens. Environ. 185, 95–107.

Craig, S.E., Jones, C.T., Li, W.K.W., Lazin, G., Horne, E., Caverhill, C., Cullen, J.J., 2012. Deriving optical metrics of coastal phytoplankton biomass from ocean colour. Remote Sens. Environ. 119, 72–83.

Dall'Olmo, G., Gitelson, A.A., Rundquist, D.C., Leavitt, B., Barrow, T., Holz, J.C., 2005. Assessing the potential of SeaWiFS and MODIS for estimating chlorophyll concentration in turbid productive waters using red and near-infrared bands. Remote Sens. Environ. 96, 176–187.

Dekker, A.G., Peters, S.W.M., 1993. The use of the Thematic Mapper for the analysis of eutrophic lakes: a case study in the Netherlands. Int. J. Remote Sens. 14, 799–821.

Downing, J.A., Prairie, Y.T., Cole, J.J., Duarte, C.M., Tranvik, L.J., Striegl, R.G., McDowell, W.H., Kortelainen, P., Caraco, N.F., Melack, J.M., Middelburg, J.J., 2006. The global abundance and size distribution of lakes, ponds, and impoundments. Limnol. Oceanogr. 51, 2388–2397.

Duan, H., Zhang, Y., Zhang, B., Song, K., Wang, Z., 2007. Assessment of chlorophyll-a concentration and trophic state for Lake Chagan using Landsat TM and field spectral data. Environ. Monit. Assess. 129, 295–308.

Duan, H., Ma, R., Xu, X., Kong, F., Zhang, S., Kong, W., Hao, J., Shang, L., 2009. Two-decade reconstruction of algal blooms in China's Lake Taihu. Environ. Sci. Technol. 43, 3522–3528.

Duan, H., Ma, R., Hu, C., 2012. Evaluation of remote sensing algorithms for cyanobacterial pigment retrievals during spring bloom formation in several lakes of East China. Remote Sens. Environ. 126, 126–135.

Fan, Y., Li, W., Gatebe, C.K., Jamet, C., Zibordi, G., Schroeder, T., Stamnes, K., 2017. Atmospheric correction over coastal waters using multilayer neural networks. Remote Sens. Environ. 199, 218–240.

Feng, L., Hu, C., Chen, X., Tian, L., Chen, L., 2012. Human induced turbidity changes in Poyang Lake between 2000 and 2010: observations from MODIS. J. Geophys. Res. 117.

Feng, L., Hu, C., Han, X., Chen, X., Qi, L., 2014. Long-term distribution patterns of chlorophyll-a concentration in China's largest freshwater Lake: MERIS full-resolution observations with a practical approach. Remote Sens. 7, 275–299.

Feng, L., Hou, X., Zheng, Y., 2019. Monitoring and understanding the water transparency changes of fifty large lakes on the Yangtze Plain based on long-term MODIS observations. Remote Sens. Environ. 221, 675–686.

Franz, B.A., Bailey, S.W., Kuring, N., Werdell, P.J., 2015. Ocean color measurements with the operational land imager on Landsat-8: implementation and evaluation in SeaDAS. J. Appl. Remote. Sens. 9.

Ghatkar, J.G., Singh, R.K., Shanmugam, P., 2019. Classification of algal bloom species from remote sensing data using an extreme gradient boosted decision tree model. Int. J. Remote Sens. 40, 9412–9438.

Giardino, C., Pepe, M., Brivio, P.A., Ghezzi, P., Zilioli, E., 2001. Detecting chlorophyll, Secchi disk depth and surface temperature in a sub-alpine lake using Landsat imagery. Sci. Total Environ. 268, 19–29.

Gilerson, A.A., Gitelson, A.A., Zhou, J., Gurlin, D., Moses, W., Ioannou, I., Ahmed, S.A., 2010. Algorithms for remote estimation of chlorophyll-a in coastal and inland waters using red and near infrared bands. Opt. Express 18, 24109–24125.

Gitelson, A., 1992. The peak near 700 nm on radiance spectra of algae and water: relationships of its magnitude and position with chlorophyll concentration. Int. J. Remote Sens. 13, 3367–3373.

Gitelson, A.A., Dall'Olmo, G., Moses, W., Rundquist, D.C., Barrow, T., Fisher, T.R., Gurlin, D., Holz, J., 2008. A simple semi-analytical model for remote estimation of chlorophyll-a in turbid waters: validation. Remote Sens. Environ. 112, 3582–3593.

Gons, H.J., 1999. Optical teledetection of chlorophyll a in turbid inland waters. Environ. Sci. Technol. 33, 1127–1132.

Gons, H.J., Auer, M.T., Effler, S.W., 2008. MERIS satellite chlorophyll mapping of oligotrophic and eutrophic waters in the Laurentian Great Lakes. Remote Sens. Environ. 112, 4098–4106.

Ha, N.T.T., Koike, K., Nhuan, M.T., Canh, B.D., Thao, N.T.P., Parsons, M., 2017. Landsat 8/OLI two bands ratio algorithm for chlorophyll-a concentration mapping in hypertrophic waters: an application to West Lake in Hanoi (Vietnam). IEEE . Sel. Top. Appl. Earth Observ. Remote Sensing 10, 4919–4929.

Hou, X., Feng, L., Duan, H., Chen, X., Sun, D., Shi, K., 2017. Fifteen-year monitoring of the turbidity dynamics in large lakes and reservoirs in the middle and lower basin of the Yangtze River, China. Remote Sens. Environ. 190, 107–121.

Hu, C., Lee, Z., Ma, R., Yu, K., Li, D., Shang, S., 2010. Moderate resolution imaging Spectroradiometer (MODIS) observations of cyanobacteria blooms in Taihu Lake, China. J. Geophys. Res. Oceans 115, C04002.

Hunter, P.D., Tyler, A.N., Gilvear, D.J., Willby, N.J., 2009. Using remote sensing to aid the assessment of human health risks from blooms of potentially toxic cyanobacteria. Environ. Sci. Technol. 43, 2627–2633.

Ilori, C.O., Pahlevan, N., Knudby, A., 2019. Analyzing performances of different atmospheric correction techniques for Landsat 8: application for coastal remote sensing. Remote Sensing 11, 469.

Jeffrey, S.t., Humphrey, G., 1975. New spectrophotometric equations for determining chlorophylls a, b, c1 and c2 in higher plants, algae and natural phytoplankton. Biochem. Physiol. Pflanz. 167, 191–194.

Kravitz, J., Matthews, M., Bernard, S., Griffith, D., 2020. Application of Sentinel 3 OLCI for chl-a retrieval over small inland water targets: successes and challenges. Remote Sens. Environ. 237, 111562.

Kuhn, C., de Matos Valerio, A., Ward, N., Loken, L., Sawakuchi, H.O., Kampel, M., Richey, J., Stadler, P., Crawford, J., Striegl, R., Vermote, E., Pahlevan, N., Butman, D., 2019. Performance of Landsat-8 and Sentinel-2 surface reflectance products for river remote sensing retrievals of chlorophyll-a and turbidity. Remote Sens. Environ. 224, 104–118.

Le, C., Li, Y., Zha, Y., Sun, D., Huang, C., Lu, H., 2009. A four-band semi-analytical model for estimating chlorophyll a in highly turbid lakes: the case of Taihu Lake, China. Remote Sens. Environ. 113, 1175–1182.

Li, J., Roy, D.P., 2017. A global analysis of sentinel-2A, sentinel-2B and Landsat-8 data revisit intervals and implications for terrestrial monitoring. Remote Sens. 9, 902.

Li, J., Sheng, Y., 2012. An automated scheme for glacial lake dynamics mapping using Landsat imagery and digital elevation models: a case study in the Himalayas. Int. J. Remote Sens. 33, 5194–5213.

Liu, D., Duan, H., Yu, S., Shen, M., Xue, K., 2019. Human-induced eutrophication dominates the bio-optical compositions of suspended particles in shallow lakes: implications for remote sensing. Sci. Total Environ. 667, 112–123.

Liu, G., Li, L., Song, K., Li, Y., Lyu, H., Wen, Z., Fang, C., Bi, S., Sun, X., Wang, Z., Cao, Z., Shang, Y., Yu, G., Zheng, Z., Huang, C., Xu, Y., Shi, K., 2020. An OLCI-based algorithm for semi-empirically partitioning absorption coefficient and estimating chlorophyll a concentration in various turbid case-2 waters. Remote Sens. Environ. 239, 111648.

Lorenzen, C.J., 1967. Determination of chlorophyll and pheo-pigments: spectrophotometric equations. Limnol. Oceanogr. 12, 343–346.

Ma, R., Tang, J., Dai, J., 2006. Bio-optical model with optimal parameter suitable for Taihu Lake in water colour remote sensing. Int. J. Remote Sens. 27, 4305–4328.

Ma, R., Yang, G., Duan, H., Jiang, J., Wang, S., Feng, X., Li, A., Kong, F., Xue, B., Wu, J., Li, S., 2011. China's lakes at present: number, area and spatial distribution. Sci. China Earth Sci. 54, 283–289.

Mishra, S., Mishra, D.R., 2012. Normalized difference chlorophyll index: a novel model for remote estimation of chlorophyll-a concentration in turbid productive waters. Remote Sens. Environ. 117, 394–406.

Mobley, C.D., 1999. Estimation of the remote-sensing reflectance from above-surface measurements. Appl. Opt. 38, 7442–7455.

Mueller, J.L., McClain, C.R., Fargion, G.S., Bidigare, R., Trees, C., Balch, W., Dore, J., Drapeau, D., Karl, D., Van, L., 2003. Ocean optics protocols for satellite ocean color sensor validation, revision 5, volume V: biogeochemical and bio-optical measurements and data analysis protocols. NASA Tech. Memo 211621, 36.

Neil, C., Spyrakos, E., Hunter, P.D., Tyler, A.N., 2019. A global approach for chlorophyll-a retrieval across optically complex inland waters based on optical water types. Remote Sens. Environ. 229, 159–178.

Page, B.P., Kumar, A., Mishra, D.R., 2018. A novel cross-satellite based assessment of the spatio-temporal development of a cyanobacterial harmful algal bloom. Int. J. Appl. Earth Obs. Geoinf. 66, 69–81.

Pahlevan, N., Lee, Z.P., Wei, J.W., Schaaf, C.B., Schott, J.R., Berk, A., 2014. On-orbit radiometric characterization of OLI (Landsat-8) for applications in aquatic remote sensing. Remote Sens. Environ. 154, 272–284.

Pahlevan, N., Roger, J.C., Ahmad, Z., 2017a. Revisiting short-wave-infrared (SWIR) bands for atmospheric correction in coastal waters. Opt. Express 25, 6015–6035.

Pahlevan, N., Schott, J.R., Franz, B.A., Zibordi, G., Markham, B., Bailey, S., Schaaf, C.B., Ondrusek, M., Greb, S., Strait, C.M., 2017b. Landsat 8 remote sensing reflectance (Rrs) products: evaluations, intercomparisons, and enhancements. Remote Sens. Environ. 190, 289–301.

Pahlevan, N., Chittimalli, S.K., Balasubramanian, S.V., Vellucci, V., 2019. Sentinel-2/Landsat-8 product consistency and implications for monitoring aquatic systems. Remote Sens. Environ. 220, 19–29.

Pahlevan, N., Smith, B., Schalles, J., Binding, C., Cao, Z., Ma, R., Alikas, K., Kangro, K., Gurlin, D., Hà, N., Matsushita, B., Moses, W., Greb, S., Lehmann, M.K., Ondrusek, M., Oppelt, N., Stumpf, R., 2020. 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. 240.

Palmer, S.C.J., Kutser, T., Hunter, P.D., 2015. Remote sensing of inland waters: challenges, progress and future directions. Remote Sens. Environ. 157, 1–8.

Pinckney, J., Papa, R., Zingmark, R., 1994. Comparison of high-performance liquid chromatographic, spectrophotometric, and fluorometric methods for determining chlorophyll a concentrations in estaurine sediments. J. Microbiol. Methods 19,

59–66.
Prasad, S., Saluja, R., Garg, J.K., 2020. Assessing the efficacy of Landsat-8 OLI imagery derived models for remotely estimating chlorophyll-a concentration in the Upper Ganga River, India. Int. J. Remote Sens. 41, 2439–2456.
Pyo, J., Duan, H., Baek, S., Kim, M.S., Jeon, T., Kwon, Y.S., Lee, H., Cho, K.H., 2019. A convolutional neural network regression for quantifying cyanobacteria using hyperspectral imagery. Remote Sens. Environ. 233.
Qi, L., Hu, C., Duan, H., Barnes, B.B., Ma, R., 2014. An EOF-based algorithm to estimate chlorophyll a concentrations in Taihu Lake from MODIS land-band measurements: implications for near real-time applications and forecasting models. Remote Sens. 6, 10694–10715.
Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., Prabhat, 2019. Deep learning and process understanding for data-driven Earth system science. Nature 566, 195–204.
Ruddick, K.G., Ovidio, F., Rijkeboer, M., 2000. Atmospheric correction of SeaWiFS imagery for turbid coastal and inland waters. Appl. Opt. 39, 897–912.
Ruddick, K.G., Gons, H.J., Rijkeboer, M., Tilstone, G., 2001. Optical remote sensing of chlorophyll a in case 2 waters by use of an adaptive two-band algorithm with optimal error properties. Appl. Opt. 40, 3575–3585.
Sayers, M.J., Grimm, A.G., Shuchman, R.A., Deines, A.M., Bunnell, D.B., Raymer, Z.B., Rogers, M.W., Woelmer, W., Bennion, D.H., Brooks, C.N., Whitley, M.A., Warner, D.M., Mychek-Londer, J., 2015. A new method to generate a high-resolution global distribution map of lake chlorophyll. Int. J. Remote Sens. 36, 1942–1964.
Seegers, B.N., Stumpf, R.P., Schaeffer, B.A., Loftin, K.A., Werdell, P.J., 2018. Performance metrics for the assessment of satellite data products: an ocean color case study. Opt. Express 26, 7404–7422.
Shi, K., Zhang, Y., Xu, H., Zhu, G., Qin, B., Huang, C., Liu, X., Zhou, Y., Lv, H., 2015. Long-term satellite observations of microcystin concentrations in Lake Taihu during Cyanobacterial bloom periods. Environ. Sci. Technol. 49, 6448–6456.
Siegel, D.A., Wang, M., Maritorena, S., Robinson, W., 2000. Atmospheric correction of satellite ocean color imagery: the black pixel assumption. Appl. Opt. 39, 3582–3591.
Smith, M.E., Robertson Lain, L., Bernard, S., 2018. An optimized Chlorophyll a switching algorithm for MERIS and OLCI in phytoplankton-dominated waters. Remote Sens. Environ. 215, 217–227.
Spyrakos, E., O'Donnell, R., Hunter, P.D., Miller, C., Scott, M., Simis, S.G.H., Neil, C., Barbosa, C.C.F., Binding, C.E., Bradt, S., Bresciani, M., Dall'Olmo, G., Giardino, C., Gitelson, A.A., Kutser, T., Li, L., Matsushita, B., Martinez-Vicente, V., Matthews, M.W., Ogashawara, I., Ruiz-Verdú, A., Schalles, J.F., Tebbs, E., Zhang, Y., Tyler, A.N., 2017. Optical types of inland and coastal waters. Limnol. Oceanogr. 63, 846–870.
Vanhellemont, Q., Ruddick, K., 2018. Atmospheric correction of metre-scale optical satellite data for inland and coastal water applications. Remote Sens. Environ. 216, 586–597.
Wang, M., Jiang, L., 2018. Atmospheric correction using the information from the short blue band. IEEE Trans. Geosci. Remote Sens. 1–14.
Wang, S., Li, J., Zhang, B., Spyrakos, E., Tyler, A.N., Shen, Q., Zhang, F., Kuster, T., Lehmann, M.K., Wu, Y., Peng, D., 2018. Trophic state assessment of global inland waters using a MODIS-derived Forel-Ule index. Remote Sens. Environ. 217, 444–460.
Wang, D., Ma, R., Xue, K., Loiselle, S.A., 2019. The assessment of Landsat-8 OLI atmospheric correction algorithms for inland waters. Remote Sens. 11, 169.
Watanabe, F., Alcantara, E., Rodrigues, T., Rotta, L., Bernardo, N., Imai, N., 2018. Remote sensing of the chlorophyll-a based on OLI/Landsat-8 and MSI/Sentinel-2A (Barra Bonita reservoir, Brazil). Anais da Academia Brasileira de Ciencias 90, 1987–2000.
Wright, S.W., Jeffrey, S.W., Mantoura, R.F.C., Llewellyn, C.A., Bjørnland, T., Repeta, D., Welschmeyer, N., 1991. Improved HPLC method for the analysis of chlorophylls and carotenoids from marine phytoplankton. Mar. Ecol. Prog. Ser. 77, 183–196.
Wu, Z., Cai, Y., Liu, X., Xu, C.P., Chen, Y., Zhang, L., 2013. Temporal and spatial variability of phytoplankton in Lake Poyang: the largest freshwater lake in China. J. Great Lakes Res. 39, 476–483.