



Inland harmful cyanobacterial bloom prediction in the eutrophic Tri An Reservoir using satellite band ratio and machine learning approaches

Hao-Quang Nguyen¹ · Nam-Thang Ha^{2,3} · Thanh-Luu Pham^{4,5}

Received: 26 March 2019 / Accepted: 23 December 2019 / Published online: 8 January 2020
© Springer-Verlag GmbH Germany, part of Springer Nature 2020

Abstract

In recent years, Tri An, a drinking water reservoir for millions of people in southern Vietnam, has been affected by harmful cyanobacterial blooms (HCBs), raising concerns about public health. It is, therefore, crucial to gain insights into the outbreak mechanism of HCBs and understand the spatiotemporal variations of chlorophyll-a (Chl-a) in this highly turbid and productive water. This study aims to evaluate the predictable performance of both approaches using satellite band ratio and machine learning for Chl-a concentration retrieval—a proxy of HCBs. The monthly water quality samples collected from 2016 to 2018 and 23 cloud free Sentinel-2A/B scenes were used to develop Chl-a retrieval models. For the band ratio approach, a strong linear relationship with in situ Chl-a was found for two-band algorithm of Green-NIR. The band ratio-based model accounts for 72% of variation in Chl-a concentration from 2016 to 2018 datasets with an RMSE of 5.95 $\mu\text{g/L}$. For the machine learning approach, Gaussian process regression (GPR) yielded superior results for Chl-a prediction from water quality parameters with the values of 0.79 (R^2) and 3.06 $\mu\text{g/L}$ (RMSE). Among various climatic parameters, a high correlation ($R^2 = 0.54$) between the monthly total precipitation and Chl-a concentration was found. Our analysis also found nitrogen-rich water and TSS in the rainy season as the driving factors of observed HCBs in the eutrophic Tri An Reservoir (TAR), which offer important solutions to the management of HCBs in the future.

Keywords Tri An Reservoir · Sentinel-2A/B · Chlorophyll-a · Harmful cyanobacterial blooms · Band ratio regression · Machine learning · Gaussian process regression

Responsible editor: Vitor Manuel Oliveira Vasconcelos

✉ Thanh-Luu Pham
pt.luu@hutech.edu.vn

¹ Graduate School of Systems and Information Engineering, University of Tsukuba, Tsukuba 3058573, Japan

² Environmental Research Institute, School of Science, University of Waikato, Hamilton 3260, New Zealand

³ Faculty of Fisheries, University of Agriculture and Forestry, Hue University, Thua Thien Hue 530000, Vietnam

⁴ Ho Chi Minh City University of Technology (HUTECH), 475A Dien Bien Phu Street, Ward 25, Binh Thanh District, Ho Chi Minh City 70000, Vietnam

⁵ Vietnam Academy of Science and Technology (VAST), Institute of Tropical Biology, 85 Tran Quoc Toan Street, District 3, Ho Chi Minh City 70000, Vietnam

Introduction

Among the most common lake/reservoir problems is harmful cyanobacterial blooms (HCBs), the consequence of excessive release of nutrients and pollutants from anthropogenic activities (Paerl 2017). HCBs have been recognized as an emerging issue, causing a broad range of environmental, social, and economic damage (Lee et al. 2015; Schaeffer et al. 2018). HCBs, for instance, may lead to incidents of hypoxic or anoxic conditions causing mortality (Chorus and Bartram 1999). Surface blooms have caused the degradation of water quality and have had negative effects on recreational opportunities and the economy (Paerl and Huisman 2008). HCBs are also well known for their toxic secondary metabolites, known as cyanotoxins, including hepatotoxins, neurotoxins, and dermatotoxic compounds. These toxins have had detrimental effects on higher trophic levels, mortality, and illness in aquatic animals as well as adverse health risks to humans (Pham and Utsumi 2018). The impacts of HCBs on human life have

been exacerbated as a result of eutrophication and global warming (Paerl and Paul 2012; Visser et al. 2016).

It is, therefore, crucial to monitor and understand the spatial and temporal variations of Chl-a concentration, representing HCBs, in the raw water storage units. Recently, remote sensing has been strongly suggested as a practical approach not only for Chl-a long-term observation but also water quality analysis, mainly because of its capability to capture synoptic data of a large area during the algal bloom (Bresciani et al. 2018). This new approach contrasts with the traditional field-based methods, which are usually costly, labor intensive, and have a low frequency of in situ measurement (Mu et al. 2019; Quang et al. 2017; Schaeffer et al. 2018). An accurate remote estimation of Chl-a concentration in turbid productive waters is essential for large-scale and multi-temporal studies. However, the deficiency of appropriate satellite sensors and Chl-a retrieval model have left researchers with unresolved challenges (Le et al. 2009; Mishra and Mishra 2012; Toming et al. 2016).

Numerous studies have used satellite data to monitor the occurrence of algae blooms in coastal and inland waters, most of which follow models based on the correlation between the inherent optical properties (or apparent optical properties) and the water quality parameters (Bresciani et al. 2018; Lins et al. 2017; Quang et al. 2017; Zhang et al. 2016). It has been recognized that, as a result of the complex interaction between the inner and outer constituents, the variation of Chl-a concentration in water usually results in a nonlinear relationship between phytoplankton abundance and a group of water quality, hydrology, and meteorology factors (Lou et al. 2016; Yi et al. 2018a). Moreover, due to the presence of multiple constituents such as detritus, non-algal particles (NAPs), and colored dissolved organic matter (CDOM), the use of remote sensing for monitoring Chl-a in inland waters has been far less successful compared to their application in open oceans (Chen et al. 2017; Li et al. 2018; Liu and Tang 2012). To overcome such limitations, the local-based satellite band ratio has been preferred, or most recently, the advanced machine learning methods have been contributing various practical models to Chl-a retrieval in lakes/reservoirs.

In general, the input data of machine learning models consist of either remote sensing reflectance (as reviewed above) or water quality parameters. Normally, the latter approach with water quality data inheres in the apparent advantages due to the certainty of water sampling and analysis, which consequently assures the accuracy of the input data for the model's performance. The selected research papers in this group include artificial neural network (ANN) with back propagation and/or support vector machine regression (SVR) (Chen et al. 2017; Kown et al. 2018; Park et al. 2015; Wang et al. 2018; Xie et al. 2012), principle component analysis and multivariate linear regression (Keller et al. 2018); more recently, extreme learning machine has considerably contributed to

research in this field (Lou et al. 2016; Yi et al. 2018a). In most cases, SVR is preferred, mainly because of its proven advantages with an accepted accuracy for both training and test phase (Karamizadeh et al. 2014). To reduce the amount of input data, Li et al. (2018) applied a minimum redundancy/maximum relevance (mRMR) and random forest to select the key factor for random forest and support vector machine models.

In Vietnam, HCBs occur consistently and at a higher frequency in both inland and coastal waters. However, the prediction of HCBs has only been examined using physical and band ratio-based models from the satellite data (Dippner et al. 2011; Ha et al. 2017a; Ha et al. 2017b; Liu and Tang 2012; Tang et al. 2004). The machine learning methods have seldom been used for this task, despite their proven performance (Blix et al. 2017; Blix and Eltoft 2018b; Bui et al. 2017; Keller et al. 2018). In addition, the question of using a linear model for Chl-a prediction remains valid in case of complex optical properties of water. Hence, there is a notable deficiency of knowledge about bio-optical variability in freshwater systems, which is exploiting machine learning.

Using Tri An as a typical case study for eutrophic deep reservoir, a detailed assessment of machine learning and satellite band ratio regression approaches was performed to evaluate the predictabilities of diverse ensemble models. The aims of this study are as follows: (a) to predict Chl-a concentration using band ratio regression, extracted from remote sensing data and machine learning algorithms, exploiting water quality parameters and comparing their results to recommend the one with better performance and (b) to analyze the spatiotemporal variation and elucidate the mechanism of HCBs in TAR. It is hoped that this work will contribute to an initial assessment of the variability of HCBs in the highly turbid reservoirs of Vietnam.

Methods

Study site description

The Tri An Reservoir is one of the biggest reservoirs in Vietnam, located in Dinh Quan district, Dong Nai province, within a quadrat bounded by 11°05′–11°17′N, 106°58′–107°16′E (Fig. 1). The reservoir is designed for multiple purposes, involving drinking and industrial water supply, agricultural irrigation and fisheries, recreational and tourist resources, flood control, and hydropower operation. Its surface area, maximum depth, mean depth, and volume are respectively 320 km², 27 m, 8.5 m, and 2.7 billion m³. The annual mean values of rainfall, air temperature, and wind speed are roughly estimated to be 2400 mm, 33 °C, and 9 m/s, respectively. During the past decade, a high frequency of HCBs has been recorded, dominated by *Microcystis* and *Anabaena* colonies

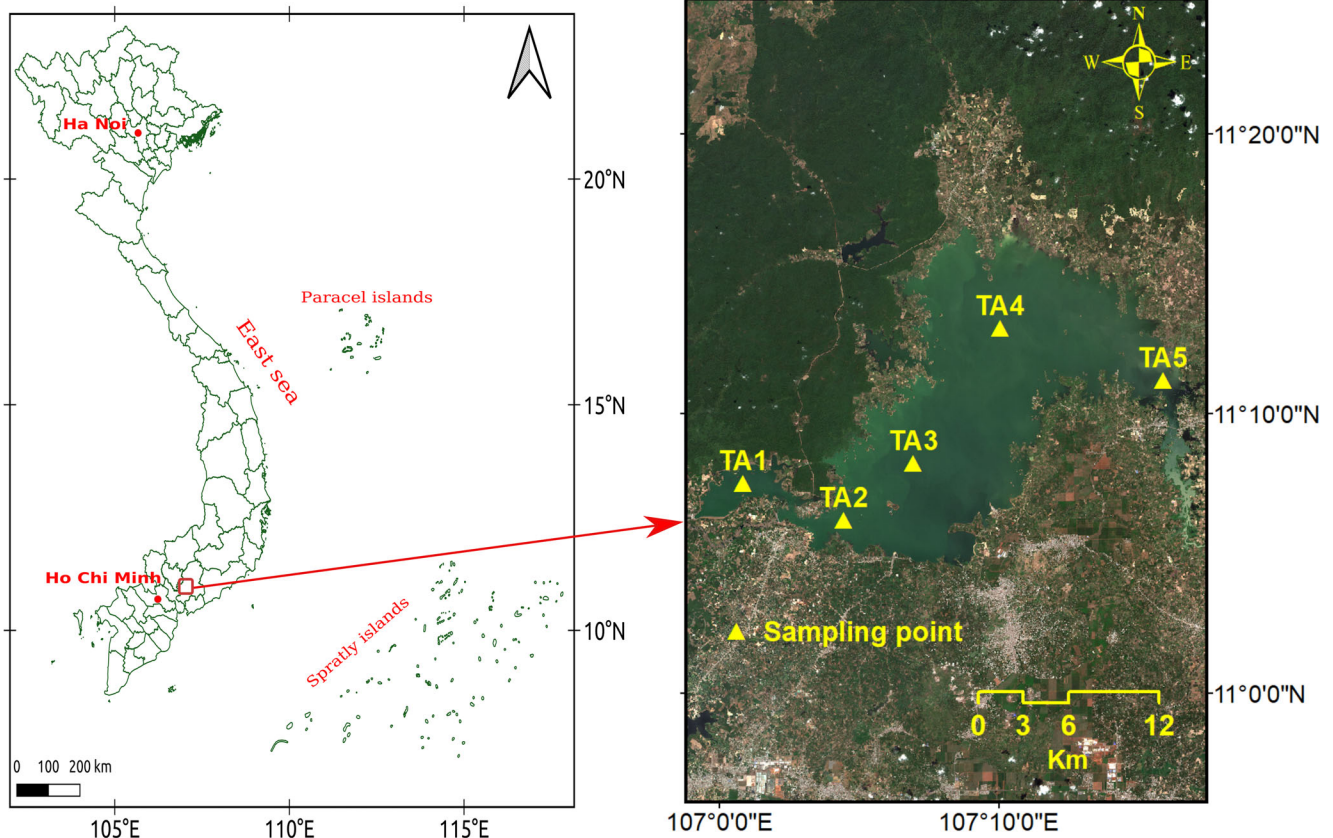


Fig. 1 Location of the eutrophic Tri An Reservoir and the five monthly sampling stations (yellow triangulars)

with the presence of cyanotoxins (Dao et al. 2016). Based on the data of TN (0.25–1.3 mg/L) and TP (0.05–0.14 mg/L) concentrations, TAR falls into the eutrophic category.

Water parameters measurement

The data on water quality were collected monthly from five monitoring stations at a 2-m depth from April 2016 to February 2018 (Fig. 1). The samples were preserved with ice in the field until further processing in the laboratory on the same day. Water pH, temperature, and DO were measured in situ with a multi-detector (WTW Multi 3320, Weilheim, Germany), while Secchi disk was used for determining transparency. To identify Chl-a fraction in water samples, a known volume of raw water samples (100–300 mL) was filtered through glass-fiber filters (Whatman GF/C, England), then Chl-a was extracted using 90% acetone overnight in the dark at 4 °C. After centrifugation, Chl-a concentration was measured at 630–750 nm using a spectrophotometer (UV-VIS, Harch, 500) and calculated using the trichromatic equations (APHA, 2005).

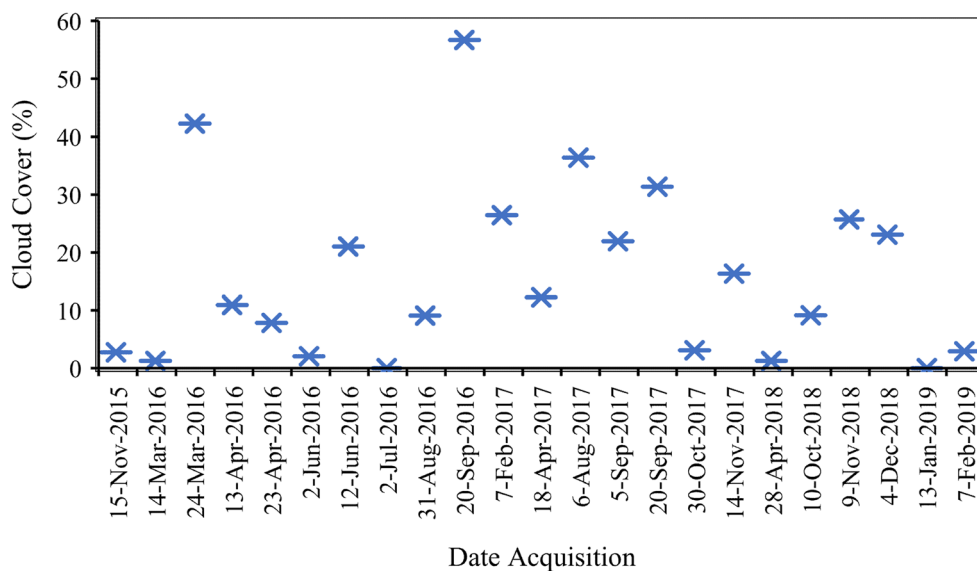
Chemical parameters were analyzed colorimetrically in triplicate with a spectrophotometer (Hach DR/2010) using the following APHA (2005) methods: nitrate 4500NO₃⁻ (B), phosphate 4500PO₄³⁻ (B), total nitrogen Kjeldahl, 4500 N (C), and total phosphorous 4500P (D). To measure the total

suspended solids (TSS), 300–400 mL of raw water samples were filtered into a pre-weighed glass-fiber filter and dried completely at 95 ± 5 °C. The TSS concentration was estimated gravimetrically. In addition, the monthly rainfall and wind speed data, published by the Southern Regional HydroMeteorological Center (Vietnam), were collected in order to elucidate the driving factors, generating a high Chl-a concentration variation and HCBs mechanism in TAR.

Image processing

In this study, band ratio-based model was developed using Sentinel 2A/B and in situ Chl-a data. The Multispectral Instrument (MSI), launched on 23 June 2015 for 2A and 07 March 2017 for 2B, was a filter-based push-broom-type imager, acquiring imagery every 5 days. The MSI sensor observes the Earth at 13 spectral bands, spreads over the VNIR and SWIR domains (443–2190 nm) with spatial resolutions, ranging from 10 to 60 m (Gascon et al. 2017). Level-1C, orthorectified georeferenced, and radiometrically calibrated to Top-Of-Atmosphere (TOA) reflectance image was downloaded from Sentinels Scientific Data Hub and performed on the Sentinel Application Platform (SNAP) version 6.0 on Windows 10 (64-bit). In particular, a series of 23 cloud free images (Fig. 2), acquired from November 2015 to

Fig. 2 Date acquisition and cloud cover (%) of obtained Sentinel-2A/B images



February 2019, were used to develop the model and analyze the spatial patterns of Chl-a concentration. In this study, cloud cover was estimated for the whole image with no cloud above the water body (Fig. 2). After cloud masking and removal using ArcGIS, the cloud-free time series were used to process next steps.

Atmospheric correction was carried out in order to remove the noises from the aerosol particles in the atmosphere. There are several atmospheric correction methods, including Sen2Cor, 6SV, ACOLITE, DOS (Dark Object Subtraction), and ATCOR, and the evaluation of the best atmospheric corrections is still ongoing in the scientific community (Martins et al. 2017; Chen et al. 2017; Sola et al. 2018). In this study, we employed Sen2Cor to perform correction of atmospheric effects, since it commonly outperforms in the highly turbid waters (Grendaité et al. 2018; Mueller-Wilm et al. 2018; Sola et al. 2018). Furthermore, the ATCOR algorithm-based Sen2Cor processor has recently been renovated to improve accuracy for deriving the surface reflectance over water by using the surfaces of the Climate Change Initiative Land Cover (Mueller-Wilm et al. 2018). Hence, Sen2Cor was used to calibrate TOA reflectance to surface water reflectance (R_w). Then, the R_w values from band 1 to band 7 were used as input for the band ratio model approach (Fig. 3a).

We do not have in situ reflectance measurements from the reservoir under investigation carried out simultaneously with the Sentinel-2 overpass. In order to validate the results obtained by Sen2Cor method, we used Sentinel-2 Level-2A atmospherically corrected data commenced from the Open Hub on 2 May 2017 and subsequently on the ServHub (Adriana and Richard 2017). In total, 60 points were randomly extracted from two images Level-2A on Jan. 13, 2019, and Jan. 28, 2019 to validate the robustness of atmospheric correction in TAR. The correlation between surface water reflectance

calculated from Sen2Cor and the one extracted from Level-2A atmospherically corrected data supplied by the ServHub indicates a very good atmospheric correction for water pixels in TAR (Fig. 3b).

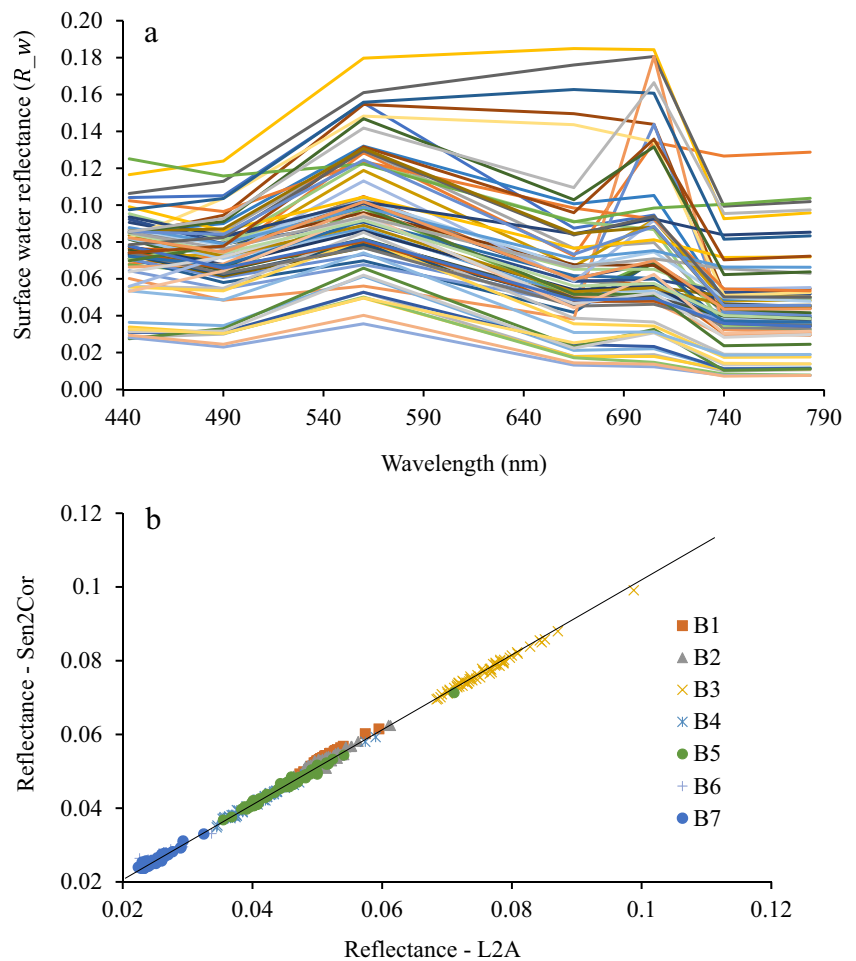
Band ratio regression model development

Several studies have discussed the time gap between in situ measurements and the satellite overpass, indicating a maximum of ± 8 days is reasonable in case of the stationary condition of the water environment (Johnson et al. 2013; Tan et al. 2017; Maeda et al. 2019). A 5-day lag, therefore, was acceptable to develop Chl-a retrieval algorithm in the present study, since the cyanobacterial blooms in TAR have extended for several weeks (based on field observation).

To eliminate the distortion on water surface reflectance, the average value using a 3×3 pixel box, centered on each sample station, was calibrated to perform a direct comparison with the in situ measurements (Quang et al. 2017).

To select the best satellite bands for the band ratio model, the statistical relationship between a color index (i.e., band ratio) and the in situ measurement of Chl-a was tested (Chen et al. 2017; Grendaité et al. 2018; Mishra et al. 2017). Among the satellite bands, the two- and three-band empirical models using reflectance in red and near-infrared (NIR) spectral regions have commonly been reviewed in inland waters. Therefore, a wide range of spectral band from blue to NIR regions was adopted to develop the linear model for Chl-a retrieval in this study. A total of 22 observations which were below 5 days different from satellite overpass (roughly 32%) were carried out for model development. The remaining 47 observations (roughly 68%) were selected to validate the model's performance.

Fig. 3 The surface water reflectance (R_w) calculated by Sen2Cor in the Tri An Reservoir (a). The correlation between reflectance spectra calculated by Sen2Cor and Sentinel-2 Level-2A atmospherically corrected data supplied by the ServHub (b)



Machine learning regression model performance

In total, 117 observations of nine (9) water quality parameters, including total phosphorous (TP), total nitrogen (TN), total suspended solid (TSS), nitrate (NO_3^-), phosphate (PO_4^{3-}), pH, temperature (temp), transparency (trans), and dissolved oxygen (DO), were considered input data for machine learning models. Due to a wide variation of water quality values, Chl-a concentration, and the difference in measurement unit, all the parameters were log-transformed to keep a normal distribution of the input data. The inputs for the model were selected from a correlation analysis with a coefficient of Pearson correlation higher than 0.3. In this research, four machine learning methods and one multivariate linear regression model (MLR) were compared to test their performance in terms of Chl-a prediction. The machine learning models include Gaussian Processor Regressor (GPR), Random Forest Regressor (RFR), Support Vector Machine Regressor (SVR), and Multi-layer Perceptron Regressor (MLP). Details about the algorithms and their operation can be found at <https://scikit-learn.org/stable/> and in the research paper (Pedregosa et al. 2011). The performance of five models was adapted in

Python environment using scikit-learn library within a two-step processing. At first, the hyper-parameters were tuned using a grid search with a fivefold cross-validation for the RFR, SVR, and MPL models. In case of the GPR model, various kernels were randomly tested to select the best kernel that corresponded to the dataset in this research. Then, the performance of the involved models was tested through the training and validation phases in a 10 cross-validation using Shuffle Split technique for the sampling. For the total dataset, 60 and 40% were respectively divided for training and validation phases.

Evaluation criteria

The performance of the Chl-a retrieval models in TAR was evaluated using the following indicators: square of correlation coefficient (R^2), which provides the variability measure for the data reproduced in the model; root-mean-square error (RMSE) and mean absolute error (MAE), which measure residual errors, providing a global idea of the difference between observation and modeling; explained variance score (EVS) returns a score for the explanation of the variance of the

measured and modeled values, and Bias explains the discrepancies between the measured and simulated Chl-a concentration. See Eqs. (1), (2), (3), (4) and (5) below for the formulas.

$$R^2 = 1 - \frac{\sum (x_i^{measured} - x_i^{estimated})^2}{\sum (x_i^{measured} - x_{mean}^{measured})^2} \tag{1}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^N (x_i^{estimated} - x_i^{measured})^2}{N}} \tag{2}$$

$$MAE = \frac{1}{N} \sum_{i=1}^N (x_i^{estimated} - x_i^{measured})^2 \tag{3}$$

$$Bias = \frac{\sum_{i=1}^N x_i^{estimated} - x_i^{measured}}{N} \tag{4}$$

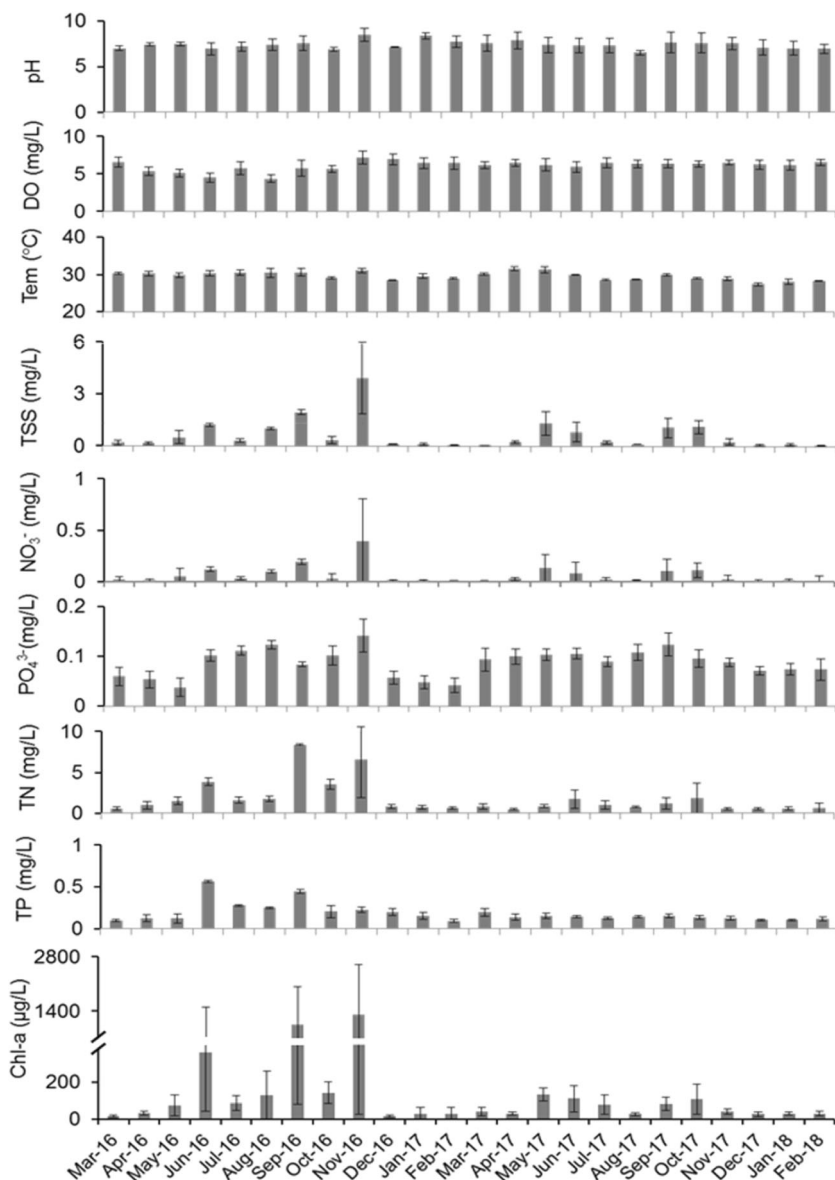
$$EVS = 1 - \frac{Var(x_i^{measured} - x_i^{estimated})}{Var(x_i^{measured})} \tag{5}$$

Results

The variation of water quality parameters

The monthly mean and standard deviation of water quality variables from March 2016 to February 2018 in TAR are shown in Fig. 4. The surface water temperature in TAR did not vary much over the sampling period, ranging from 27.1 to 32.5 °C; however, pH largely changed from 6.0 to 9.0. Transparency exhibited a wide variation from 36 to 191 cm

Fig. 4 Mean values of water quality parameters in the Tri An Reservoir (from Mar 2016 to Feb 2018). Data were presented as mean values ± SD



among five sampling sites, which was a similar trend to TSS (from 0.05 to 3.92 mg/L). The mean dissolved oxygen ranged from 4.4 to 7.2 mg/L. TAR is nitrogen eutrophicated with a higher and wider range of nitrate (0.17–0.6 mg/L) and TN (0.47–8.4 mg/L) compared to phosphate (0.04–0.14 ml/L) and TP (0.09–0.57 mg/L), respectively. A wide range of Chl-a (mean value varied from 15.48 to 1310 $\mu\text{g/L}$) was observed during the sampling period. The highest mean Chl-a concentrations were recorded during May to November and reached several peaks in June (360 $\mu\text{g/L}$), September (500 $\mu\text{g/L}$), and November 2016 (1310 $\mu\text{g/L}$) as heavy blooms occurred (Fig. 5a–c). A repeated cycle began at a low concentration in the dry months (Fig. 5d), with a minimum of 31 $\mu\text{g/L}$ in January 2017 followed by an increasing in May 2017 (Fig. 4).

Algorithms for Chl-a concentration retrieval

Band ratio-based linear regression model

Through all the cross-regression analysis, the relationship between Chl-a and blue-green ratio was low, approximated by linear function. By following the same principle of the blue-green ratio model and considering the ratio between reflectance in the near-infrared and reflectance in the red region, various algorithms were tested for Chl-a retrieval in this study (Table 1). As noted previously, the linear regression-based algorithms for Chl-a retrieving from remote sensing data are

preferred (Grendaitè et al. 2018; Ha et al. 2013; Ha et al. 2017a; Lins et al. 2017; Ritchie et al. 2003). This means that, among various models including linear, quadratic, and exponential functions with similar R^2 and RMSE, the linear regression algorithms will be the optimal selection. For this reason, the model No. 3 with R^2 of 0.72 and RMSE of 5.95 $\mu\text{g/L}$ (highlighted in *italic*) will be used as Chl-a prediction model (Figs. 6 and 7). It is clear that the estimated Chl-a has a small RMSE value compared to the mean in situ Chl-a (varied from 15.48 to 1310 $\mu\text{g/L}$), confirming the appropriateness of our model for estimating Chl-a in TAR when the acquisition times differ. Despite a better performance ($R^2 > 0.74$) of the ratio B3/B6 vs. Chl-a concentration found in this study, this model was not exploited for Chl-a retrieval due to high RMSE (187.03 $\mu\text{g/L}$).

The performance of our new linear model (No. 3, Table 1) was then validated using 47 in situ measurements of Chl-a concentration, retrieved during field campaigns performed synchronous to satellite overpasses, with a maximum of 5-day time difference (Fig. 8). It was also noted that 47 these observations were different from those in the training phase (22 observations). This study demonstrated that the Green - NIR band ratio model was successful in the prediction of Chl-a concentration in TAR, with R^2 greater than 0.70 and with very small values of RMSE and Bias. These results attest to our model's ability to handle the relationship between water surface reflectance and Chl-a concentration.

Fig. 5 Heavy bloom of cyanobacteria in June (a), September (b), November, 2016 (c) and water without bloom (d) from the Tri An Reservoir



Table 1 Regression models for chlorophyll-a retrieval (B1 = 443 nm, B2 = 490 nm, B3 = 560 nm, B4 = 665 nm, B5 = 705 nm, B6 = 740 nm, and B7 = 783 nm)

No.	No. Samples for training	No. Samples for testing	Variables (x)	Model	R^2	RMSE ($\mu\text{g/L}$)	Bias	Mean Chl-a ($\mu\text{g/L}$)
1	22	47	B1/B3 vs $\log_{10}\text{Chl-a}$	$y = -1.0668x + 2.4018$	0.27	3.12	0.02	26
2			B6/B3 vs $\log_{10}\text{Chl-a}$	$y = -1.5854x + 2.3435$	0.69	4.22	0.14	
3			B3/B6 vs $\log_{10}\text{Chl-a}$	$y = 0.3438x + 0.7736$	0.72	5.95	0.24	
4			B3/B6 vs $\log_{10}\text{Chl-a}$	$y = -0.0538x^2 + 0.6149x + 0.469$	0.73	4.53	0.16	
5			B3/B6 vs Chl-a	$y = 36.363x - 36.111$	0.74	187.03	-20.43	
6			B3/B7 vs Chl-a	$y = 37.271x - 37.804$	0.71	186.82	-19.45	
7			B3/B7 vs $\log_{10}\text{Chl}$	$y = 0.355x + 0.7514$	0.70	6.42	0.25	
8			B7/B3 vs $\log_{10}\text{Chl}$	$y = -1.5224x + 2.314$	0.65	5.00	0.15	
9			B2/B6	$y = 51.278x - 39.882$	0.69	185.21	-29.56	
10			B2/B7 vs Chl-a	$y = 52.037x - 40.838$	0.65	185.10	-28.63	
11			B2/B6 vs $\log_{10}\text{Chl-a}$	$y = 0.4699x + 0.7636$	0.63	5.00	0.15	
12			B2/B7 vs $\log_{10}\text{Chl-a}$	$y = 0.481x + 0.7478$	0.61	4.35	0.16	
13			B6 vs Chl-a	$y = 1.322x^{-1.101}$	0.49	202.24	-12.73	
14			B7 vs Chl-a	$y = 131.65x^2 - 24.824x + 2.4481$	0.49	4.74	0.20	
15			(B5 + B6)/B4 vs $\log_{10}\text{Chl-a}$	$y = -0.4811x + 2.4517$	0.20	4.62	-0.03	
16			B5/B4 vs $\log_{10}\text{Chl-a}$	$y = 1.5412x^{-0.023}$	0.0002	3.25	-0.03	
17			B5-(B4 + B6)/2 vs $\log_{10}\text{Chl-a}$	$y = 2.3932x^{0.1022}$	0.13	2.92	-0.03	
18			(B1-B2)/(B1 + B2) vs $\log_{10}\text{Chl-a}$	$y = -2.4056x + 1.5881$	0.22	3.08	-0.05	

RMSE value has been converted into unit of $\mu\text{g/L}$ with the variables using \log_{10} Chl-a. The italic line in the table indicates the best performance of the linear model.

Water quality parameters-based machine learning for Chl-a retrieval

Input feature selection

In this paper, a feature selection was conducted using the correlation analysis to reduce the inputs from nine (9) to five (5) water quality parameters. Despite a good option for this

selection through the OOB (Out Of Bag) score of random forest model, as suggested by Li et al. (2018), the correlation analysis between the \log_{10} Chl-a and water parameters was used to create homogeneous inputs for all the compared models (Table 2).

The data for TAR presented a high correlation between total nitrogen and Chl-a concentration with a coefficient of 0.86. The total phosphorous and total suspended solid

Fig. 6 Scatter plots comparing satellite retrieved and observed Chl-a using 22 training data observations

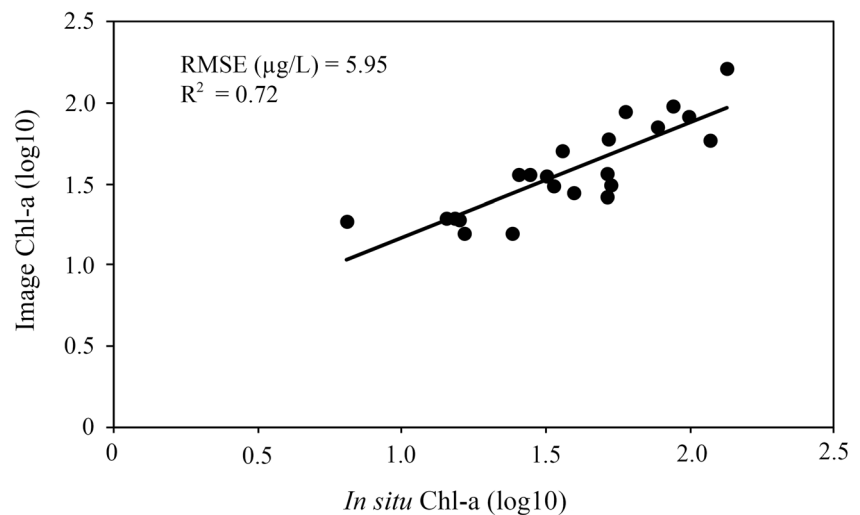
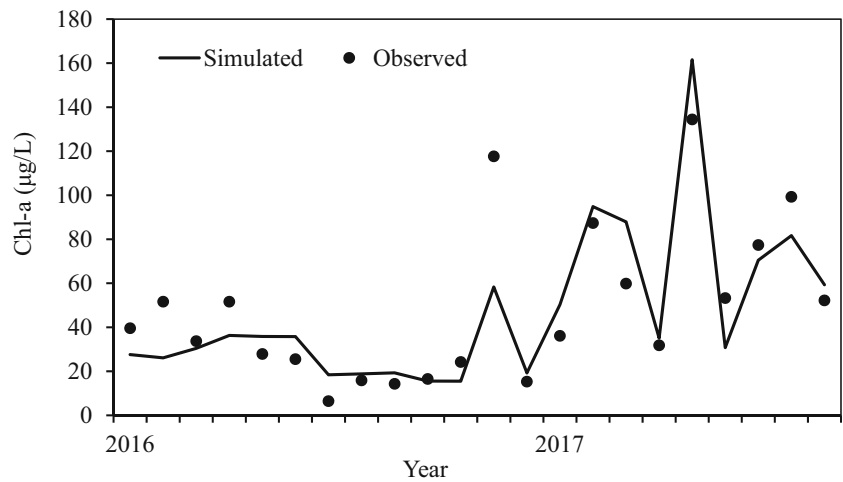


Fig. 7 Observed and simulated Chlorophyll-a concentration for the training dataset of the prediction model



establish a lower relationship with Chl-a (0.43 and 0.57, respectively) (Table 2). On the other hand, pH and temperature help to explain the variation of Chl-a. Nonetheless, this contribution is not significant with a low coefficient of Pearson correlation. As described in the methodology section, the water quality parameters with a Pearson correlation coefficient > 0.3 were selected. Hence, TP, TSS, TN, pH, and temperature are considered the input for the models of Chl-a prediction.

Model performance and comparison

The tuned parameters and selected kernel for machine learning models (Table 3) were preserved during the cross-validation running. Five water quality parameters, involving TP, TN, TSS, pH, and surface water temperature, were selected as the input to train and validate the linear and machine learning models. The best model for Chl-a prediction was a model with (a) the highest values of R^2 , EVS and the lowest values of RMSE, MAE (Table 4), and (b) the closest permutation of training and test scores in the learning curve (Fig. 9).

In addition, the model was perceived better, as it presented a smaller standard deviation of the score in the learning curve.

As presented in Fig. 9, the learning curves indicated a similar performance of the GPR, MLR, and MLP models, although the slight differences were observed. For the mentioned indicators, GPR was determined as the best model for Chl-a prediction from water quality parameters. This model was able to explain more than 79% of Chl-a variation at the study site with the lowest RMSE (3.06 µg/L) (significantly lower than the mean of Chl-a value of 26 µg/L) (Table 4). The stability and outperforming of the GPR model were also confirmed by a small standard deviation and closed permutation of the training and cross-validation score. The MLR model also demonstrated a good performance with a high R^2 and low RMSE (0.79 and 3.09 µg/L, respectively). However, the test score had a wider standard deviation compared to the GPR model (Fig. 9). The SVR and MLP models were both good in the permutation of the test score. MLP was stable during the learning process (Fig. 9); however, it was less accurate than the GPR and MLR models. Conversely, RFR was the only model that the training and test scores were not permuted

Fig. 8 Observed and simulated Chlorophyll-a concentration for the validation dataset

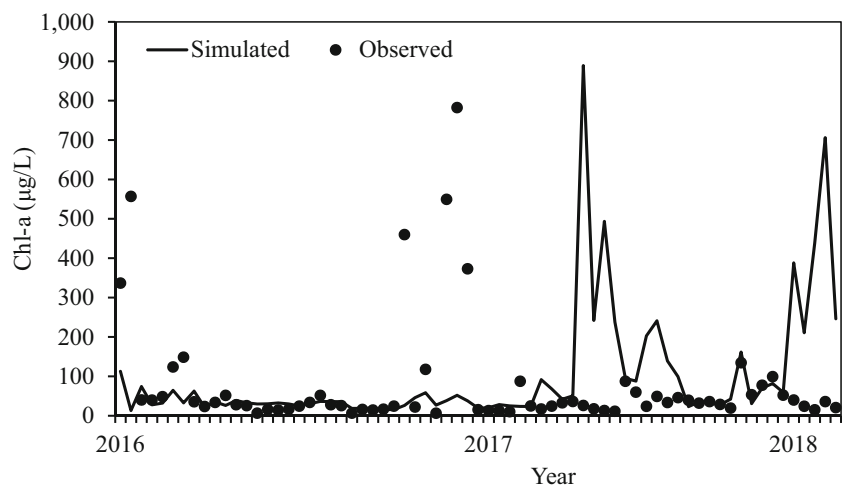


Table 2 Pearson correlation coefficients of transformed water quality parameters and Chl-a concentration. Boldface type indicated high correlation of Chl-a with other factors

	TSS	TN	pH	Temp	TP	DO	Trans	NO ₃ ⁻	PO ₄ ³⁻	Chl-a
TSS	1.00	0.51	0.13	0.41	0.31	-0.22	-0.46	0.07	0.33	<i>0.57</i>
TN		1.00	0.22	0.32	0.45	-0.24	-0.27	0.43	0.23	<i>0.86</i>
pH			1.00	0.29	0.11	0.35	0.36	-0.02	-0.05	<i>0.34</i>
Temp				1.00	0.29	-0.19	-0.29	0.04	0.23	<i>0.38</i>
TP					1.00	-0.20	-0.25	0.19	0.32	<i>0.43</i>
DO						1.00	0.60	-0.26	-0.01	-0.18
Trans							1.00	-0.26	-0.22	-0.22
NO ₃ ⁻								1.00	-0.02	0.29
PO ₄ ³⁻									1.00	0.29
Chl-a										1.00

during the learning, although the coefficient of determination reached 0.72 (Fig. 9).

According to Table 4, our results presented an accepted accuracy for Chl-a prediction with all the compared models. The linear and machine learning models were rational to explain a wide range of Chl-a variation (roughly 6–4600 µg/L) in TAR with the highest coefficient of R^2 (0.79) and the lowest RMSE (3.06 µg/L) of the GPR model. This RMSE was significantly lower than the mean value of Chl-a (26 µg/L), which determined the outperforming of the model for Chl-a prediction. Nonetheless, all the involved models shared a slightly large standard deviation of the validation scores due to a wide range of Chl-a concentrations in TAR. The comparison of the model's performance also answered the question in the introduction of this research, in that the multivariate linear model still works well in the case of the existing cause-effect relationship between the input (water quality) and output (Chl-a concentration) parameters. The previous analyses illustrated

that the variation of Chl-a concentration can be interpreted by the key factors, such as TSS and TN in TAR.

Spatiotemporal variation in Chl-a concentration from 2015 to 2019

Totally, 23 Sentinel-2A/B scenes that were acquired under the lowest cloud coverage conditions from the late 2015 to the early 2019 were applied to the linear model developed above to clarify the spatiotemporal variation in Chl-a concentration. These images were classified into rainy (May–October) and dry seasons (November–April). To study the spatial distribution in different seasons, retrieving Chl-a from the pixels of the Sentinel-2A/B was taken into account. Then, the relationship between Chl-a concentration and its frequency was obtained in the rainy and dry seasons (Fig. 10). In both seasons, Chl-a values significantly varied from 6.46 to 4626.02 µg/L. However, the difference in the frequency of Chl-a concentration between the two seasons was only exactly in the range from 8 to 400 µg/L. Chl-a ranging from 20 to 60 µg/L was the most frequent value in the rainy season with the highest frequency of 39%. In contrast, the most frequent value of Chl-a in the dry season was around 30 µg/L with a frequency of 33% (Fig. 10).

The distribution maps determining the specific regions of high Chl-a concentration in TAR are shown in Figs. 11 and 12,

Table 3 The best parameters and kernel for machine learning models

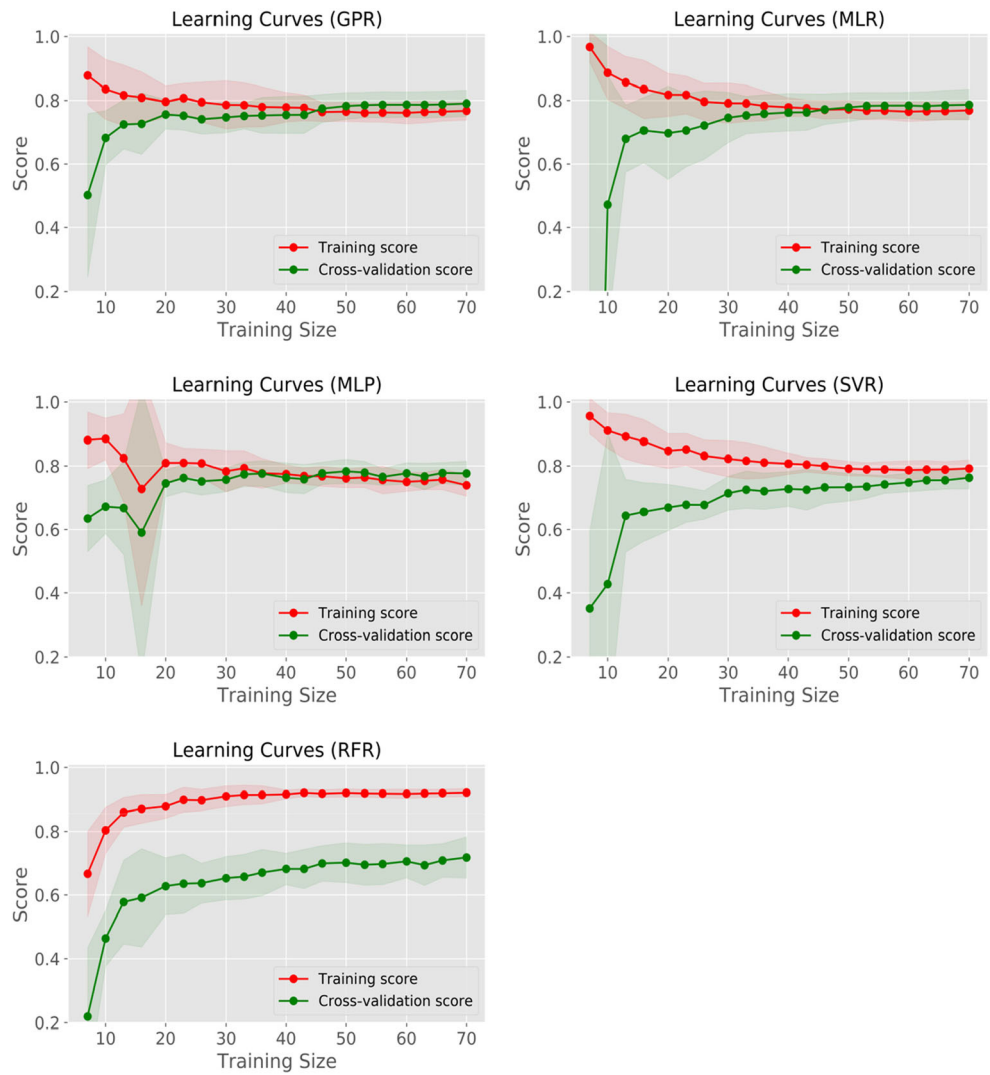
	Model			
Parameter	RFR		SVR	
Max depth	30	C	1000	
Bootstrap	True	Gamma	0.01	
Number of tree	50	Kernel	RBF	
Max feature	Auto			
Min sample leaf	2			
Min sample split	3			
MLP	GPR			
Hidden layer size	50	Alpha	0.0001	
Solver	Adam	Iteration for optimizer	10	
Activation	Relu	Kernel	RBF and WhiteKernel	
Max iteration	500			

Table 4 Validation scores for \log_{10} Chl-a prediction from water quality parameters

	GPR	RFR	SVR	MLP	MLR
R^2	0.79	0.72	0.76	0.77	0.79
RMSE	3.06	3.64	3.28	3.19	3.09
MAE	2.46	2.89	2.58	2.57	2.43
EVS	0.79	0.72	0.77	0.78	0.79

RMSE and MAE values were converted into unit of µg/L

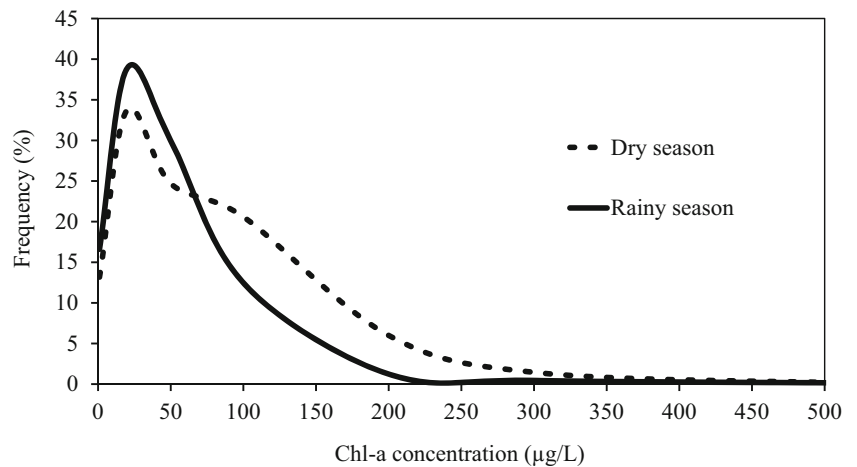
Fig. 9 Learning curves of machine learning and multivariate linear models. The light red and green areas indicate the standard deviation of the learning curves. The more narrow area determines a better performance of the model. Score is coefficient of determination of the learning process. The red and green lines are the mean value of the score



corresponding to the rainy and dry seasons. Due to high cloud coverage, there were not enough satellite images for a continuous, long-term observation. It is, therefore, a challenge to

generalize the overall variation of Chl-a in TAR. In the maps of 10 scenes in the rainy season and 13 scenes in the dry season observed from 2015 to 2019, the patterns of Chl-a

Fig. 10 Frequency distribution of Chl-a in the Tri An Reservoir by dry and rainy season



distribution were not homogeneous but mixed by high and low areas of Chl-a, due to the influences of meteorological and hydrological parameters. In both seasons, most areas of TAR suffered from HCBs with a high Chl-a concentration, ranging from 20 to below 5000 $\mu\text{g/L}$. The blooms were directly observed during the field campaigns on April 23, 2016 and September 20, 2016, which are consistent with the patterns in Figs. 11 and 12. Algae tends to bloom more strongly in the south and southwest areas, which are the downstream of TAR. It is worth noting that there are two major rivers located in the nearby station, TA5 (Fig. 1), discharging water into the reservoir especially after heavy rains, which implies that HCBs usually occur in the downstream compared to the north of the study area. In the rainy season, Chl-a concentration tends to be higher from the early days of June to the end of the season (i.e., October), and these values are specific to the dry season. Conversely, higher values were observed in the early days of the dry season (November to early February), and Chl-a values began declining in the following months (from March to June). It is clear that the most serious blooms first occurred in the southern or central part of the reservoir and gradually moved toward the southwest and northwest areas (pair images of September 5, 2017 vs. September 20, 2017). The higher values were in the southern and southwestern parts, while the lower values were in the southeast and northeast areas of the reservoir. During the years from 2015 to 2019, the maximum bloom area occupied more than 50% of the total area.

With a R^2 of 0.72, the linear model explains with great certainty the general variation of Chl-a in TAR in both the

dry and rainy seasons. Nonetheless, a small area was recorded with a very high Chl-a concentration (over 5000 $\mu\text{g/L}$) in the narrow corner of the reservoir. These values may arise from very high turbidity and/or strong effects of bottom reflectance in shallow waters, leading to the anomalous values of the surface water reflectance in bands 3 and 6 of Sentinel-2 imagery.

Discussion

Similarity to other studies

The optimal position of the band determined for Chl-a estimation algorithm in this study is consistent with the finding of other research papers. Spectral bands beyond 650 nm are appropriate for the development of Chl-a retrieval models for inland waters, particularly in case 2 waters where Chl-a concentration is above 10 $\mu\text{g/L}$ (Richardson and LeDrew 2006; Le et al. 2009; Mishra et al. 2017). Especially, in case of large variation of Chl-a concentrations, Zimba and Gitelson (2006) proved that the wavelengths of 650 nm, 710 nm, and 740 nm were the optimal choice. The specific wavelength varies with water constituents and their optical properties (Le et al. 2009). These findings demonstrate that the red-near-infrared wavelengths are the appropriate regions for the development of Chl-a retrieval algorithms in TAR.

To our knowledge, this is the third study using remote sensing for Chl-a concentration in the lakes/reservoirs in Vietnam and the first in the southern area. Our model (R^2 of

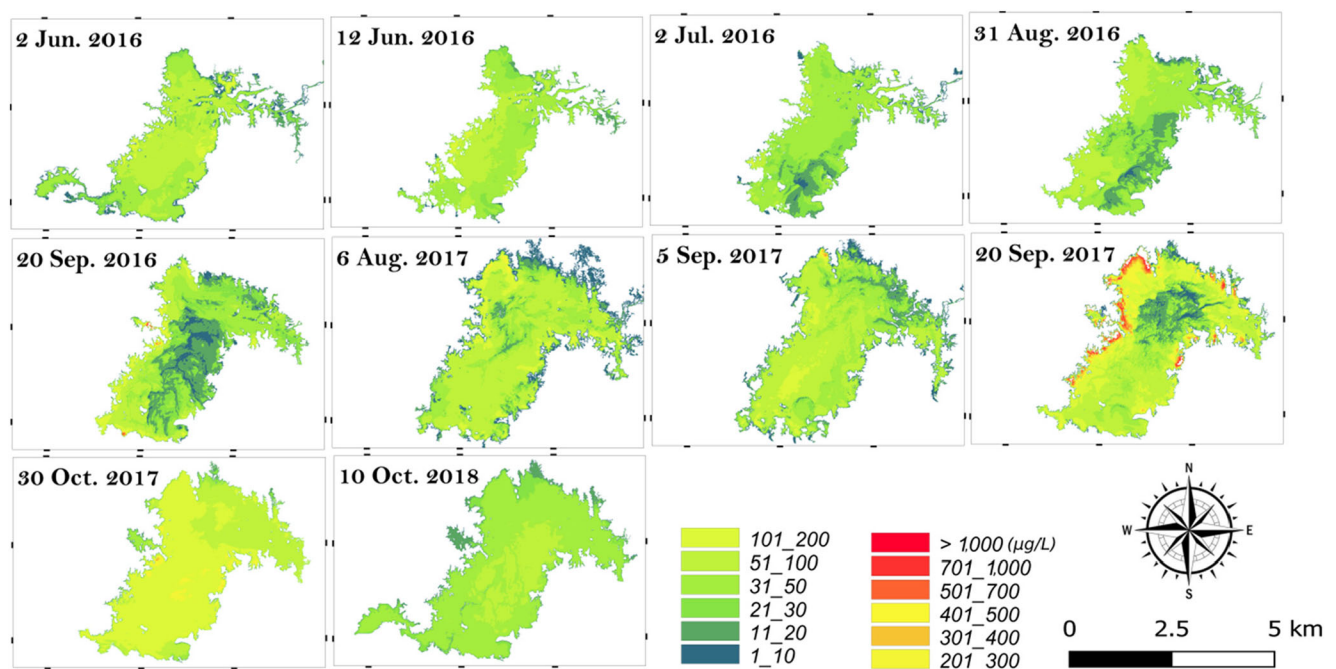


Fig. 11 Spatiotemporal distribution of HCBs in the Tri An Reservoir in rainy season

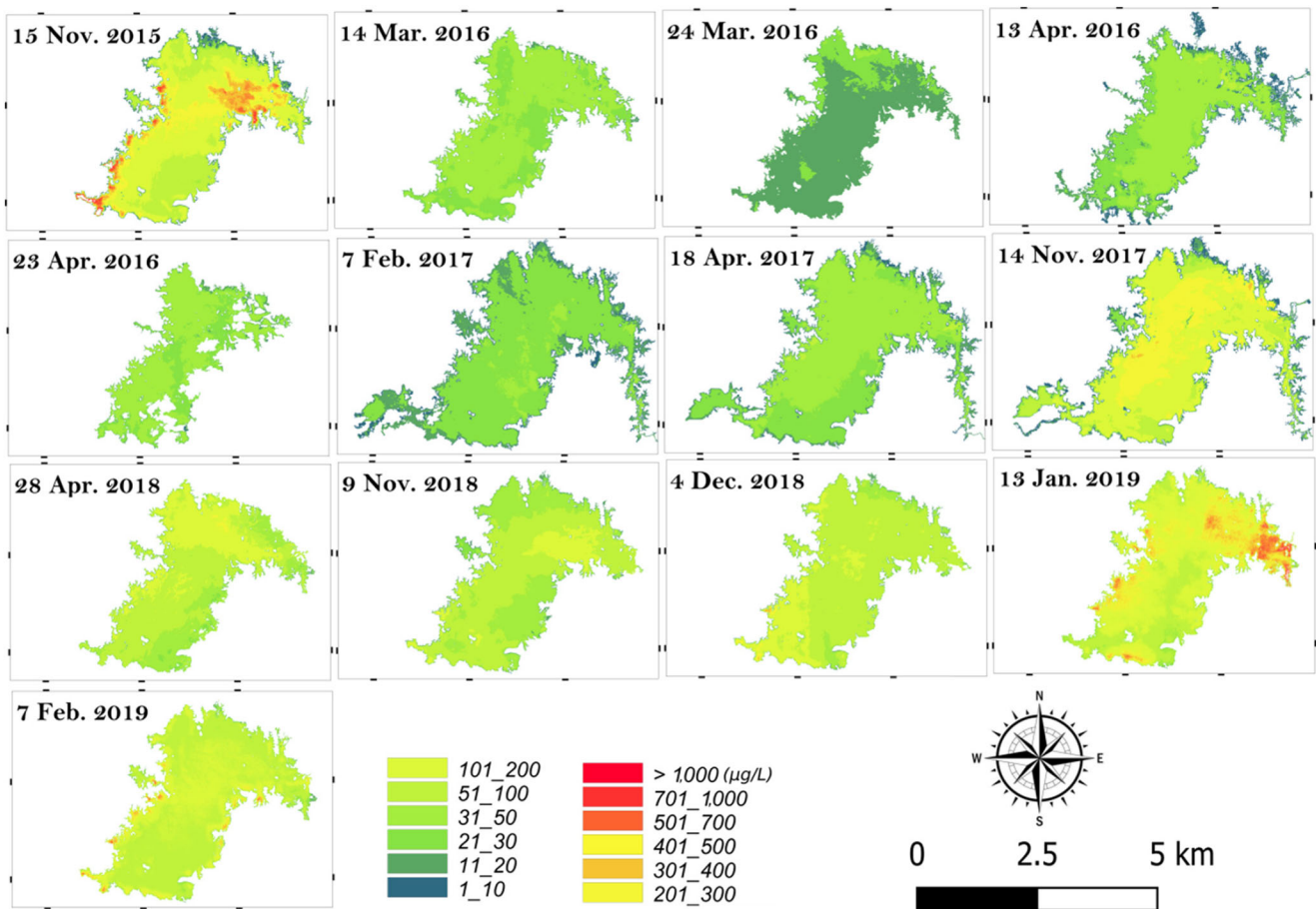


Fig. 12 Spatiotemporal distribution of HCBs in the Tri An Reservoir in dry season

0.72) outperformed the model applied in Ba Be lake with an R^2 of 0.68. It is worth noting that TAR exhibited a large variation in Chl-a (mean value from 15.48 to 1310 µg/L), compared to a very small range from 1.58 to 6 µg/L in Ba Be water case (Ha et al. 2017b). The other study used Landsat 8/OLI two bands ratio algorithm for Chl-a concentration mapping in West Lake in Hanoi with a range from 42 to 258 µg/L (Ha et al. 2017a). However, this study used the exponential function for Chl-a retrieval with an archived R^2 from 0.64 to 0.82, which may result in uncertainties compared to linear regression method applied in our paper (Grendaitè et al. 2018; Ha et al. 2013; Lins et al. 2017; Pham et al. 2019; Quang et al. 2017).

Machine learning with other case studies

The results of this study also support the practicality of using machine learning models for the retrieval of bio-optical parameters. Compared to other research papers, additional results were found in the case study in TAR. For log-transformed data, a high correlation coefficient with Chl-a was detected for TN (0.86) and TSS (0.57), which contrasts

with various observed datasets (Li et al. 2018; Lou et al. 2016). In addition, the GPR model outperformed other competing models and provided a novel solution to the task of Chl-a prediction in the freshwater environment. Compared to the application of random forest (Li et al. 2018), extreme learning (Lou et al. 2016), ANN and support vector machine (Park et al. 2015; Xie et al. 2012), and M5P model tree and despite a lower value of R^2 (Yi et al. 2018b), GPR still shows very good RMSE and MAE values in the case study of TAR. A similar outstanding performance of the GPR model was also identified for biophysical parameter retrieval, particularly for the oceanic Chl-a estimation (Blix and Eltoft 2018a; Blix and Eltoft 2018b; Verrelst et al. 2012). More interestingly, the multivariable linear model proved itself as a profitable predictor for Chl-a variation when a cause-effect relationship exists between the inputs (water quality parameters) and the output (Chl-a concentration). This performance is worthy of attention, mainly because of the simplicity and low consumption of the computer power of the model, compared to other machine learning approaches.

In comparison with the linear model for Chl-a retrieval from the remotely sensing data, the machine learning

approach using water quality may not be an optimal method for monitoring the HCBs in TAR in spite of a higher R^2 and lower RMSE compared to the band ratio regression models. This issue emanates from the limitations of the machine learning model in practice, which requires substantial amounts of input data, has complex inherent algorithms with various parameters, and poses the challenge of precisely interpreting the results from the applied models. In addition, the unavailability of water quality data may further complicate model validation and prediction for the early warnings of HCBs. As a result, our proposed linear model may be construed as an optimal selection for the further monitoring of HCBs in inland waters of Vietnam.

Mechanisms of HCBs in TAR

The blooms of harmful cyanobacteria in inland waters are influenced by multiple factors, including, but not limited to, light, temperature, turbidity, precipitation, wind speed, water residence time, and nutrient composition (Mu et al. 2019; Paerl 2017). In Lake Vancouver (Canada), phosphate was found to be the key factor regulating HCBs and toxins

concentrations (Lee et al. 2015), whereas high nutrients, low water clarity, and warmer surface temperatures were identified as the three most influential environmental factors correlated with cyanobacterial composition in the US lakes and reservoirs (Beaver et al. 2018). High water temperatures have been known to trigger the development of cyanobacterial bloom in temperate zones (Imai et al. 2008; Kosten et al. 2012), while the occurrence of HCBs in tropical areas has been reported to be regulated by many factors (Bui et al. 2017; Pham et al., 2017).

In TAR, Chl-a concentration was strongly correlated with TN, TSS, and, to a less extent, TP; however, it was negatively correlated with transparency. It is difficult to draw strong conclusions on blooming. However, an overall trend is evident since the rainy months and the early dry months (from May to November) tend to correspond to the highest bloom frequencies. The discharge of nitrogen-rich water, as dominated by the intensification of agriculture and land runoff in the rainy season from the TAR catchment area, is very likely to create a favorable condition for the development of HCBs. It is suggested that cyanobacteria blooms in TAR are attributed to the conditions of high nutrient concentration and low transparency in the rainy season.

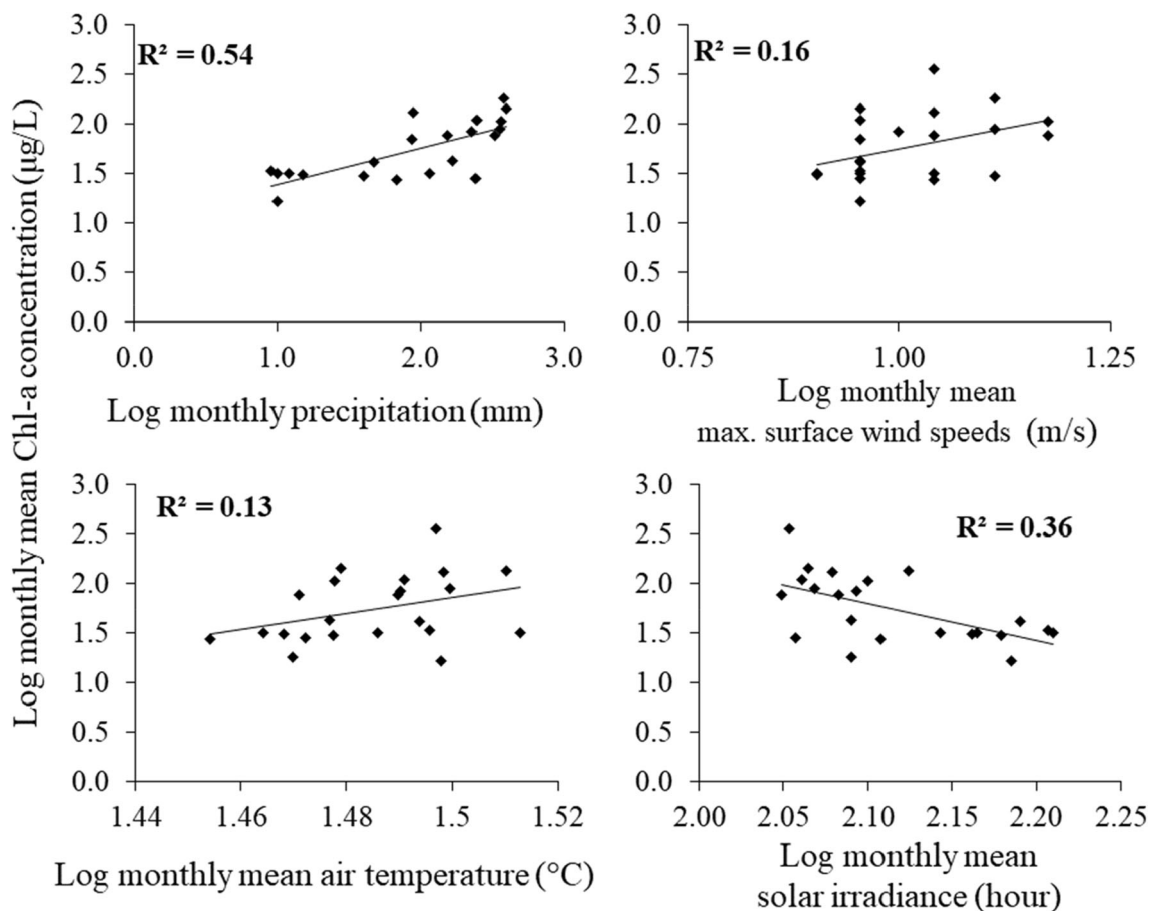


Fig. 13 Correlation between Chl-a concentration and climatic parameters

The relationships between the monthly mean Chl-a concentration and climatic parameters measured at the Tri An station are shown in Fig. 13. In general, heavy rain will lead to a temporary increase of nutrients, which is beneficial for algae blooms (Mu et al. 2019). The results show that monthly precipitation exhibited statistically significant positive correlation with average monthly Chl-a concentration ($R^2 = 0.54$, $p < 0.01$). This suggests that monthly precipitation has a strong effect on blooms in the short term in the TAR. This is consistent with observations of Mu et al. (2019) in the Dianchi Lake, China. Followed by the rainfall, the monthly mean solar irradiance is also considered as a contributor to HCBs in TAR ($R^2 = 0.36$, $p < 0.01$). In contrast, wind speed and air temperature have less impacts on HCBs with R^2 of 0.16, and 0.13, respectively (Fig. 13).

Conclusions

To our best knowledge, this study is the first attempt to calculate Chl-a concentration in TAR using the data extracted from both Sentinel-2A/B data and water quality collected from field campaigns. We constructed the models using both approaches, involving traditional band ratio regression with the simulated Sentinel-2A/B data and the state-of-the-art machine learning with water quality data to estimate and map Chl-a in a eutrophic tropical reservoir. The combination of NIR and green band ratio with an R^2 of 0.72 $\mu\text{g/L}$ and an RMSE of 5.95 $\mu\text{g/L}$ was identified as the optimal model for quantifying Chl-a in such tropical inland waters. By comparison, both methods have satisfactory performance for Chl-a retrieval ($R^2 > 0.7$). Specifically, GPR yields better results with the highest R^2 of 0.79 and the lowest RMSE of 3.06 $\mu\text{g/L}$, compared to the other machine learning models.

The results illustrate the benefit of using machine learning models when it is hard to obtain cloud free satellite images, particularly in tropical regions. Both band ratio regression and machine learning model can support each other and be applied to water quality and environment management studies. Future studies are advised to repeat our sampling strategy to validate the model and link remote sensing reflectance data to machine learning algorithms.

Next, studies investigating the factors influencing the occurrence of HCBs using monthly water quality data had been carried out to elucidate the spatiotemporal variation. High TN and TSS in the rainy season were considered the driving factors of HCBs with Chl-a concentration greater than 20 $\mu\text{g/L}$ in most areas of TAR. Regarding climatic parameters, Chl-a concentration was high due to a large amount of precipitation entering the study area, followed by solar irradiance while air temperature and wind speed made small contribution to HCBs.

The early prediction of HCBs is necessary to support a healthy practice of water usage in the community. The promising results of this study offer various approaches to the task of monitoring HCBs using satellite-based modeling or machine learning with water quality data. However, longitudinal cyanobacterial data collection is recommended for an accurate prediction of HCBs in TAR.

Acknowledgments We thank the editor and anonymous reviewers for their constructive comments, which helped us to improve the manuscript. We also thank to Ms. Nguyen Hong Van who provided us valuable climatic data.

Funding information This study was funded by Vietnam Academy of Science and Technology (VAST) under grant number “KHCBSS.02/19-21”.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

- Adriana GC, Richard K (2017) Sentinel Data Access 2017 Annual Report vol 1. ESA
- APHA (2005) Standard methods for the examination of water and wastewater, 21st edn. American Public Health Association/American Water Works Association/Water Environment Federation, Washington DC
- Beaver JR, Tausz CE, Scotese KC, Pollard AI, Mitchell RM (2018) Environmental factors influencing the quantitative distribution of microcystin and common potentially toxigenic cyanobacteria in U.S. lakes and reservoirs. *Harmful Algae* 78:118–128. <https://doi.org/10.1016/j.hal.2018.08.004>
- Blix K, Camps-Valls G, Janssen R (2017) Gaussian process sensitivity analysis for oceanic chlorophyll estimation. *IEEE J Sel Top Appl Earth Obs Remote Sens* 10:1265–1277. <https://doi.org/10.1109/JSTARS.2016.2641583>
- Blix K, Eltoft T (2018a) Evaluation of feature ranking and regression methods for oceanic chlorophyll-a estimation. *IEEE J Sel Top Appl Earth Obs Remote Sens* 11:1403–1418. <https://doi.org/10.1109/JSTARS.2018.2810704>
- Blix K, Eltoft T (2018b) Machine learning automatic model selection algorithm for oceanic chlorophyll-a content retrieval. *Remote Sens* 10:775. <https://doi.org/10.3390/rs10050775>
- Bresciani M, Cazzaniga I, Austoni M, Sforzi T, Buzzi F, Morabito G, Giardino C (2018) Mapping phytoplankton blooms in deep subalpine lakes from Sentinel-2A and Landsat-8. *Hydrobiologia* 824: 197–214. <https://doi.org/10.1007/s10750-017-3462-2>
- Bui M-H, Pham T-L, Dao T-S (2017) Prediction of cyanobacterial blooms in the Dau Tieng reservoir using an artificial neural network. *Mar Freshw Res* 68:2070. <https://doi.org/10.1071/MF16327>
- Chen J, Zhu W, Tian YQ, Yu Q, Zheng Y, Huang L (2017) Remote estimation of colored dissolved organic matter and chlorophyll-a in Lake Huron using Sentinel-2 measurements. *J Appl Remote Sens* 11:1. <https://doi.org/10.1117/1.JRS.11.036007>
- Chorus I, Bartram J (1999) Toxic cyanobacteria in water: a guide to their public health consequences, monitoring and management, published on behalf of WHO. Sporn Press, London, 416 pp

- Dao T-S, Nimptsch J, Wiegand C (2016) Dynamics of cyanobacteria and cyanobacterial toxins and their correlation with environmental parameters in Tri An Reservoir, Vietnam. *J Water Health* 14:669–712
- Dippner JW, Nguyen-Ngoc L, Doan-Nhu H, Subramaniam A (2011) A model for the prediction of harmful algae blooms in the Vietnamese upwelling area. *Harmful Algae* 10:606–611. <https://doi.org/10.1016/j.hal.2011.04.012>
- Gascon F, Bouzinac C, Thépaut O, Jung M, Francesconi B, Louis J, Lonjou V, Lafrance B, Massera S, Gaudel-Vacaresse A, Languille F, Alhammoud B, Viallefont F, Pflug B, Bieniarz J, Clerc S, Pessiot L, Trémas T, Cadau E, De Bonis R, Isola C, Martimort P, Fernandez V, Copernicus (2017) Sentinel-2A calibration and products validation status. *Remote Sens* 9:584. doi:<https://doi.org/10.3390/rs9060584>
- Grendaitė D, Stonevičius E, Karosienė J, Savadova K, Kasperovičienė J (2018) Chlorophyll-a concentration retrieval in eutrophic lakes in Lithuania from Sentinel-2 data. *Geologija Geografija* 4:15–28. <https://doi.org/10.6001/geol-geogr.v4i1.3720>
- Ha NTT, Koike K, Nhuan MT (2013) Improved accuracy of chlorophyll-a concentration estimates from MODIS imagery using a two-band ratio algorithm and geostatistics: as applied to the monitoring of eutrophication processes over Tien Yen Bay (Northern Vietnam). *Remote Sens* 6:421–442. <https://doi.org/10.3390/rs6010421>
- Ha NTT, Koike K, Nhuan MT, Canh BD, Thao NTP, Parsons M (2017a) Landsat 8/OLI two bands ratio algorithm for chlorophyll-a concentration mapping in hypertrophic waters: An application to west lake in Hanoi (Vietnam). *IEEE J Sel Top Appl Earth Obs Remote Sens* 10:4919–4929. <https://doi.org/10.1109/JSTARS.2017.2739184>
- Ha NTT, Thao NTP, Koike K, Nhuan MT (2017b) Selecting the best band ratio to estimate chlorophyll-a concentration in a tropical freshwater lake using sentinel 2A images from a case study of Lake Ba Be (northern Vietnam). *ISPRS Int J Geo Inf* 6:290. <https://doi.org/10.3390/ijgi6090290>
- Imai H, Chang KH, Kusaba M, Si N (2008) Temperature-dependent dominance of *Microcystis* (Cyanophyceae) species: *M. aeruginosa* and *M. wessenbergii*. *J Plankton Res* 31:171–178. <https://doi.org/10.1093/plankt/fbn110>
- Johnson R, Strutton PG, Wright SW, McMinn A, Meiners KM (2013) Three improved satellite chlorophyll algorithms for the Southern Ocean. *J Geophys Res-Oceans* 118(7):3694–3703. <https://doi.org/10.1002/jgrc.20270>
- Karamzadeh S, Abdullah SM, Halimi M, Shayan J, Rajabi M (2014) Advantage and drawback of support vector machine functionality. In: 2014 international conference on computer, communications, and control technology (I4CT), 2014/09/2014. IEEE, Langkawi, Malaysia, pp 63–65. <https://doi.org/10.1109/I4CT.2014.6914146>
- Keller S, Maier PM, Riese FM, Norra S, Holbach A, Börsig N, Wilhelms A, Moldaenke C, Zaake A, Hinz S (2018) Hyperspectral data and machine learning for estimating CDOM, chlorophyll a, diatoms, green algae and turbidity. *Int J Environ Res Public Health* 15:1881. <https://doi.org/10.3390/ijerph15091881>
- Kosten S, Huszar VLM, Bécares E, Costa LS, van Donk E, Hansson L-A, Jeppesen E, Kruk C, Lacerot G, Mazzeo N, De Meester L, Moss B, Lüring M, Nöges T, Romo S, Scheffer M (2012) Warmer climates boost cyanobacterial dominance in shallow lakes. *Glob Chang Biol* 18(1):118–126. <https://doi.org/10.1111/j.1365-2486.2011.02488.x>
- Kwon Y, Baek S, Lim Y, Pyo J, Ligaray M, Park Y, Cho K (2018) Monitoring coastal chlorophyll-a concentrations in coastal areas using machine learning models. *Water* 10:1020. <https://doi.org/10.3390/w10081020>
- 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. <https://doi.org/10.1016/j.rse.2009.02.005>
- Lee TA, Rollwagen-Bollens G, Bollens SM, Faber-Hammond JJ (2015) Environmental influence on cyanobacteria abundance and microcystin toxin production in a shallow temperate lake. *Ecotox Environ Safe* 114:318–325. <https://doi.org/10.1016/j.ecoenv.2014.05.004>
- Li X, Sha J, Wang Z-L (2018) Application of feature selection and regression models for chlorophyll-a prediction in a shallow lake. *Environ Sci Pollut Res* 25:19488–19498. <https://doi.org/10.1007/s11356-018-2147-3>
- Lins R, Martinez J-M, Motta Marques D, Cirilo J, Fragoso C (2017) Assessment of chlorophyll-a remote sensing algorithms in a productive tropical estuarine-lagoon system. *Remote Sens* 9:516. <https://doi.org/10.3390/rs9060516>
- Liu C, Tang D (2012) Spatial and temporal variations in algal blooms in the coastal waters of the western South China Sea. *J Hydro-Environ Res* 6:239–247. <https://doi.org/10.1016/j.jher.2012.02.002>
- Lou I, Xie Z, Ung WK, Mok KM (2016) Freshwater algal bloom prediction by extreme learning machine in Macau storage. *Neural Comput & Applic* 27:19–26. <https://doi.org/10.1007/s00521-013-1538-0>
- Maeda EE, Lisboa F, Kaikkonen L, Kallio K, Koponen S, Brotas V, Kuikka S (2019) Temporal patterns of phytoplankton phenology across high latitude lakes unveiled by long-term time series of satellite data. *Remote Sens Environ* 221:609–620. <https://doi.org/10.1016/j.rse.2018.12.006>
- Martins V, Barbosa C, de Carvalho L, Jorge D, Lobo F, Novo E (2017) Assessment of atmospheric correction methods for sentinel-2 MSI images applied to Amazon floodplain lakes. *Remote Sens* 9:322. <https://doi.org/10.3390/rs9040322>
- Mishra DR, Ogashawara I, Gitelson AA (2017) Remote sensing of inland waters. Bio-optical modeling and remote sensing of inland waters. Elsevier, In, pp 1–24
- Mishra S, Mishra DR (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. <https://doi.org/10.1016/j.rse.2011.10.016>
- Mu M, Wu C, Li Y, Lyu H, Fang S, Yan X, Liu G, Zheng Z, Du C, Bi S (2019) Long-term observation of cyanobacteria blooms using multi-source satellite images: a case study on a cloudy and rainy lake. *Environ Sci Pollut Res* 26:11012–11028. <https://doi.org/10.1007/s11356-019-04522-6>
- Mueller-Wilm U, Devignot O, Pessiot L (2018) Sen2Cor configuration and user manual vol 2. ESA
- Paerl HW (2017) Controlling cyanobacterial harmful blooms in freshwater ecosystems. *Microb Biotechnol* 10:1106–1110. <https://doi.org/10.1111/1751-7915.12725>
- Paerl HW, Huisman J (2008) Climate: blooms like it hot. *Science* 320:57–58. <https://doi.org/10.1126/science.1155398>
- Paerl HW, Paul VJ (2012) Climate change: links to global expansion of harmful cyanobacteria. *Water Res* 46:1349–1363. <https://doi.org/10.1016/j.watres.2011.08.002>
- Park Y, Cho KH, Park J, Cha SM, Kim JH (2015) Development of early-warning protocol for predicting chlorophyll-a concentration using machine learning models in freshwater and estuarine reservoirs, Korea. *Sci Total Environ* 502:31–41. <https://doi.org/10.1016/j.scitotenv.2014.09.005>
- Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E (2011) Scikit-learn: machine learning in Python. *J Mach Learn Res* 12:2825–2830
- Pham T-L, Dao T-S, Tran N-D, Nimptsch J, Wiegand C, Motoo U (2017) Influence of environmental factors on cyanobacterial biomass and microcystin concentration in the Dau Tieng reservoir, a tropical eutrophic water body in Vietnam. *Ann Limnol Int J Limnol* 53:89–100. <https://doi.org/10.1051/limn/2016038>
- Pham T-L, Utsumi M (2018) An overview of the accumulation of microcystins in aquatic ecosystems. *J Environ Manag* 213:520–529. <https://doi.org/10.1016/j.jenvman.2018.01.077>

- Pham T, Yokoya N, Bui D, Yoshino K, Friess D (2019) Remote sensing approaches for monitoring mangrove species, structure, and biomass: opportunities and challenges. *Remote Sens* 11:230. <https://doi.org/10.3390/rs11030230>
- Quang N, Sasaki J, Higa H, Huan N (2017) Spatiotemporal variation of turbidity based on landsat 8 OLI in Cam Ranh Bay and Thuy Trieu lagoon, Vietnam. *Water* 9:570. <https://doi.org/10.3390/w9080570>
- Richardson LL, LeDrew EF (2006) Remote sensing of aquatic coastal ecosystem processes vol 9. Remote Sensing and Digital Image Processing. Springer Netherlands, Dordrecht
- Ritchie JC, Zimba PV, Everitt JH (2003) Remote sensing techniques to assess water quality. *Photogramm Eng Remote Sens* 69:695–704. <https://doi.org/10.14358/PERS.69.6.695>
- Schaeffer BA, Bailey SW, Conmy RN, Galvin M, Ignatius AR, Johnston JM, Keith DJ, Lunetta RS, Parmar R, Stumpf RP, Urquhart EA, Werdell PJ, Wolfe K (2018) Mobile device application for monitoring cyanobacteria harmful algal blooms using Sentinel-3 satellite ocean and land colour instruments. *Environ Model Softw* 109:93–103. <https://doi.org/10.1016/j.envsoft.2018.08.015>
- Sola I, García-Martín A, Sandoñis-Pozo L, Álvarez-Mozos J, Pérez-Cabello F, González-Audicana M, Montorio Llovería R (2018) Assessment of atmospheric correction methods for Sentinel-2 images in Mediterranean landscapes. *Int J Appl Earth Obs Geoinf* 73:63–76. <https://doi.org/10.1016/j.jag.2018.05.020>
- Tan W, Liu P, Liu Y, Yang S, Feng S (2017) A 30-year assessment of phytoplankton blooms in Erhai Lake using Landsat imagery: 1987 to 2016. *Remote Sens* 9(12):1265. <https://doi.org/10.3390/rs9121265>
- Tang DL, Kawamura H, Doan-Nhu H, Takahashi W (2004) Remote sensing oceanography of a harmful algal bloom off the coast of southeastern Vietnam. *J Geophys Res Oceans* 109(C3). <https://doi.org/10.1029/2003JC002045>
- Toming K, Kutser T, Laas A, Sepp M, Paavel B, Nõges T (2016) First experiences in mapping lake water quality parameters with Sentinel-2 MSI imagery. *Remote Sens* 8:640. <https://doi.org/10.3390/rs8080640>
- Verrelst J, Muñoz J, Alonso L, Delegido J, Rivera JP, Camps-Valls G, Moreno J (2012) Machine learning regression algorithms for biophysical parameter retrieval: opportunities for Sentinel-2 and -3. *Remote Sens Environ* 118:127–139. <https://doi.org/10.1016/j.rse.2011.11.002>
- Visser PM, Verspagen JMH, Sandrini G, Stal LJ, Matthijs HCP, Davis TW, Paerl HW, Huisman J (2016) How rising CO₂ and global warming may stimulate harmful cyanobacterial blooms. *Harmful Algae* 54:145–159. <https://doi.org/10.1016/j.hal.2015.12.006>
- Wang X, Gong Z, Pu R (2018) Estimation of chlorophyll a content in inland turbidity waters using WorldView-2 imagery: a case study of the Guanting reservoir, Beijing, China. *Environ Monit Assess* 190:620. <https://doi.org/10.1007/s10661-018-6978-7>
- Xie Z, Lou I, Ung WK, Mok KM (2012) Freshwater algal bloom prediction by support vector machine in Macau storage reservoirs. *Math Probl Eng* 2012:1–12. <https://doi.org/10.1155/2012/397473>
- Yi H-S, Lee B, Park S, Kwak K-C, An K-G (2018a) Prediction of short-term algal bloom using the M5P model-tree and extreme learning machine. *Environ Eng Res* 24:404–411. <https://doi.org/10.4491/eer.2018.245>
- Yi H-S, Lee B, Park S, Kwak K-C, An K-G (2018b) Short-term algal bloom prediction in Juksan weir using M5P model-tree and extreme learning machine. *Environ Eng Res*. <https://doi.org/10.4491/eer.2018.245>
- Zhang Y, Zhang Y, Shi K, Zha Y, Zhou Y, Liu M (2016) A Landsat 8 OLI-based, semianalytical model for estimating the total suspended matter concentration in the slightly turbid Xin'anjiang reservoir (China). *IEEE J Sel Top Appl Earth Obs Remote Sens* 9:398–413. <https://doi.org/10.1109/JSTARS.2015.2509469>
- Zimba PV, Gitelson A (2006) Remote estimation of chlorophyll concentration in hyper-eutrophic aquatic systems: model tuning and accuracy optimization. *Aquaculture* 256:272–286. <https://doi.org/10.1016/j.aquaculture.2006.02.038>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.