PLOS ONE

RESEARCH ARTICLE

Predictive performance of regression models to estimate Chlorophyll-a concentration based on Landsat imagery

Miguel Ángel Matus-Hernández[1☯], Norma Yolanda Hernández-Saavedra[1☯], Raúl Octavio Martínez-Rincón[2☯*]

1 Centro de Investigaciones Biológicas del Noroeste, La Paz, Baja California Sur, México, **2** CONACYT—Centro de Investigaciones Biológicas del Noroeste, La Paz, Baja California Sur, México

☯ These authors contributed equally to this work.
* raul.martinez.rincon@gmail.com

OPEN ACCESS

Citation: Matus-Hernández MÁ, Hernández-Saavedra NY, Martínez-Rincón RO (2018) Predictive performance of regression models to estimate Chlorophyll-a concentration based on Landsat imagery. PLoS ONE 13(10): e0205682. https://doi.org/10.1371/journal.pone.0205682

Editor: Zhihua Zhang, Beijing Normal University, CHINA

Received: March 5, 2018

Accepted: September 29, 2018

Published: October 12, 2018

Data Availability Statement: All relevant data are within the paper.

Funding: This work was supported by CONACyT (PD-CPN-213849) who also provided a scholarship to the first author, MAMH. The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interest exist.

Abstract

Chlorophyll-a (Chl-a) concentration is a key parameter to describe water quality in marine and freshwater environments. Nowadays, several products with Chl-a have derived from satellite imagery, but they are not available or reliable sometimes for coastal and/or small water bodies. Thus, in the last decade several methods have been described to estimate Chl-a with high-resolution (30 m) satellite imagery, such as Landsat, but a standardized method to estimate Chl-a from Landsat imagery has not been accepted yet. Therefore, this study evaluated the predictive performance of regression models (Simple Linear Regression [SLR], Multiple Linear Regression [MLR] and Generalized Additive Models [GAMs]) to estimate Chl-a based on Landsat imagery, using *in situ* Chl-a data collected (synchronized with the overpass of Landsat 8 satellite) and spectral reflectance in the visible light portion (bands 1–4) and near infrared (band 5). These bands were selected because of Chl-a absorbance/reflectance properties in these wavelengths. According to goodness of fit, GAM outperformed SLR and MLR. However, the model validation showed that MLR performed better in predicting log-transformed Chl-a. Thus, MLR, constructed by using four spectral bands (1, 2, 3, and 5), was considered the best method to predict Chl-a. The coefficients of this model suggested that log-transformed Chl-a concentration had a positive linear relationship with bands 1 (coastal/aerosol), 3 (green), and 5 (NIR). On the other hand, band 2 (blue) suggested a negative relationship, which implied high coherence with Chl-a absorbance/reflectance properties measured in the laboratory, indicating that Landsat 8 images could be applied effectively to estimate Chl-a concentrations in coastal environments.

Introduction

Coastal environments are highly productive and complex marine ecosystems because they show the interaction of various natural and anthropogenic phenomena that provide an important source of nutrients for phytoplankton and aquatic organisms, as well as for various human activities. Nevertheless, during the last decades, studies have demonstrated that these

water bodies have been under significant stress due to anthropogenic alterations and climate variations that are increasingly frequent events, such as algal blooms [1,2]. In these environments, Chl-a has been considered as one of the most important parameters for measuring water quality, so it can be used as an indicator of ecosystem health [3,4].

The concentration of Chl-a varies spatially in coastal regions, and conventional methods for point-to-point studies are expensive, require time, and are usually spatially incomplete [5,6]. Therefore, several techniques have been proposed recently to use remote sensors as a viable option for monitoring environmental parameters at local spatial scales through images with high spatial resolution. Several algorithms have been developed to measure Chl-a based on the relationship that exists between the reflectance of different wavelengths from sensors specifically designed for monitoring Chl-a in marine environments, such as Coastal Zone Color Scanner (CZCS) with a spatial resolution of 825 m [7]; Sea-Viewing Wide Field-of-View Sensor (SeaWiFS) of 1130 m [8,9]; Medium Resolution Imaging Spectrometer (MERIS) of 300 m [10]; and Moderate Resolution Imaging Spectroradiometer (MODIS) of 250 m, 500 m and 1000 m [11,12]. Nonetheless, several difficulties have been reported when performing adequate monitoring of these environments, among which those of low resolution can only be applied effectively in homogeneous open sea areas but not for spatially complex coastal environments, such as bays or estuaries that require a higher spatial resolution for their study [1,5].

Landsat satellite series have provided a temporary record of multispectral images of the longest land surface in history since 1972, registry widely used for several governmental, public, and private applications [13]. The last satellite of this series is Landsat 8, which consists of two sensors, one called Operational Land Imager (OLI) and the other one Thermal Infrared Sensor (TIRS). Both of them obtain data jointly to provide land surface images, including coastal regions, polar ice, islands, and continental zones [14]. Although this satellite was designed for the study of terrestrial processes and limited to spectral and temporal resolution for oceanic applications, its high spatial resolution (30 m) makes it ideal for applications in small water bodies [15].

Recent studies have demonstrated the broad potential of Landsat images in lakes and coastal environments (bays and inlets) based on the existent correlations between band reflectance and different water properties, such as: Secchi disk transparency (SDT)[16–18]; concentration of suspended sediments [2,19–21]; turbidity [18,22–24]; studies of colored dissolved organic matter (CDOM) [15,23,25,26] and macroalgal blooms [27]; and quantifying Chl-a concentrations [18,28–32]. Therefore, the main objective of this study was to select the best model to estimate Chl-a concentration from *in situ* measurements and Landsat 8 satellite images in the Bahía de La Paz, Mexico by using multiple linear regression models to evaluate their possible application in monitoring Chl-a concentration in coastal water bodies.

Materials and methods

Study area

The study area is located within the Bahía de La Paz on the western coast of Baja California Sur, Mexico between 24˚ 09' and 24˚ 47' N, and 110˚ 45' and 110˚ 18' W (Fig 1). It is a coastal water body of about 90 km long, 60 km wide and 4500 km^2 with two water mouths that connect it with the western region of the Gulf of California. The main water mouth is wide and 300 m in depth located to the northwest while to the east, the water mouth (small mouth or the San Lorenzo Canal) is narrow and shallow associated with 20-m deep channels [33,34].

The Ensenada de La Paz is a coastal lagoon located in the southern part of the Bahía de La Paz between 24˚ 06' and 24˚ 11' N, and 110˚ 19' and 110˚ 25' W. It is a protected coastal water body separated from the Bahía de La Paz by a marine sandy barrier called "El Mogote", approximately 11 km long in east-western direction and 2.7 km in its widest part [35]. Ensenada de

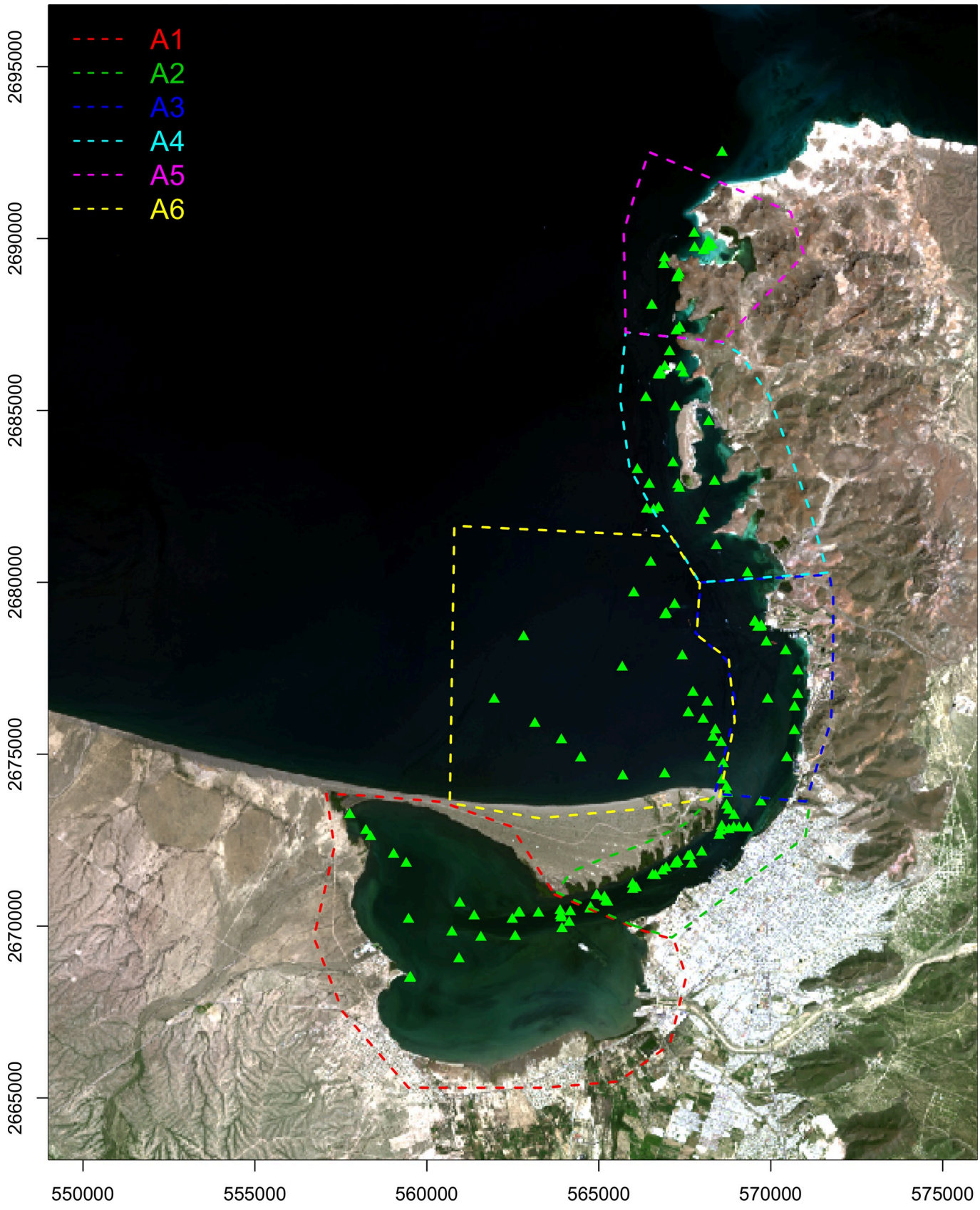
A1
A2
A3
A4
A5
A6
2695000
2690000
2685000
2680000
2675000
2670000
2665000
550000
555000
560000
565000
570000
575000

Fig 1. Map of the geographical location of the study area. Distribution of the sampling sites (green triangles) and arbitrary areas (polygons) used for time series analysis. Map generated in programming language R using Landsat 8 image from 2016-09-09.

https://doi.org/10.1371/journal.pone.0205682.g001

La Paz is 12 km in length, 5 km in width in an area of 45 km^2 with respect to sea level average. Morphologically speaking, the water mouth is formed by two parallel channels in their connection with the Bahía de La Paz of approximately 4 km in length and 0.6 km in width in total with an average depth of 7.0 m [36].

Field data collection

The *in situ* data collection was done synchronously with the overpass of Landsat 8 satellite, which passes by this zone every 16 days at approximately 17:47 UTC. Thus, field trips were made two hours before and after 17:47 UTC to avoid the effect of Chl-a variability related to tides and local currents. The Chl-a concentration was measured near the surface (~ 50 cm deep), taken with the multi-parameter sensor RBRmaestro model XRX-420 produced by RBR Ltd in Ottawa, Canada. No specific permissions were required for our study locations/activities since Chl-a data was collected in non-protected or private locations of the study area.

Twelve field campaigns were performed for over one year of monitoring due to bad weather (mainly high cloudiness), and field trips did not take place some dates of the period of study. Table 1 shows the dates of the field trips made, as well as some descriptive statistics of Chl-a measured *in situ* for the six arbitrary areas (polygons) used for time series analysis (see Fig 1 for details).

Satellite data (Landsat 8 images)

Landsat 8 Level 1 data products were used in this study, which were included in the Landsat 8 OLI/TIRS C1 Level-1 data set and downloaded from the US Geological Survey server (USGS, https://www.usgs.gov) using Earth Explorer platform (https://earthexplorer.usgs.gov). Landsat 8 has two sensors onboard Operational Land Imager (OLI) and Thermal Infrared Sensor (TIRS). In total these sensors had 11 spectral bands, nine of the OLI sensor and two of the TIRS sensor. The spatial resolution of bands 1–7 and 9 was 30 m; band 8 (panchromatic) was 15 m and 100 m for bands 10–11. In this study, the following spectral bands of the visible light portion and near infrared (NIR) were used; B1 (coastal/aerosol: 0.435–0.451 μm); B2 (blue: 0.452–0.512 μm); B3 (green: 0.533–0.590 μm); B4 (red: 0.636–0.673 μm); and B5 (NIR: 0.851–0.879 μm).

The study area was in the Landsat ID scene: LC8034043 (Path = 34, Row = 43). The images were acquired from 2016-08-24 to 2017-06-08, and only were those without cloud cover selected and cropped to highlight the study area (Fig 1). Landsat 8 images were imported and processed with the *raster* library [37] from programming language R [38] version 3.3.2 to obtain water pixel remote sensing reflectance of each one of the selected bands, which was calculated by using the equations in Landsat 8 user manual [13].

Statistical modeling

This study used linear regression (LR) and generalized additive models (GAM) to develop a model to estimate Chl-a concentrations from *in situ* data and spectral reflectance of Landsat 8 bands 1–5 images. LR explored the linear relationship between response and predictor variables; GAM explored linear or non-linear relationships between response and predictor variables throughout smooth functions (e.g. thin plate regression spline). Assuming that error terms (residuals) were independent of the predictor variables, normally distributed with mean 0 and homoscedastic [39, 40].

Table 1. Descriptive statistics of *in situ*-measured Chlorophyll-a concentrations ($\mu g^{*}l^{-1}$).

Date	A1			A2			A3		
	Min	Mean	Max	Min	Mean	Max	Min	Mean	Max
2016-08-24				0.33	0.44	0.55	0.18	0.19	0.20
2016-09-09	1.37	1.37	1.37	0.43	0.72	1.34	0.37	0.51	0.61
2016-09-25	0.95	1.14	1.41	0.61	0.79	1.05			
2016-10-27	0.41	0.58	0.79	0.30	0.33	0.38			
2016-11-28	0.58	0.58	0.58	0.51	0.54	0.56			
2017-01-31	0.42	0.47	0.52	0.39	0.39	0.39	0.52	0.52	0.52
2017-02-16	0.37	0.37	0.37	0.38	0.38	0.38	0.36	0.36	0.36
2017-03-20	0.51	0.85	1.04	0.59	0.75	0.99	0.33	0.33	0.34
2017-04-05	0.33	0.33	0.33	0.39	0.44	0.47	0.19	0.19	0.19
2017-04-21	0.50	0.59	0.68	0.52	0.52	0.53			
2017-05-23	0.35	0.47	0.59	0.26	0.31	0.37	0.33	0.33	0.33
2017-06-08	2.12	2.12	2.12	2.11	2.23	2.37	1.42	2.10	2.71
Date	A4			A5			A6		
	Min	Mean	Max	Min	Mean	Max	Min	Mean	Max
2016-08-24	0.16	0.28	0.66				0.25	0.25	0.25
2016-09-09	0.61	0.61	0.61						
2016-09-25							0.17	0.24	0.41
2016-10-27	0.18	0.18	0.18				0.20	0.23	0.27
2016-11-28	0.40	0.41	0.41	0.38	0.42	0.45	0.34	0.36	0.37
2017-01-31	0.42	0.42	0.42	0.33	0.36	0.41	0.43	0.43	0.43
2017-02-16	0.49	0.60	0.71	0.64	0.66	0.68	0.25	0.25	0.25
2017-03-20	0.49	0.51	0.53	0.43	0.50	0.58			
2017-04-05	0.24	0.28	0.33	0.36	0.36	0.36	0.22	0.22	0.22
2017-04-21	0.16	0.16	0.17	0.15	0.17	0.20	0.18	0.25	0.40
2017-05-23	0.14	0.17	0.24	0.14	0.17	0.23	0.22	0.22	0.22
2017-06-08	0.46	0.94	1.52						

Min, Minimum; Max, maximum; SD. A1 to A6 arbitrary areas (see Fig 1 for details).

https://doi.org/10.1371/journal.pone.0205682.t001

For modeling we used 147 records of log-transformed Chl-a as response variable and spectral bands of the visible light portion (B1 [coastal/aerosol], B2 [blue], B3 [green], and B4 [red]) and near infrared (NIR, B5) as predictor variables. Values of Chl-a concentration were logarithmically transformed and used as dependent variable because some authors have widely described that chlorophyll showed a non-linear relationship with Landsat bands [41,42].

LR models were constructed using a single predictor variable or more than one to highlight the number of predictor variables; we named Simple Linear Regression (SLR) when one predictor variable was used in the model and Multiple Linear Regression (MLR) when two or more predictor variables were used. All data processing and statistical models were conducted in R [38]; *mgcv* library was used for GAM [43].

A different band combination was tested for SLR, MLR, and GAM since previous studies on Chl-a estimation from Landsat imagery suggested that addition, multiplication, proportion, or quadratic transformation of bands 1–5 gave good results in SLR [5,23,42,44,45]. Up to two bands were combined through permutation to be used as predictor variables in SLR. Thus, for SLR we tested a total of 250 different models (S1 Table). For MLR we used band combinations with two, three, four and five predictor variables using spectral bands 1–5 without transformation. For MLR we tested a total of 26 different models (S2 Table). For GAM we used each

single band and band combinations using two, three, four and five predictor variables and tested a total of 31 different models (S3 Table). To date, GAM had not been used to estimate Chl-a from Landsat imagery.

In general, the models can be represented as follows:

$$y_i = \alpha + \beta X + \varepsilon_i \quad (1)$$

$$y_i = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_p X_p + \varepsilon_i \quad (2)$$

$$y_i = \alpha + f_1 X_1 + f_2 X_2 + f_3 X_3 + \cdots + f_p X_p + \varepsilon_i \quad (3)$$

where y_i was the expected value of log-transformed Chl-a concentration ($\mu g^* l^{-1}$); α = intercept; $ß_p$ were the coefficients of predictor variables (X_p), which were bands 1–5; and f_p were smooth functions (thin plate regression spline) of the covariates; ε_i the error terms (residuals) were independent of X and they were assumed to be normally distributed with mean 0 and homoscedasticity. GAMs were used with Gaussian error distribution and identity link function.

Goodness of fit and predictive performance

The quality of model fit was assessed using the R squared (R^2) and adjusted R squared (adj. R^2). These statistics were used to describe the proportion of variance explained and could take values between 0 and 1. The difference between R^2 and adj. R^2 relied in that the latter used a penalization based on the number of parameters (predictor variables); this variation of R^2 was used because it had been demonstrated that R^2 always increased when a new predictor variable was added to the model, causing an overparameterization of models. Therefore, adj. R^2 was preferred for models with two or more predictor variables.

Data splitting is an effective method for evaluating predictive performance of a given model in which a portion of the data is used to estimate model coefficients, and the remainder of the data is used to measure prediction accuracy of the model. In this study data were arbitrarily separated into two subsets, training data (2016-08-24, 2016-09-09, 2016-09-25, 2016-10-27, 2016-11-28, 2017-01-31, 2017-02-16 and 2017-03-20) and test data (2017-04-05, 2017-04-21, 2017–05–23 and 2017-06-08). Thus, after fitting models on the training data, their performance was measured against test data.

Predictive performance of the models was evaluated using root-mean-square error (RMSE) and correlation coefficient (R) (S4–S6 Tables); these values were calculated using observed and predicted data from the test dataset. Predicted data was computed using coefficients or smooth functions of all models; thus RMSE of all models was on the same scale (log-transformed Chl-a). RMSE was defined as:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(X_i^{observed} - X_i^{predicted}\right)^2}$$

Finally, to evaluate model assumptions, we analyzed residuals of the best fitted SLR, MLR, and GAM, respectively to identify if they were normally distributed, homoscedastic and had the presence of outliers.

Chlorophyll-a estimation from Landsat 8

Using the best fitted model and all Landsat 8 scenes available and cloud free for the study area, we obtained the estimated log-transformed Chl-a for the period May 2013—October 2017; log-transformed Chl-a was returned to its original scale by applying the exponential function.

To describe spatial and temporal variability of estimated Chl-a, we analyzed time-series data on six arbitrary areas (polygons) in the study area (Fig 1). These polygons were defined considering geographical features: (1) a semi-closed coastal lagoon (Ensenada de La Paz); (2) the channel that communicates Ensenada de La Paz with Bahía de La Paz; (3) the northernmost part of the city of La Paz; (4) the port city; (5) the second largest presence of mangroves of the study area; and (6) the southernmost portion of Bahía de La Paz locally known as El Mogote.

Results

in situ Chlorophyll-a data

Table 1 shows the descriptive statistics of Chl-a concentration measured during field trips, which ranged from 0.136 to 2.714 $\mu g^{*}l^{-1}$; the highest average value (1.652 $\mu g^{*}l^{-1}$) was recorded in 2017-06-08 and the lowest one (0.252 $\mu g^{*}l^{-1}$) in 2017-05-23; the highest values were recorded during a red tide event. As shown in Table 1, average Chl-a values were usually lower than 0.7.

Fig 2 shows spatial and temporal variability of the observed Chl-a during the survey period. As it can be observed, Chl-a values were usually higher in area 1 and 2 with respect to others. The highest ones were recorded in June 8, 2017 in areas 1–3 and corresponded to the red tide event.

Selection of the best fitted model

A total of 307 models were constructed and evaluated to identify which of them could explain the highest variance proportion of log-transformed Chl-a, inferred from R^2 and adjusted R^2. Table 2 shows the three best fitted SLR, MLR, and GAM, respectively. SLR resulted in the lowest proportion of variance explained (R^2 = 0.000–0.542; adj. R^2 = -0.009–0.538); MLR in a higher proportion of variance explained (R^2 = 0.061–0.764; adj. R^2 = -0.044–0.753) while GAM resulted in the highest proportion of variance explained (adj. R^2 = 0.216–0.854). The SLR with the highest R^2 and adjusted R^2 was the ratio between B4 (red) and $B1^2$ (coastal/aersol), explaining 54.2% of total variance. The MLR with the highest R^2 and adjusted R^2 was that which included the five spectral bands, explaining 76.4% of total variance. The GAM with the highest adjusted R^2 was that which included four spectral bands (B1 [coastal/aerosol], B2 [blue], B3 [green], and B4 [red]), explaining 88.7% of total deviance.

Predictive performance

As mentioned in Methods, predictive performance was evaluated using Pearson coefficient of correlation (R) and root-mean-square error (RMSE) obtained from predictions of the training model on an independent data set. As shown in Table 3, the SLR with the highest R and lowest RMSE was the model with the ratio between B4 and $B1^2$; the MLR with the highest R and lowest RMSE was the one that included four bands (B1 [coastal/aerosol], B2 [blue], B3 [green], and B5 [NIR]); and the GAM with the highest R and lowest RMSE was the one that included B1, B2, and B3. These results indicated that three modeling approaches could predict log-transformed Chl-a with high accuracy.

Fig 3 shows residual patterns of best fitted SLR, MLR, and GAM, respectively. As it can be observed, residuals of the three models seemed to have normal distribution with mean 0, no marked trend (homoscedasticity) in residuals versus fitted, and absence of outliers. Therefore, model assumptions seemed to be fine.

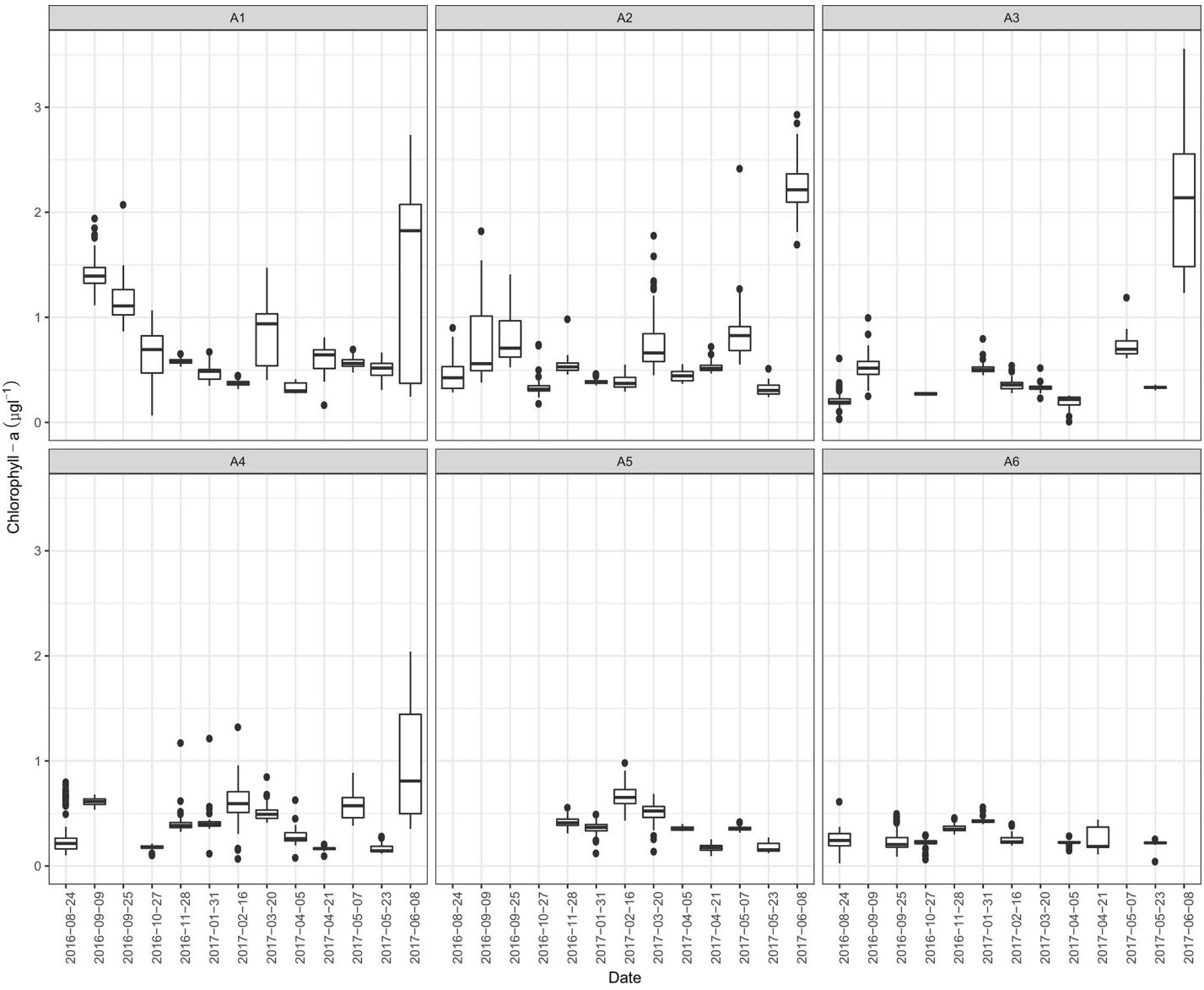

Fig 2. Observed chlorophyll-a values during the survey period. Solid lines represent medians; boxes the interquartile ranges; whiskers minimum and maximum or 1.5 times the interquartile range (when outliers were present); points represent the outliers. A1-A6 arbitrary areas (see Fig 1 for details).

https://doi.org/10.1371/journal.pone.0205682.g002

Optimal model for estimating Chl-a

Given the results of Tables 2 and 3, the MLR with four bands (B1 [coastal/aerosol], B2 [blue], B3 [green], and B5 [NIR]) was considered as the best model to predict log-transformed Chl-a in the study area. Table 4 shows coefficients of this model, from which, we could infer that log-transformed Chl-a had a positive linear relationship with bands B1, B3, and B5 and negative linear relationship with band B2. Coefficients of this model suggested that B2, B3, and B1 had the strongest linear relationship with log-transformed Chl-a; on the contrary B5 had the weakest linear relationship with log-transformed Chl-a.

Table 2. Goodness of fit of the three best fitted SLR, MLR, and GAM, respectively, for log-transformed Chl-a estimation.

Model	R^2	adj.R^2
Simple Linear Regression		
y = 1.84–6.54*(B1^2/B4)	0.506	0.502
y = -3.6 + 1.09*(B4/B1^2)	**0.542**	**0.538**
y = -3.06 + 5.55*(B4/B2)	0.477	0.473
Multiple Linear Regression		
y = 0.94 + 88.45*B1–194.77*B2 + 97.55*B3 + 10.79*B4	0.735	0.725
y = 1.54 + 79.56*B1–191.62*B2 + 102.22*B3 + 13.17*B5	0.757	0.748
y = 1.05 + 103.37*B1–221.63*B2 + 119.1*B3–19.09*B4 + 21.39*B5	**0.764**	**0.753**
Generalized Additive Models		
y = f(B1) + f(B2) + f(B3) + f(B4)		**0.854**
y = f(B1) + f(B2) + f(B3) + f(B5)		0.848
y = f(B1) + f(B2) + f(B3) + f(B4) + f(B5)		0.847

R^2, Coefficient of determination; adj. R^2, adjusted coefficient of determination. In bold the best fitted SLR, MLR, and GAM, respectively.

https://doi.org/10.1371/journal.pone.0205682.t002

Temporal variability of predicted Chlorophyll-a

Predicted values of Chl-a for the period May 2013—October 2017 within the six arbitrary areas defined in this study are shown in Fig 4, ranging from 0.11 to 1.58 μg * l^{-1} and displaying high seasonal variability with peaks of maximum Chl-a during May and June; the lowest values were predicted for December and January and the highest ones corresponded to the red tide event observed in June 8, 2017; however, it is important to notice from model predictions, that this event was not present in area 1 (Ensenada de La Paz).

Spatial variability of predicted Chlorophyll-a

Fig 5 represents the predicted Chl-a concentration of the study area corresponding to the month of June from 2013 to 2017 where the values obtained in each of the dates were

Table 3. Predictive performance of the three best fitted SLR, MLR, and GAM, respectively, applied for log-transformed Chl-a estimation.

Model	R	MSRE
Simple Linear Regression		
y = 1.84–6.54*(B1^2/B4)	0.791	0.288
y = -3.6 + 1.09*(B4/B1^2)	**0.858**	**0.255**
y = -3.06 + 5.55*(B4/B2)	0.849	0.303
Multiple Linear Regression		
y = 1.48 + 81.53*B1–187.8*B2 + 98.22*B3	0.856	0.209
y = 2.13 + 66.02*B1–174.51*B2 + 95.85*B3 + 5.23*B5	**0.875**	**0.191**
y = 1.39 + 95.99*B1–211.35*B2 + 117.22*B3–22.85*B4 + 13.2*B5	0.866	0.194
Generalized Additive Models		
y = f(B1) + f(B2) + f(B3)	**0.835**	**0.239**
y = f(B2) + f(B3) + f(B4)	0.821	0.286
y = f(B1) + f(B2) + f(B3) + f(B4)	0.798	0.274

R, Pearson coefficient of correlation; RMSE, root-mean-square error. In bold SLR, MLR, and GAM, respectively, with the highest predictive performance.

https://doi.org/10.1371/journal.pone.0205682.t003

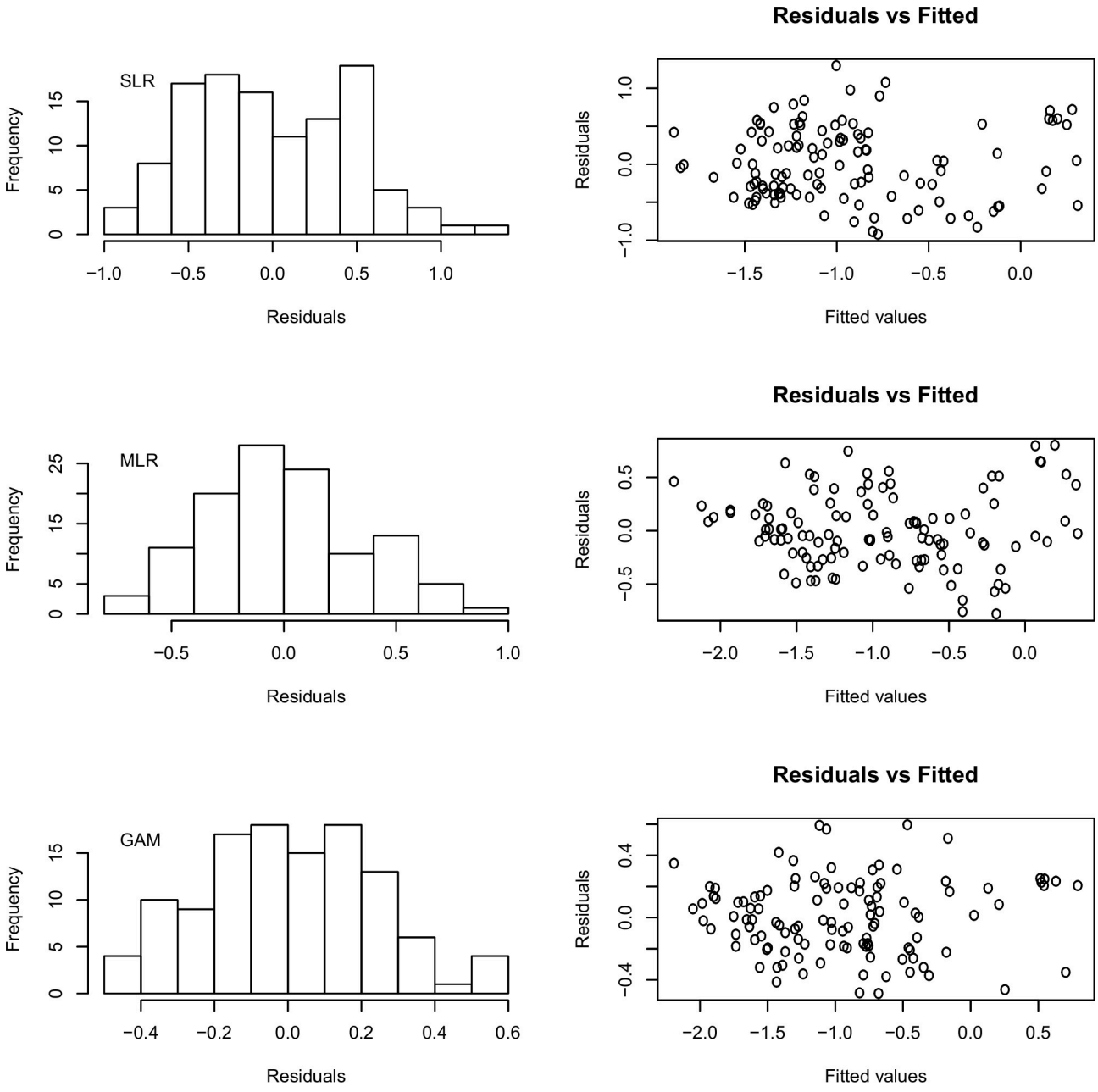

Fig 3. Residual analysis of the best fitted SLR (top), MLR (center), and GAM (bottom), respectively.

https://doi.org/10.1371/journal.pone.0205682.g003

Table 4. Descriptive statistics of coefficients of the best-fitted model.

	Coefficient	Standard error	T value	P
Intercept	1.54	0.77	2.13	0.036
B1 (c/a)	79.56	14.96	5.32	<0.001
B2 (blue)	-191.62	15.55	-11.58	<0.001
B3 (green)	102.22	6.65	15.36	<0.001
B5 (NIR)	13.17	3.70	3.56	<0.001

https://doi.org/10.1371/journal.pone.0205682.t004

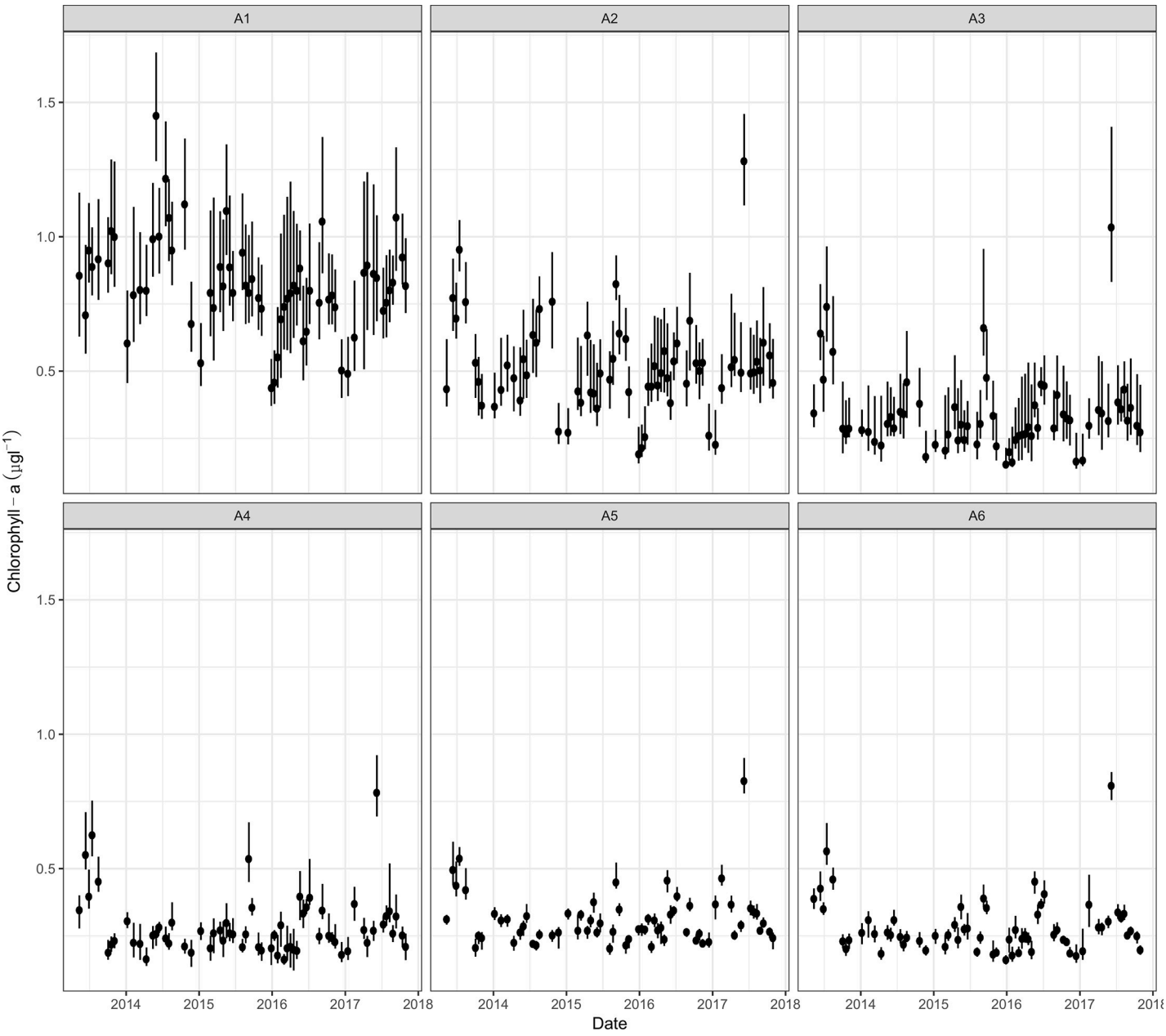

Fig 4. Predictions of Chl-a ($\mu g^{*}l^{-1}$) obtained from the best-fitted model and the Landsat imagery set for the period 2013–2017. Points represent the means; whiskers represent the interquartile ranges. A1-A6 arbitrary areas (see Fig 1 for details).

https://doi.org/10.1371/journal.pone.0205682.g004

compared; high values were observed in 2017-06-08 because this image showed conditions detected during a proliferation event registered in the study area. In the same way, high values were observed in the image corresponding to 2013-06-13 compared with the rest of the images because they corresponded to values prior to the proliferation event reported in literature from June 18 to 20, 2013. It indicated that based on the data obtained *in situ* during a year of monitoring, it was possible to set reference values through those generated by the model; Chl-a anomalies can be detected and can be used as indicators of possible algal proliferation events.

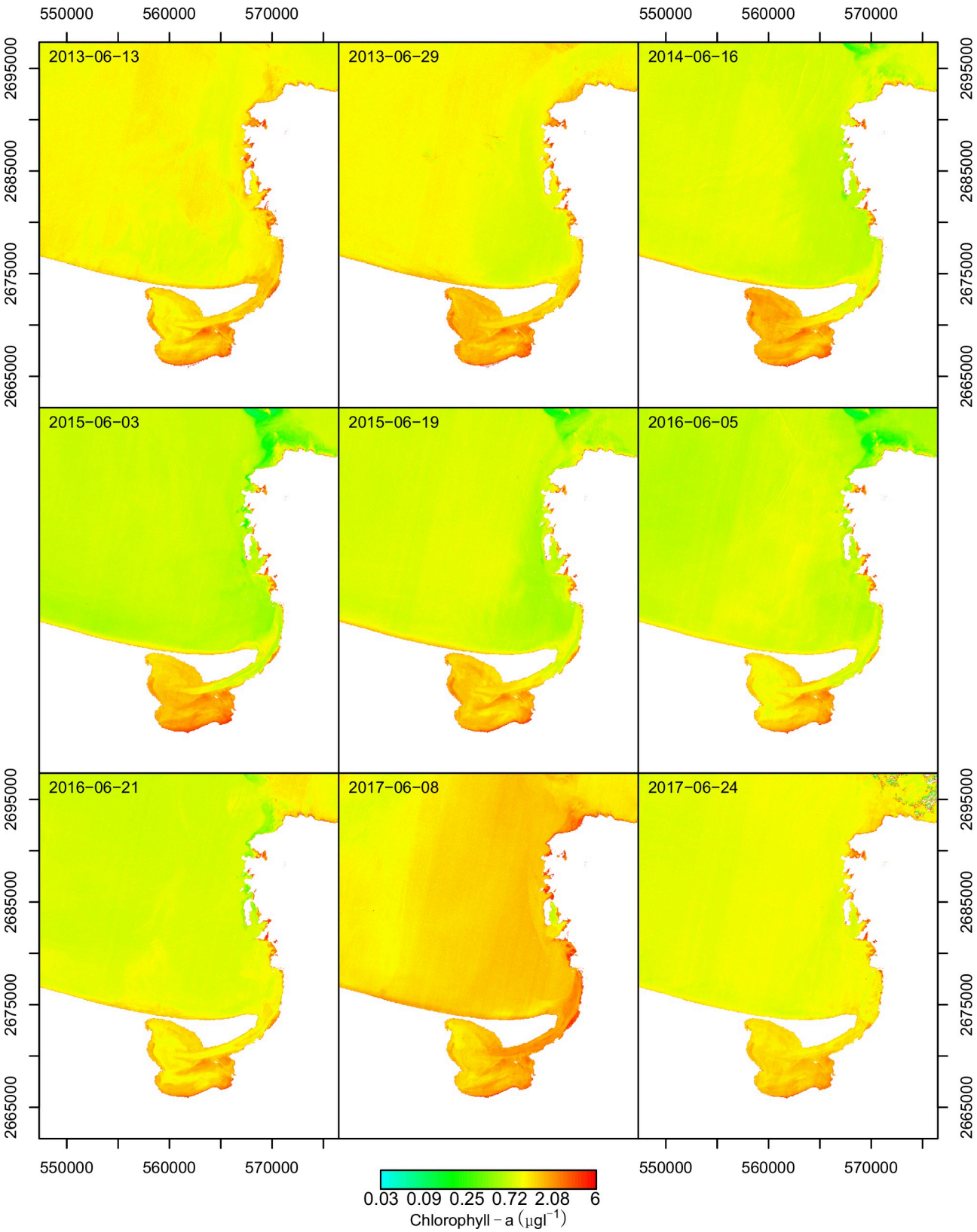

Fig 5. Predicted Chlorophyll-a ($\mu g^{*}l^{-1}$) in the study area, corresponding to June 2013 to 2017. Maps generated in programming language R.

https://doi.org/10.1371/journal.pone.0205682.g005

Discussion

This study evaluated the use of Landsat 8 for estimating Chl-a concentration in the coastal water body located in northwestern Mexico by field data collection, simple linear regression, multiple linear regression and generalized additive models, using as response variable log-transformed Chl-a and reflectance values as spectral predictive variables of the visible part of light and NIR. The results obtained suggested that the use of spectral bands 1 (coastal/aerosol), 2 (blue), 3 (green), and 5 (NIR), from the MLR model, allowed us to reliably estimate the concentrations of Chl-a in a coastal environment.

To date, a large number of available publications have demonstrated that Chl-a can be estimated using Landsat satellite images by *in situ* data collection and using simple or multiple linear regression models, but something that attracted our attention was the great diversity of approaches that have been used for this purpose. For example, some authors have used simple models where the predictive variable was one of the spectral bands [31,32,46]; other authors suggested that the ratio of two spectral bands could be used as a good predictor of Chl-a [18,28,29,45,47]; finally, other authors suggested that various combinations of spectral bands in multiple linear regression models allowed a better estimation of Chl-a in aquatic environments [23,32,45–47].

These approaches generated uncertainty as to which was the best one to estimate Chl-a in aquatic environments. This study used a statistical approach and the Chl-a absorption/reflection theory to create the best possible model, in such a way, that MLR was constructed using spectral bands where Chl-a had its greater absorption/reflection. According to what several researchers have demonstrated, Chl-a had its highest light absorption at wavelengths from 400–500 nm (blue) and 680 nm (red) and its maximum reflection up to 550 nm (green) and 700 nm (NIR). Thus, a negative correlation was expected between Chl-a and reflectance in the blue band; that is, the higher the concentration of Chl-a, the lower reflectance in this wavelength. On the other hand, a positive correlation between Chl-a and reflectance in the green and NIR bands was expected; in other words, the higher the concentration of Chl-a, the higher reflectance in these wavelengths [45,48–51].

Initially, this study evaluated the linear correlation between log-transformed Chl-a and spectral reflectance in five wavelengths (coastal/aerosol, blue, green, red, and NIR) using Pearson correlation coefficients; however, two things that attracted our attention in the results obtained were (1) low correlation ($r < 0.2$) between Chl-a and the selected spectral bands and (2) correlation between Chl-a and the red and NIR bands, which had an opposite sign than that expected. In this regard, several authors have found higher or lower values of Pearson correlation coefficient and Chl-a; for example, Lim & Choi [44] found correlation values greater than 0.6 among the blue, green, red, and NIR bands and Chl-a; however, their results suggested inverse relationships because all correlation values were negative. On the other hand, Patra et al. [45] found correlations smaller than 0.5 and positive among blue, green, red and NIR bands and Chl-a. In both cases, the estimation of Chl-a was performed in freshwater bodies (rivers and lakes), which could have generated these differences with what was found in our study. Usually in freshwater bodies, such as rivers and lakes, turbidity (caused by particulate organic matter) is several times greater than in marine bodies [15,52,53].

Other authors have suggested that the combination of spectral bands by way of ratio (e.g. NIR/red) had a higher correlation with Chl-a [18,44–46,54]. In this regard, our study found higher correlation values between Chl-a and the red and squared transformed coastal/aerosol (B4/B12) band ratio. Another interesting point in this study was the use of the coastal aerosol band (B1) because when it was included, the models increased the correlation value (R) and decreased the value of RMSE. This band was constituted by wavelengths that detect deep blue

and violet very similar to the blue band characterized by low reflectance in environments with high Chl-a concentration. According to Slonecker et al. [26] and Loyd [55], this feature makes the band potentially important for investigating coastal phenomena.

As mentioned above, the selected MLR was used to perform statistical inference, in this particular case, to evaluate the linear relationship between spectral bands and Chl-a by using the coefficients of multiple linear regression models. Our results showed a high concordance between the observed (model) and expected (absorption/reflection properties of Chl-a) results, specifically the negative coefficient of the blue (B2) band and positive coefficients of the green (B3) and NIR (B5) bands. In this regard, Brivio et al. [48] and Lim & Choi [47] used multiple linear regression models to estimate Chl-a (among others) through the use of different Landsat spectral bands; however, the coefficients of the best model used to estimate Chl-a showed the opposite expected signs. For example, positive values with the blue (B2) and negative with green (B3) bands.

To date, many studies have addressed estimating Chl-a from images acquired by satellites, both in freshwater bodies and marine environments, obtaining promising results. Nonetheless, when comparing methods and results, they have shown a great discrepancy in the way Chl-a has been estimated from Landsat images; it may be due to the wavelength used (or proportions among them), the type of statistical method, or the type of environment where the study was performed. All indicates that the methods applied in a specific place or environment cannot be replicated similarly in another place and/or different environment, which suggests the need for greater field validation and spatial or temporal coverage, or plainly and simply a comparison of Chl-a estimation with a standardized method in different types of the aquatic environments required.

Conclusions

This study has evaluated the performance of simple and multiple linear regression and generalized additive models to estimate Chl-a concentration, using the first five bands of Landsat 8 images in the Bahía de La Paz, Baja California Sur, Mexico. The obtained results indicated that this method provided a reliable estimation of Chl-a in small coastal water bodies because of the high coherence found in model coefficients with the absorption/reflection properties of Chl-a evaluated in the laboratory under controlled conditions. Therefore, remote sensing has shown to represent an ideal opportunity to develop regional scale research on various parameters in environments estimated in small coastal water bodies to allow a constant monitoring at low cost and high-quality spatial scale.

Supporting information

S1 Table. Goodness of fit of the SLR models.
(DOCX)

S2 Table. Goodness of fit of the MLR models.
(DOCX)

S3 Table. Goodness of fit of the GAM models.
(DOCX)

S4 Table. Predictive performance of the SLR models.
(DOCX)

S5 Table. Predictive performance of the MLR models.
(DOCX)

S6 Table. Predictive performance of the GAM models.
(DOCX)

Acknowledgments

The authors thank Mario Cota Castro for help provided during the sampling survey and Diana Fischer for editing and improving English in this manuscript. First author would like to thank CONACyT for the scholarship provided.

Author Contributions

Conceptualization: Raúl Octavio Martínez-Rincón.

Formal analysis: Miguel Ángel Matus-Hernández, Raúl Octavio Martínez-Rincón.

Funding acquisition: Norma Yolanda Hernández-Saavedra.

Methodology: Miguel Ángel Matus-Hernández, Raúl Octavio Martínez-Rincón.

Project administration: Norma Yolanda Hernández-Saavedra.

Supervision: Norma Yolanda Hernández-Saavedra, Raúl Octavio Martínez-Rincón.

Writing – original draft: Miguel Ángel Matus-Hernández, Norma Yolanda Hernández-Saavedra, Raúl Octavio Martínez-Rincón.

References

1. Lee Z, Shang S, Qi L, Yan J, Lin G. A semi-analytical scheme to estimate Secchi-disk depth from Landsat-8 measurements. Remote Sens Environ. Elsevier Inc. 2016; 177: 101–106. https://doi.org/10.1016/j.rse.2016.02.033
2. Teodoro AC. Optical Satellite Remote Sensing of the Coastal Zone Environment—An Overview. In: Marghany M, editor. Environmental Applications of Remote Sensing. 1st ed. InTech; 2016. pp. 165–196. https://doi.org/10.5772/61974
3. Liu YS, Islam MA, Gao J. Quantification of shallow water quality parameters by means of remote sensing. Prog Phys Geogr. 2003; 27: 24–43. https://doi.org/10.1191/0309133303pp357ra
4. Boyer JN, Kelble CR, Ortner PB, Rudnick DT. Phytoplankton bloom status: Chlorophyll a biomass as an indicator of water quality condition in the southern estuaries of Florida, USA. Ecol Indic. 2009; 9: S56–S67. https://doi.org/10.1016/j.ecolind.2008.11.013
5. Nazeer M, Nichol JE. Development and application of a remote sensing-based Chlorophyll-a concentration prediction model for complex coastal waters of Hong Kong. J Hydrol. Elsevier B.V. 2016; 532: 80–89. https://doi.org/10.1016/j.jhydrol.2015.11.037
6. Guo Q, Wu X, Bing Q, Pan Y, Wang Z, Fu Y, et al. Study on retrieval of chlorophyll-a concentration based on Landsat OLI Imagery in the Haihe River, China. Sustain Switz. 2016; 8. https://doi.org/10.3390/su8080758
7. Antoine D, André J-M, Morel A. Oceanic primary production: 2. Estimation at global scale from satellite (Coastal Zone Color Scanner) chlorophyll. Glob Biogeochem Cycles. 1996; 10: 57–69. https://doi.org/10.1029/95GB02832
8. O'Reilly JE, Maritorena S, Mitchell BG, Siegel DA, Carder KL, Garver SA, Kahru MCM, Kahru M, McClain C. Ocean color chlorophyll algorithms for SeaWiFS. J Geophys Res. 1998; 103: 24937–24953.
9. O'Reilly JE, Maritorena S, O'Brien M, Siegel DA, Toole D, Menzies D, et al. SeaWiFS Postlaunch Calibration and Validation Analyses, Part 3. In: Hooker SB, editor. SeaWiFS Postlaunch Technical Report Series. 2000. pp. 1–49.
10. González L, EJMT. Neural network estimation of chlorophyll a from MERIS full resolution data for the coastal waters of Galician rias (NW Spain). Remote Sens Environ. 2011; 115: 524–535. https://doi.org/10.1016/j.rse.2010.09.021

11. Gitelson AA, Dall'Olmo G, Moses WM, Rundquist DC, Barrow T, Fisher TR, Gurlin DHH. A simple semi-analytical model for remote estimation of chlorophyll-a in turbid waters: Validation. Remote Sensingof Enviroment. 2008; 112: 3582–3593. https://doi.org/10.1016/j.rse.2008.04.015
12. Liu D, Wang Y. Trends of satellite derived chlorophyll-a (1997–2011) in the Bohai and Yellow Seas, China: Effects of bathymetry on seasonal and inter-annual patterns. Prog Oceanogr. Elsevier Ltd. 2013; 116: 154–166. https://doi.org/10.1016/j.pocean.2013.07.003
13. Zanter K. Landsat 8 (L8) Data Users Handbook Version 1.0 June 2015. 2015;8.
14. Jiménez-Muñoz JC, Sobrino JA, Skoković D, Mattar C, Cristóbal J. Land surface temperature retrieval methods from Landsat-8 thermal infrared sensor data. Geosci Remote Sens Lett IEEE. 2014; 11: 1840–1843. https://doi.org/10.1109/LGRS.2014.2312032
15. Joshi I, D'Sa EJ. Seasonal variation of colored dissolved organic matter in barataria bay, Louisiana, using combined landsat and field data. Remote Sens. 2015; 7: 12478–12502. https://doi.org/10.3390/rs70912478
16. Kloiber SM, Brezonik PL, Olmanson LG, Bauer ME. A procedure for regional lake water clarity assessment using Landsat multispectral data. Remote Sens Environ. 2002; 82: 38–47. https://doi.org/10.1016/S0034-4257(02)00022-6
17. Dekker AG, Peters SWM. The use of the Thematic Mapper for the analysis of eutrophic lakes: a case study in the Netherlands. Int J Remote Sens. Taylor & Francis Group. 1993; 14: 799–821. https://doi.org/10.1080/01431169308904379
18. Masocha M, Dube T, Nhiwatiwa T, Choruma D. Testing utility of Landsat 8 for remote assessment of water quality in two subtropical African reservoirs with contrasting trophic states. Geocarto Int. Taylor & Francis. 2017; 6049: 1–14. https://doi.org/10.1080/10106049.2017.1289561
19. Schiebe FR, Harrington JA, Ritchie JC. Remote sensing of suspended sediments: the Lake Chicot, Arkansas project. Int J Remote Sens. Taylor & Francis Group. 1992; 13: 1487–1509. https://doi.org/10.1080/01431169208904204
20. Ouillon S, Douillet P, Petrenko A, Neveux J, Dupouy C, Froidefond J-M, et al. Optical Algorithms at Satellite Wavelengths for Total Suspended Matter in Tropical Coastal Waters. Sensors. Multidisciplinary Digital Publishing Institute (MDPI); 2008; 8: 4165–4185. https://doi.org/10.3390/s8074165 PMID: 27879929
21. Kong J-L, Sun X-M, Wong D, Chen Y, Yang J, Yan Y, et al. A Semi-Analytical Model for Remote Sensing Retrieval of Suspended Sediment Concentration in the Gulf of Bohai, China. Remote Sens. Multidisciplinary Digital Publishing Institute. 2015; 7: 5373–5397. https://doi.org/10.3390/rs70505373
22. Hellweger FL, Schlosser P, Lall U, Weissel JK. Use of satellite imagery for water quality studies in New York Harbor. Estuar Coast Shelf Sci. 2004; 61: 437–448. https://doi.org/10.1016/j.ecss.2004.06.019
23. Brezonik P, Menken KD, Bauer M. Landsat-based Remote Sensing of Lake Water Quality Characteristics, Including Chlorophyll and Colored Dissolved Organic Matter (CDOM). Lake Reserv Manag. 2005; 21: 373–382. https://doi.org/10.1080/07438140509354442
24. Olmanson LG, Bauer ME, Brezonik PL. A 20-year Landsat water clarity census of Minnesota's 10,000 lakes. Remote Sens Environ J. 2008; 112: 4086–4097. https://doi.org/10.1016/j.rse.2007.12.013
25. Griffin CG, Frey KE, Rogan J, Holmes RM. Spatial and interannual variability of dissolved organic matter in the Kolyma River, East Siberia, observed using satellite imagery. J Geophys Res. 2011; 116: G03018. https://doi.org/10.1029/2010JG001634
26. Slonecker ET, Jones DK, Pellerin BA. The new Landsat 8 potential for remote sensing of colored dissolved organic matter (CDOM). Mar Pollut Bull. Elsevier B.V.; 2016; 107: 518–527. https://doi.org/10.1016/j.marpolbul.2016.02.076 PMID: 27004998
27. Xing Q, Hu C. Mapping macroalgal blooms in the Yellow Sea and East China Sea using HJ-1 and Landsat data: Application of a virtual baseline reflectance height technique. Remote Sens Environ. Elsevier Inc.; 2016; 178: 113–126. https://doi.org/10.1016/j.rse.2016.02.065
28. Giardino C, Pepe M, Brivio PA, Ghezzi P, Zilioli E. Detecting chlorophyll, Secchi disk depth and surface temperature in a sub-alpine lake using Landsat imagery. Sci Total Environ. 2001; 268: 19–29. https://doi.org/10.1016/S0048-9697(00)00692-6 PMID: 11315741
29. Han L, Jordan KJ. Estimating and mapping chlorophyll- *a* concentration in Pensacola Bay, Florida using Landsat ETM+ data. Int J Remote Sens. 2005; 26: 5245–5254. https://doi.org/10.1080/01431160500219182
30. Allan MG, Hamilton DP, Hicks BJ, Brabyn L. Landsat remote sensing of chlorophyll *a* concentrations in central North Island lakes of New Zealand. Int J Remote Sens. 2011; 32: 2037–2055. https://doi.org/10.1080/01431161003645840

31. Allan MG, Hamilton DP, Hicks B, Brabyn L. Empirical and semi-analytical chlorophyll a algorithms for multi-temporal monitoring of New Zealand lakes using Landsat. Environ Monit Assess. 2015;187. https://doi.org/10.1007/s10661-015-4397-6

32. Kim HH, Ko BC, Nam JY. Predicting chlorophyll- *a* using Landsat 8 OLI sensor data and the non-linear RANSAC method–a case study of Nakdong River, South Korea. Int J Remote Sens. Taylor & Francis. 2016; 37: 3255–3271. https://doi.org/10.1080/01431161.2016.1196839

33. Álvarez-Arellano AD, Rojas-Soriano H, Prieto-Mendoza JJ. Geología de la Bahía de La Paz y áreas adyacentes. In: Urbán-Ramírez J, Ramírez-Rodríguez M, editors. La Bahía de La Paz, investigación y conservación. México: Universidad Autónoma de Baja California Sur. 1997. p. 345.

34. Jiménez-Illescas AR, Obeso-Nieblas M, Salas-de-León DA. Oceanografía física de La Bahía de La Paz, B.C.S. In: Urbán-Ramírez J, Ramírez-Rodríguez M, editors. La Bahía de La Paz, investigación y conservación. México: Universidad Autónoma de Baja California Sur.; 1997. p. 345.

35. Cervantes-Duarte R, Guerrero-Godínez R. Variación espacio-temporal de nutrientes de la Ensenada de La Paz, B.C.S. Inst Cienc Mar Limnol. 1987; 15: 129–142.

36. Jiménez-Illescas AR, Alatorre-Mendieta MA, Obeso-Nieblas M, Shirasago-Germán B, Garcia-Escobar H. Efectos de la construcción de un canal artificial entre la Ensenada y la Bahía de La Paz. 2008; Available: http://repositoriodigital.ipn.mx//handle/123456789/14242

37. Hijmans RJ. Raster: Geographic Data Analysis and Modeling. 2017. R package version 2.6–7

38. R Core Team. R: A language and environment for statistical computing. 2018. R Foundation for Statistical Computing, Vienna, Austria.

39. Zuur AF, Ieno EN, Walker N, Saveliev AA, Smith GM. Mixed effects models and extensions in ecology with R. New York: Springer. 2009.

40. Hastie TJ, Tibshirani RJ. Generalized additive models. New York: Chapman & Hall/CRC. 1990.

41. Chang K-W, Shen Y, Chen P-C. Predicting algal bloom in the Techi reservoir using Landsat TM data. Int J Remote Sens. 2004; 25: 3411–3422. https://doi.org/10.1080/01431160310001620786

42. Han L, Jordan KJ. Estimating and mapping chlorophyll-a concentration in Pensacola Bay, Florida using Landsat ETM+ data. Int J Remote Sens. 2005; 26: 5245–5254. https://doi.org/10.1080/01431160500219182

43. Wood SN. Generalized Additive Models: An Introduction with R. New York: Chapman & Hall/CRC. 2006.

44. Tebbs EJ, Remedios JJ, Harper DM. Remote sensing of chlorophyll-a as a measure of cyanobacterial biomass in Lake Bogoria, a hypertrophic, saline-alkaline, flamingo lake, using Landsat ETM+. Remote Sens Environ. Elsevier Inc. 2013; 135: 92–106. https://doi.org/10.1016/j.rse.2013.03.024

45. Patra PP, Dubey SK, Trivedi RK, Suhu SK, Rout SK. Estimation of Chlorophyll-a Concentration and Trophic States for an Inland Lake from Landsat-8 OLI Data: A Case Nalban Lake of East Kalkota Wetland, India. Preprints. 2016; 18. https://doi.org/10.20944/preprints201608.0149.v1

46. Torbick N, Corbiere M. A multiscale mapping assessment of lake champlain cyanobacterial harmful algal blooms. Int J Environ Res Public Health. 2015; 12: 11560–11578. https://doi.org/10.3390/ijerph120911560 PMID: 26389930

47. Lim J, Choi M. Assessment of water quality based on Landsat 8 operational land imager associated with human activities in Korea. Environ Monit Assess. 2015; 187: 1–17. https://doi.org/10.1007/s10661-014-4167-x

48. Brivio PA, Giardino C, Zilioli E. Determination of chlorophyll concentration changes in Lake Garda using an image-based radiative transfer code for Landsat TM images. Int J Remote Sens. 2001; 22: 487–502. https://doi.org/10.1080/014311601450059

49. Singh K., Ghosh M., Sharma S.R. PK. Blue–Red–NIR Model for Chlorophyll- a Retrieval in Hypersaline–Alkaline Water Using Landsat ETM+ Sensor. IEEE J Sel Top Appl Earth Obs Remote Sens. 2014; 7: 3553–3559. https://doi.org/10.1109/JSTARS.2014.2340856

50. Arenz RF, Lewis WM, Saunders JF. Determination of chlorophyll and dissolved organic carbon from reflectance data for Colorado reservoirs. Int J Remote Sens. 1996; 17: 1547–1566. https://doi.org/doi.org/10.1080/01431169608948723

51. Dube T. Primary Productivity of Intertidal mudflats in the Wadden Sea: A Remote Sensing Method. University Twente-ITC. 2012.

52. Dalu T, Dube T, Froneman PW, Sachikonye MTB, Clegg BW, Nhiwatiwa T. An assessment of chlorophyll- *a* concentration spatio-temporal variation using Landsat satellite data, in a small tropical reservoir. Geocarto Int. 2015; 30: 1130–1143. https://doi.org/10.1080/10106049.2015.1027292

53. Del Castillo CE, Gilbes F, Coble PG, Müller-Karger FE. On the dispersal of riverine colored dissolved organic matter over the West Florida Shelf. Limnol Oceanogr. 2000; 45: 1425–1432. https://doi.org/10.4319/lo.2000.45.6.1425

54. Watanabe FSY, Alcânntara E, Rodrigues TWP, Imai NN, Barbosa CCF, Rotta LH da S. Estimation of chlorophyll-a concentration and the trophic state of the barra bonita hydroelectric reservoir using OLI/landsat-8 images. Int J Environ Res Public Health. 2015; 12: 10391–10417. https://doi.org/10.3390/ijerph120910391 PMID: 26322489

55. Loyd C. Putting Landsat 8's Bands to Work. 2013 Jun 14 [cited 29 May 2017]. In: Landsat website [internet]. Available from: https://www.mapbox.com/blog/putting-landsat-8-bands-to-work/