Article

Algal Pigment Estimation Models to Assess Bloom Toxicity in a South American Lake

Lien Rodríguez-López [1,*] , David Francisco Bustos Usta [2] , Lisandra Bravo Alvarez [3] , Iongel Duran-Llacer [4,5] , Luc Bourrel [6] , Frederic Frappart [7] , Rolando Cardenas [8] and Roberto Urrutia [9]

1 Facultad de Ingeniería, Arquitectura y Diseño, Universidad San Sebastián, Lientur 1457, Concepción 4030000, Chile
2 Facultad de Oceanografía, Universidad de Concepción, Concepción 4030000, Chile; davidbustos@udec.cl
3 Department of Electrical Engineering, Universidad de Concepción, Edmundo Larenas 219, Concepción 4030000, Chile; lisanbravo@udec.cl
4 Escuela de Ingeniería en Medio Ambiente y Sustentabilidad, Escuela de Ingeniería Forestal, Facultad de Ciencias, Ingeniería y Tecnología, Universidad Mayor, Camino La Pirámide 5750, Santiago 8580745, Chile; iongel.duran@umayor.cl
5 Hémera Centro de Observación de la Tierra, Facultad de Ciencias, Ingeniería y Tecnología, Universidad Mayor, Camino La Pirámide 5750, Santiago 8580745, Chile
6 Géosciences Environnement Toulouse, UMR 5563, Université de Toulouse, CNRS-IRD-OMP-CNES, 31400 Toulouse, France; luc.bourrel@ird.fr
7 INRAE, Bordeaux Sciences Agro, UMR 1391 ISPA, 33140 Villenave-d'Ornon, France; frederic.frappart@inrae.fr
8 Facultad de Matemática- Física- Computación, Universidad Central Martha Abreu de las Villas, Santa Clara 50100, Cuba; rcardenas@uclv.edu.cu
9 Facultad de Ciencias Ambientales, Universidad de Concepción, Concepción 4030000, Chile; rurrutia@udec.cl
* Correspondence: lien.rodriguez@uss.cl

Abstract: In this study, we build an empirical model to estimate pigments in the South American Lake Villarrica. We use data from Dirección General de Aguas de Chile during the period of 1989–2024 to analyze the behavior of limnological parameters and trophic condition in the lake. Four seasonal linear regression models were developed by us, using a set of water quality variables that explain the values of phycocyanin pigment in Lake Villarrica. In the first case, we related chlorophyll-a (Chl-a) to phycocyanin, expecting to find a direct relationship between both variables, but this was not fulfilled for all seasons of the year. In the second case, in addition to Chl-a, we included water temperature, since this parameter has a great influence on the algal photosynthesis process, and we obtained better results. We discovered a typical seasonal variability given by temperature fluctuations in Lake Villarrica, where in the spring, summer, and autumn seasons, conditions are favorable for algal blooms, while in winter, the natural seasonal conditions do not allow increases in algal productivity. For a third case, we included the turbidity variable along with the variables mentioned above and the statistical performance metrics of the models improved significantly, obtaining R^2 values of up to 0.90 in the case of the model for the fall season and a mean squared error (MSE) of 0.04 µg/L. In the last case used, we added the variable dissolved organic matter (MOD), and the models showed a slight improvement in their performance. These models may be applicable to other lakes with harmful algal blooms in order to alert the community to the potential toxicity of these events.

Keywords: phycocyanin; harmful algal bloom; lake; Chile

1. Introduction

Pigments found in algae provide fundamental information on the trophic status of aquatic ecosystems [1,2]. Key pigments, such as chlorophyll, phycocyanin, fucoxanthin, and phycobilin, play essential roles in photosynthesis and solar energy absorption in algae [3,4]. The concentration and variety of these pigments are influenced by factors such as nutrient levels, light intensity, and the presence of contaminants [5,6]. Consequently, the

analysis of algal pigments allows the assessment of primary productivity, water quality, and eutrophication of aquatic environments [7,8]. An increase in chlorophyll levels can mean an increase in algal biomass due to nutrient enrichment, which can lead to problems such as harmful algal blooms and hypoxia [9,10]. Therefore, the study of algal pigments is vital to monitor and understand the health and balance of aquatic ecosystems and to design effective management strategies for their conservation and restoration [11,12].

Low chlorophyll-a (Chl-a) concentrations in lakes are characteristic of oligotrophic systems, whereas higher concentrations are typical of mesotrophic and eutrophic systems [13–15]. This variation is closely related to nutrient availability: oligotrophic lakes are low in nutrients, whereas mesotrophic and eutrophic lakes have higher nutrient levels [16–18]. However, due to changing climatic conditions and the global increase in land surface temperatures, the frequency and intensity of algal blooms have increased in lakes and other aquatic systems [19–21]. This phenomenon is due to several factors, including increased nutrient runoff from agricultural and urban areas, increased water temperature, which favors algal growth, and increased thermal stability, which allows algae to bloom for longer periods [22–24]. As a result, even lakes historically classified as oligotrophic may now experience more frequent and severe algal blooms, raising concerns about water quality and the health of aquatic ecosystems [25,26]. It is crucial to closely monitor these changes and implement effective management measures to mitigate the adverse effects of algal blooms on human health, aquatic biodiversity, and sustainable resource use [27].

In recent years, much progress has been made in the development of indices for the detection of aquatic vegetation [28–33]. Numerous indices have been created and evaluated, ranging from those adapted from agricultural and terrestrial vegetation applications to combinations of spectral bands specifically designed to detect primary productivity in aquatic environments [30,34–36]. Despite these advances, significant challenges remain. Many indices do not perform optimally in aquatic ecosystems other than those for which they were originally developed and validated [37]. This is because individual lake conditions are influenced by a variety of factors, such as geographic, meteorological, and physicochemical characteristics [38].

In Chile, several lakes have witnessed numerous episodes of algal blooms, the most recent being Laguna Grande de San Pedro (bloom all year in 2023), Lake Laja (May 2023), and Lake Villarrica during the summer months from 1993 to 2023–2024 where the bloom continues until autumn, with records of species such as *Microcystis aeruginosa* and *Dolichospermum* sp., and *Dolichospermum circinale*, all belonging to the cyanophycean group [36,39]. In response to this problem, several studies have been carried out in recent years to detect and predict water quality parameters and the occurrence of algal blooms in the lakes of central-southern Chile [25,36,40,41]. In this context, the objective of our work is to create and validate pigment estimation algorithms to detect the pigment phycocyanin, characteristic of algal blooms of the cyanobacteria group, in Lake Villarrica. Therefore, it is crucial to fill these gaps and develop more accurate and robust methods, adaptable to the diverse conditions present in different aquatic ecosystems, integrating statistical regression techniques with data from various sources, from in situ monitoring of water quality parameters to data obtained from meteorological stations. Improving these methods will increase the capacity to monitor and manage aquatic vegetation, contributing to the conservation and sustainable management of these vital natural resources.

2. Materials and Methods

2.1. Lake Villarrica, Study Area

Lake Villarrica is located at 39°18′ S south latitude and 72°05′ west longitude in the Araucanía Region (see Figure 1). It is one of the most important lakes in Chile, recognized for its scenic beauty and its economic and tourist importance. However, due to multiple anthropogenic impacts, the lake's condition has changed from oligotrophic to mesotrophic, which has led to the implementation of Decree N°19/2013 of the Ministry of Environment, which establishes secondary environmental quality standards for the protection of

the surface continental waters of Lake Villarrica as well as the Exempt Resolution SMA N°671/2016 which proposes and justifies the modification of the water quality monitoring performed by the DGA in Lake Villarrica. In addition, algal bloom episodes have been recorded in recent decades, with an increase in severity in recent years, likely exacerbated by the effects of climate change [36].
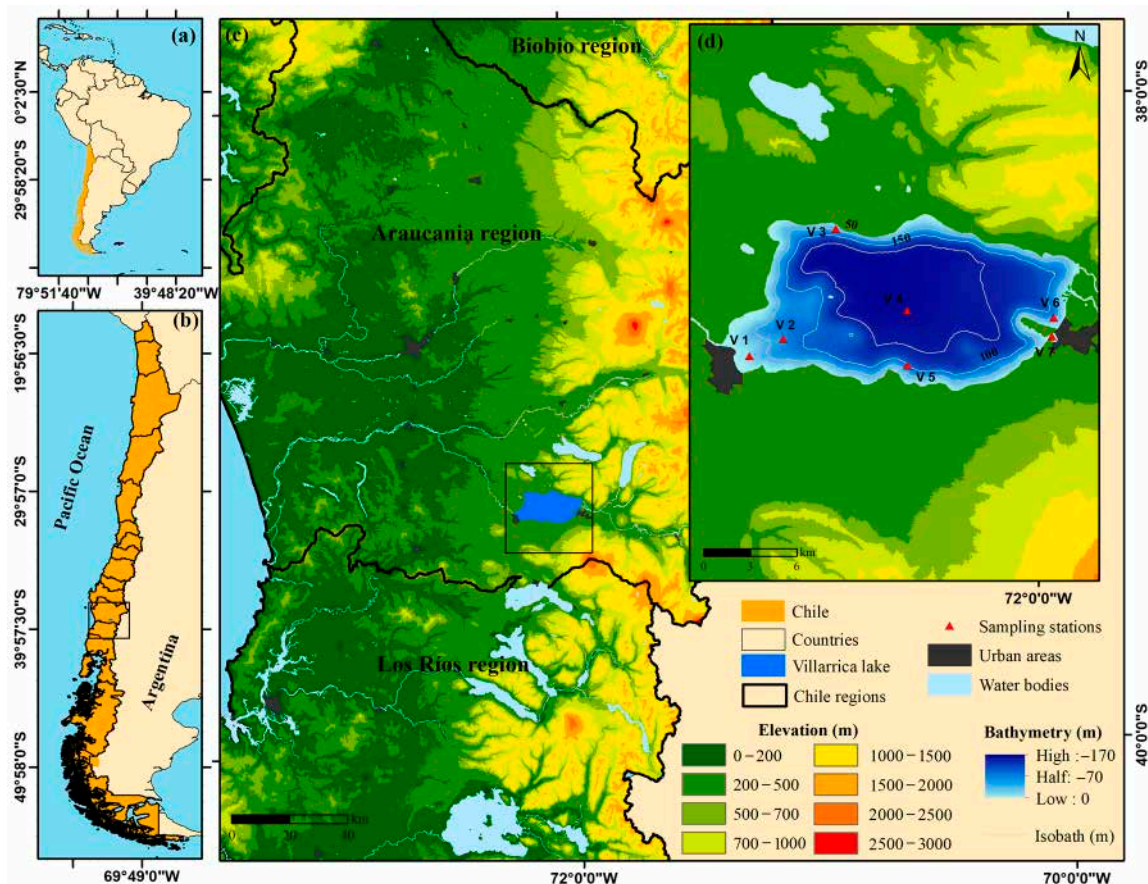

Figure 1. (**a**) South America continent, (**b**) Chile in Latin America, (**c**) Region de la Araucanía and the location of Lake Villarrica in the black box, (**d**) Lake Villarrica and seven sampling stations.

2.2. Collection of Data from Monitoring Campaigns

Water quality parameters such as water temperature (C°), chlorophyll-a (µg/L); phycocyanin (µg/L), turbidity (NTU), and dissolved oxygen (ppm) were obtained from monitoring campaigns conducted by the Dirección General de Aguas de Chile (DGA) from 1989 to 2024. The data collected at 7 monitoring stations as described in Rodríguez-López et al., 2023, cover the 4 seasons of the year from 1989 to 2009. After this date, probably due to limited resources, only the summer and spring seasons are measured until the present. The parameters monitored included surface temperature (measured using the standard thermometry method 2250 B, as given in the compendium NCh 2313), surface chlorophyll-a (Chl-a, measured using the fluorometric method), total phosphorus (determined using the standard method 4500 P B, 22nd edition, EAM), total nitrogen (determined using the standard method 4500-N C, 22nd edition, EAM), and water transparency (measured using the Secchi disk depth, SDD). Water samples were collected at five different depths at each station in the lake using a 5 L Niskin sampling bottle. Samples were stored and transported to a sample collection center. Samples were stored and transported in thermally insulated boxes, cooled with ice to maintain a temperature of approximately 5 °C, for subsequent analysis. Chemical analyses were performed at the chemical laboratory of the Dirección General de Aguas de Chile (DGA), accredited by the Instituto Nacional de Normalización under Chilean Standard NCh ISO 17,025 of 2005.

2.3. Data Processing

A methodology previously used in another Chilean aquatic system [36] was employed, in which rows with more than 80% of features with null values were removed from the data set. The data treatment process began with the treatment of outliers, data integrity analysis, and imputation of missing values. Several biogeochemical and physical variables were considered, such as chlorophyll-a (Chl-a), temperature (Temp), total nitrogen (N), total phosphorus (P), phycocyanin (PC), dissolved oxygen (O_D), turbidity (NTU), and dissolved organic matter (MOD). In addition, location variables (latitude and longitude) and temporal variables (year, month, day and time) were included. Dummy variables were created to link each measurement to its respective sampling station. In total, 4 covariates (independent variables) were used (Chl-a, Temp, O_D, and NTU) to predict phycocyanin levels (the dependent variable). To determine the number of variables, we used two criteria: First, linear correlation between covariates and the dependent variable and second, results from stepwise regression, a method for selecting optimal predictors which uses both backward and forward selection to mitigate multicollinearity [42] examining the entire range of potential models as defined by $2^k - 1$ (where k is the number of predictors, in this case, 7 covariates and 126 potential models) using statistical criterion Akaike Information Criterion (AIC) calculated for every possible model of each size [43].

Data cleaning was performed separately for each sampling station (La Poza and Litoral Pucón).

The cleaning steps were as follows

1. Drop duplicate rows: Due to the nature of the data capture process, some rows are duplicated across various years. Thus, we remove these duplicate rows to ensure data integrity.
2. Elimination of non-numeric values: Non-numeric values of the selected variables were replaced by null values.
3. Extract temporal data: The year, month, and day of each measurement were extracted and checked for consistency and completeness.
4. Imputation of null values: A sensitive imputation for null values was applied using the median, a robust measure of central tendency for each station and season considered.
5. Detection and capping of outliers: We used the Isolation Forest technique to detect multivariate outliers in the data. This method utilizes recursive partitioning within a tree structure to isolate samples that deviate from most of the data [44]. To fill the outliers, we use the kNN imputer, a technique that locates the k nearest neighbors for a missing datum from complete instances in a data set. It fills in the missing value with the most frequent neighbor value if the feature is categorical (majority rule) or with the mean if the feature is numerical [45].
6. Standardizing the numerical variables: The numerical variables (MOD, NTU, O_D, PC, Temp, and Chl-a) were standardized using the PowerTransformer method [46]. This technique transforms the numerical variables to approximate them to a uniform or Gaussian probability distribution [47].

2.4. Regression Estimation Models

To develop a set of empirical models, we used in situ measurements of Chl-a, phycocyanin (PC), temperature (Temp), dissolved oxygen (DO), and turbidity (NTU) collected during monitoring campaigns from 2021 to 2024. We tested a total of $2^7 - 1 = 127$ different models for the lake based on, first, the stepwise regression results (Section 2.3) and, second, the best correlation between phycocyanin (F) and the other variables. The best correlations were tested using linear (Lin) and multiparametric regression models defined in Equations (1) and (2), respectively.

$$R = \frac{\sum_{i=1}^{n} \left(X_i - \overline{X} \right) \left(Y_i - \overline{Y} \right)}{\sqrt{\sum_{i=1}^{n} \left(X_i - \overline{X} \right)^2 \left(Y_i - \overline{Y} \right)^2}} \tag{1}$$

where:

n is the number of data points.

X_i and Y_i are the individual sample points of variables.

$\overline{X}$ and $\overline{Y}$ are the mean values of X and Y.

$$Y = \beta_o + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon \tag{2}$$

where:

Y is the dependent variable (PC).

β_o is the intercept, representing the expected value of Y when there are no effects from covariates.

$\beta_i (i = 1, 2, \ldots, p)$ are the regression coefficients corresponding to the independent variables X_i.

$X_i (i = 1, 2, \ldots, p)$ are the independent variables of predictors (Section 2,3).

ε is the error term.

Additionally, we used the mean squared error (MSE) metric to assess the adequacy of the models.

For the estimation of coefficients from Equation (2), we used the ordinary least squares (OLS) method (Equation (3)) which minimizes the sum of squares error (SSE, Equation (4)).

$$\hat{\beta} = \left(X^T X\right)^{-1} X^T Y \tag{3}$$

where:

$\hat{\beta}$ are the estimated coefficients.

X is the design matrix of independent variables including the intercept.

Y is the vector of observed values.

$$SSE = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2 \tag{4}$$

We select this method over others such as partial least squares, weighted least squares, ridge, and lasso regression for the following reasons:

1. Linear relationship: There is a linear correlation between the dependent variable (PC) and covariates like Chl-a.
2. Gauss–Markov Theorem: Ordinary least squares (OLS) satisfies the criteria of being the best linear unbiased estimator (BLUE). It has the smallest variance among all linear unbiased estimators, being linear in its parameters, and unbiased with expected values equal to true parameter values.
3. Interpretability: Coefficients from the OLS model are easy to interpret.
4. Simplicity: OLS satisfies the principle of parsimony, which is beneficial for model simplicity.
5. Robustness and sufficient sample size: We have a sufficient sample size of at least 40 samples per station (Table 1) and season across years, meeting the requirements of the Central Limit Theorem.
6. Non-high dimensionality: With only 7 potential predictors, regularization methods are unnecessary.
7. Robustness to outliers: We implemented robust techniques such as Isolation Forest and kNN for outlier detection and outliers' imputation (see Section 2.3), reducing the need for methods like Huber regression.

The quality of the model fit was evaluated using the sample Pearson correlation coefficient (r) and the coefficient of determination (R^2). Additionally, we used the Akaike Information Criteria (AIC, Equation (5)) in the stepwise regression procedure to rank the best models.

$$AIC = n * \ln\left(\frac{SSE}{n}\right) + 2K \tag{5}$$

where:

  n is the number of data points.

  SSE is the sum of squares error defined in Equation (4).

  K is the number of parameters.

Table 1. Number of samples per each station and season used in this study.

STATION	SEASON	SAMPLES
LA POZA	Spring	125
	Summer	74
	Winter	62
	Autumn	96
PUCON	Spring	158
	Summer	145
	Winter	32
	Autumn	161

2.5. Sampling of the Algal Community

Phytoplankton monitoring was carried out on the ground at the expense of the General Directorate of Water, for which quantitative phytoplankton samples (N = 134) were collected in 500 mL bottles with a "Van Dorn" bottle at different depths and at the surface level in some effluents and tributaries of the system. Qualitative sampling was carried out, which consisted of trawling with a 60-micron phytoplankton net at each sampling station of the lake. Each sample was stored in 500 mL plastic bottles, duly labeled, and kept cool at 4 °C, subsequently preserved with 1% Lugol solution until observed.

To identify the species of microalgae present in the water samples, a qualitative analysis was carried out, through observation under a Carl Zeiss Axioskop microscope (Carl Zeiss AG, based in Oberkochen, Germany) with the 40× objective, making an inventory of all the taxa present. The samples were analyzed by experts in continental water microalgae from the Phytoplankton and Phytobenthic Laboratory of the EULA Center.

3. Results

3.1. Water Quality Behavior During Period 2021–2024

We carried out a study on the behavior of the parameters Ch-a, turbidity, water temperature, dissolved oxygen and dissolved organic matter, and phycocyanin during the period 2021–2024 as shown in Figure 2. We observed that Chl-a values increased in the autumn season until reaching 13.81 µg/L, while phycocyanin ranged between 0.17 and 4.75 µg/L in the summer, being the highest values reported (for more details, review the supporting information Tables S1–S4). On the other hand, the temperature in the last period (4 years) has slightly increased its average; however, the maximum values of 22.27 °C are being reached in the summer months coinciding with the maximums of the pigments Chl-a and PC specifically in the month of February, allowing the favorable conditions for carrying out photosynthesis of primary producers to reach a greater temporality in the year.

3.2. Analysis of Correlation Between Quality Parameters

Figure 3 depicts the correlation matrix between the dependent variable, phycocyanin pigment (PC), and the covariates. The matrix reveals a proportional increase for Chl-a, NTU, and Temp, with Chl-a showing the highest correlation coefficient (0.43) among all covariates. Conversely, O_D and MOD display inverse relationships with PC, with correlation coefficients ranging from −0.16 to −0.21. These associations support the general variability conditions between these variables as suggested by [48]. The associated *p*-values are seen in Figure S1.
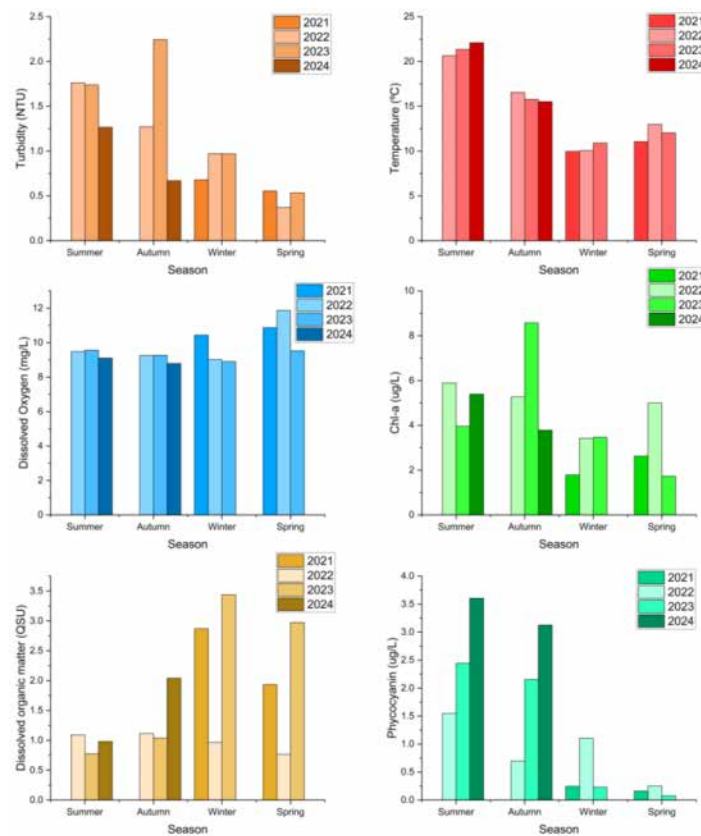
Figure 2. Behavior of limnological parameters in Lake Villarrica during the period 2021–2024.
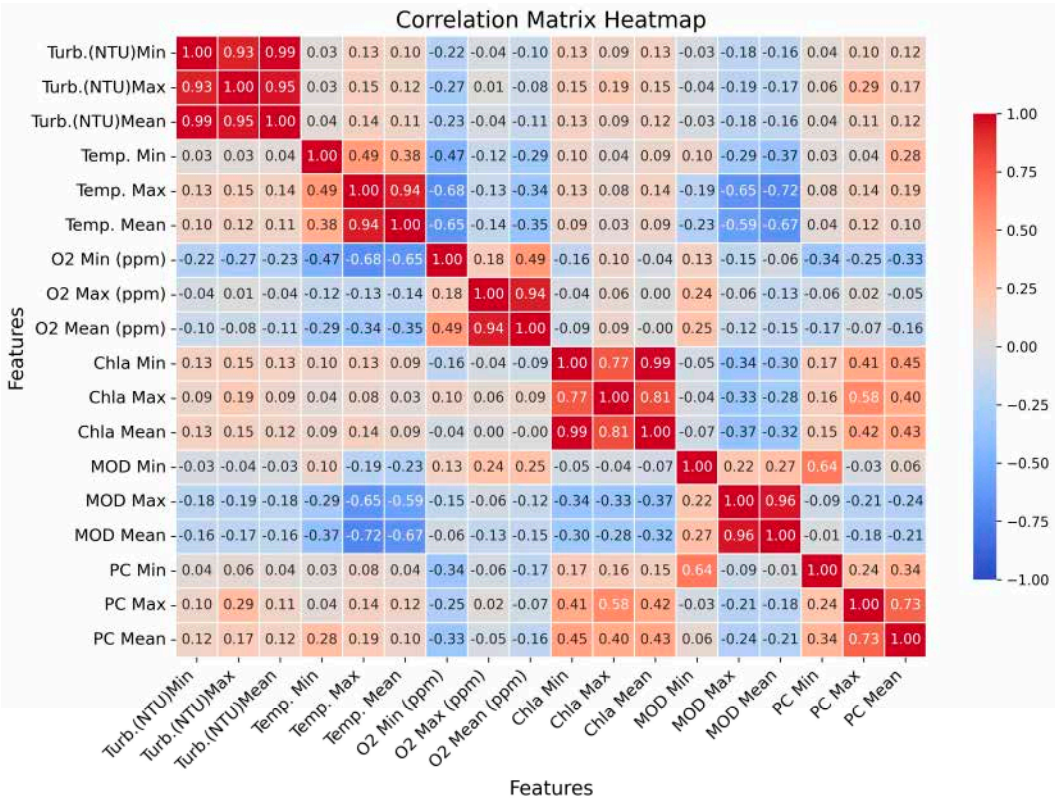

Figure 3. Correlation matrix between the predictors (NTU, Temp, O_D, Chl-a, MOD) and the dependent variable (PC) (Sections 2.3 and 2.4).

3.3. Estimation Models

With the best correlation found in in situ variables, we develop a seasonal estimation model for phycocyanin pigment.

Based on these results and the output from the stepwise regression technique, we have selected the following models as the most suitable for estimating PC, based on the best AIC values for each season considered.

Case 1: PC ~ f(Chla) (−237 < AIC < 234).

Case 2: PC ~ f(Chla+Temp) (−280 < AIC < 288).

Case 3: PC ~ f(Chla+Temp+NTU) (−544 < AIC < 243).

Case 4: PC ~ f(Chla+Temp+NTU+MOD) (−690 < AIC < 245).

O_D was excluded because including this regressor resulted in an approximately 30% increase in AIC values across all seasons considered.

The results from the best models detailed in Section 3.2 are described below.

3.3.1. Case 1: PC ~ f(Chl-a) (−237 < AIC < 234)

In this case, the determination coefficients range between 0.51 and 0.83, with autumn showing the best performance. Additionally, the MSE metrics range from 0.02 μg/L to 1.21 μg/L, with higher values observed in summer (Table 2). The coefficient for the Chl-a estimator is consistently positive across all seasons. Specifically, similar values are observed for spring and autumn (0.036 μg/L and 0.031 μg/L, respectively), while approximately ten times higher values are noted in summer and winter (0.256 and 0.305 μg/L, respectively). These results indicate a positive linear correlation between PC and Chl-a, consistent with the correlation coefficient of 0.43 observed in Figure 3.

Table 2. Linear regression models for estimating phycocyanin pigment. Each column displays the MSE and R^2 for each season, along with the model's equation in each row. Chl-a used as covariable.

SEASON	EQUATION	MSE (μG/L)	R^2	*p*-Value
SPRING	Y = 0.036 × Chl-a + 0.037	0.03	0.51	3.18×10^{-24}
SUMMER	Y = 0.256 × Chl-a − 0.285	1.21	0.63	6.45×10^{-25}
AUTUMN	Y = 0.031 × Chl-a − 0.0024	0.23	0.83	2.41×10^{-42}
WINTER	Y = 0.305 × Chl-a − 0.3598	0.02	0.62	7.44×10^{-44}

3.3.2. Case 2: PC ~ f(Chl-a+Temp) (−280 < AIC < 288)

In this scenario, including an additional variable alongside Chl-a leads to higher correlation coefficients and consequently higher determination coefficients, given the nature of R^2 as a monotonic estimator [49]. Comparing the coefficients for Chl-a reveals similarities to those obtained in Case 1 (Table 3).

Table 3. Linear regression models for estimating phycocyanin pigment. Each column displays the MSE and R^2 for each season, along with the model's equation in each row. Chl-a and Temp used as covariables.

SEASON	EQUATION	MSE (μG/L)	R^2	*p*-Value
SPRING	y = 0.040 × Chl-a − 0.011 × Temp + 0.165	0.03	0.52	4.29×10^{-24}
SUMMER	y = 0.240 × Chl-a − 0.150 × Temp + 2.847	1.09	0.67	2.05×10^{-26}
AUTUMN	y = 0.032 × Chl-a + 0.009 × Temp − 0.155	0.08	0.82	3.64×10^{-41}
WINTER	y = 0.247 × Chl-a + 0.048 × Temp − 0.783	0.04	0.65	3.72×10^{-52}

For temperature (Temp), we observe an inverse relationship with PC during spring and summer, with the strongest negative correlation observed in summer (−0.150). Conversely, during autumn and winter, the relationship appears to be positive but with lower intensity as indicated by smaller magnitude values (<0.048).

3.3.3. Case 3: PC ~ f(Chl-a+Temp+NTU) ($-544 <$ AIC < 243)

In this case, the coefficients for Chl-a and Temp are similar to those obtained in Cases 1 and 2, indicating consistency in their relationships. Another significant finding is that turbidity (NTU) exhibits a positive relationship with PC across all seasons, with higher values observed in summer and autumn (up to 0.354).

The R^2 values are consistently higher compared to the first two cases explored, as expected, and the MSE is lower for all seasons compared to the previous two cases (Table 4).

Table 4. Linear regression models for estimating phycocyanin pigment. Each column displays the MSE and R^2 for each season, along with the model's equation in each row. Chl-a, Temp, and NTU used as covariables.

SEASON	EQUATION	MSE (μG/L)	R^2	p-Value
SPRING	y = 0.037 × Chl-a − 0.018 × Temp + 0.005 × NTU + 0.250	0.03	0.53	3.99×10^{-29}
SUMMER	y = 0.201 × Chl-a − 0.143 × Temp + 0.354 × NTU + 2.421	0.74	0.78	3.62×10^{-27}
AUTUMN	y = 0.020 × Chl-a + 0.007 × Temp + 0.174 × NTU − 0.231	0.04	0.90	3.64×10^{-41}
WINTER	y = 0.101 × Chl-a − 0.014 × Temp + 0.046 × NTU + 0.037	0.02	0.78	2.04×10^{-107}

3.3.4. Case 4: PC~ f(Chla+Temp+NTU+MOD) ($-690 <$ AIC < 245)

In this final case, considering Chl-a, Temp, NTU, and MOD, we observe a proportional increase between MOD and PC during summer and autumn. In spring and winter, there appears to be an inverse relationship, albeit less clearly defined with coefficients $<|0.1|$. All R^2 and MSE metrics show slight improvements compared to those obtained in Case 3. See Table 5 for more details.

Table 5. Linear regression models for estimating phycocyanin pigment. Each column displays the MSE and R^2 for each season, along with the model's equation in each row. Chl-a, Temp, NTU, and MOD used as covariables.

SEASON	EQUATION	MSE (μg/L)	R^2	p-Value
SPRING	y = 0.032 × Chl-a − 0.035 × Temp + 0.006 × NTU −0.064 × MOD + 0.619	0.02	0.55	4.73×10^{-37}
SUMMER	y = 0.205 × Chl-a − 0.137 × Temp + 0.364 × NTU + 0.424 × MOD + 1.769	0.63	0.81	1.55×10^{-26}
AUTUMN	y = 0.020 × Chl-a + 0.007 × Temp + 0.174 × NTU + 0.040 × MOD − 0.244	0.04	0.91	6.55×10^{-49}
WINTER	y = 0.101 × Chl-a − 0.014 × Temp + 0.046 × NTU − 0.075 × MOD + 0.629	0.02	0.79	5.92×10^{-138}

3.4. Metrics Assessment

We used the results from Section 3.3 to create a final assessment between the different models evaluated (Figure 4). The time series plot (left column) illustrates the predicted values generated by each model equation, whilst the R^2 scores (top right) and mean squared errors (bottom right) provide a quantitative comparison of model performance. Notably, Model 4, which incorporates the largest number of regressors, emerges as the top-performing model. This outcome is consistent with mathematical principles, which suggest that a model with a greater number of regressors is likely to provide a better fit to the data.

3.5. Specific Composition and Relative Abundance of the Phytoplankton Community

The phytoplankton community in Lake Villarrica is composed of 7 Phyla, 10 classes, 34 genera, and 53 species, of which 23 spp. They were diatoms of the classes Bacillariophyceae (S = 19 spp.), Coscinodiscophyceae (S = 3 spp.), and Mediophyceae (S = 1); 16 spp. green algae of the classes Chlorophyceae (S = 4), Zygnematophyceae (S = 9 spp.), Trebouxiophyceae (S = 2 spp.), Klebsormidiophyceae (S = 1), 4 cyanobacteria, 7 spp. dinoflagellates, and 3 spp. Cryptophyceae (see Figure 5).
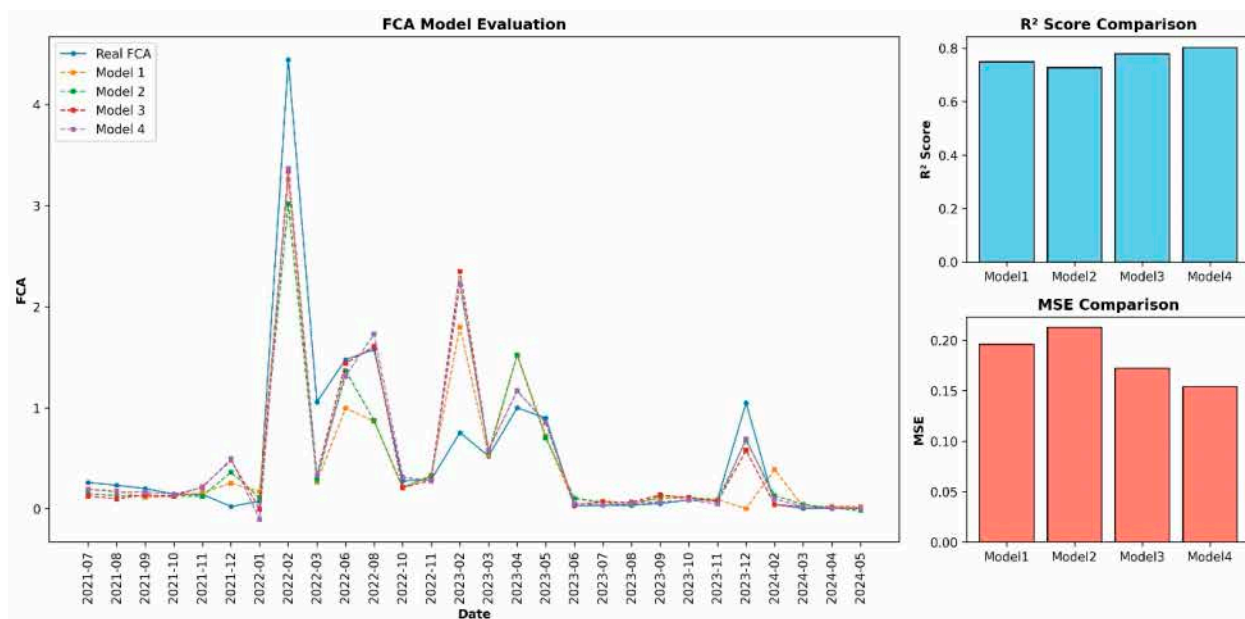
Figure 4. Model performance is evaluated through a time series plot of predicted vs. actual FCA values (left) and comparisons of R^2 scores (top right) and mean squared errors (bottom right) across different models.
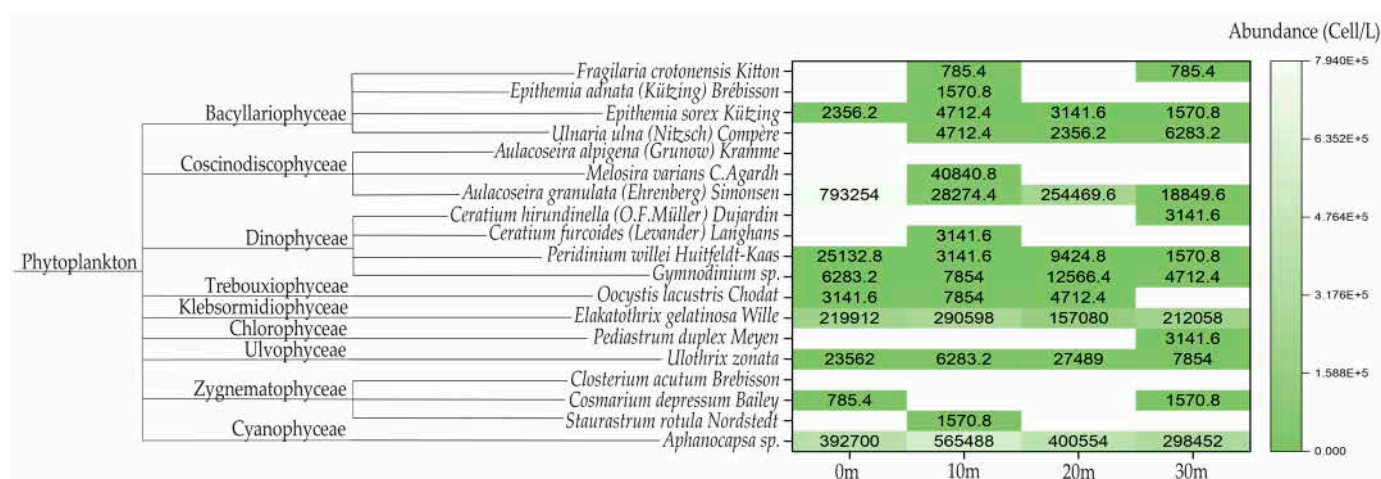

Figure 5. Phytoplankton community present in Lake Villarica and their abundance according to depth over 2021–2024.

The pennate and centric diatoms (Bacillariophyceae 54% and Coscinodiscophyceae 19%) represented the highest abundance in the community. The most abundant species were the pennate diatom *Fragilaria crotonensis* and the centric *Aulacoseira granulata*. The cyanobacterium *Dolichospermum lemmermannii* was also abundant (10%) (see Figure 6).

The dinoflagellate, an invasive species *Ceratium*, was recorded in 88% of the samples analyzed, presenting an abundance that varied between 50,266 Cell/L in spring 2021 at the surface level of the Litoral Pucon sector (see Table 6).
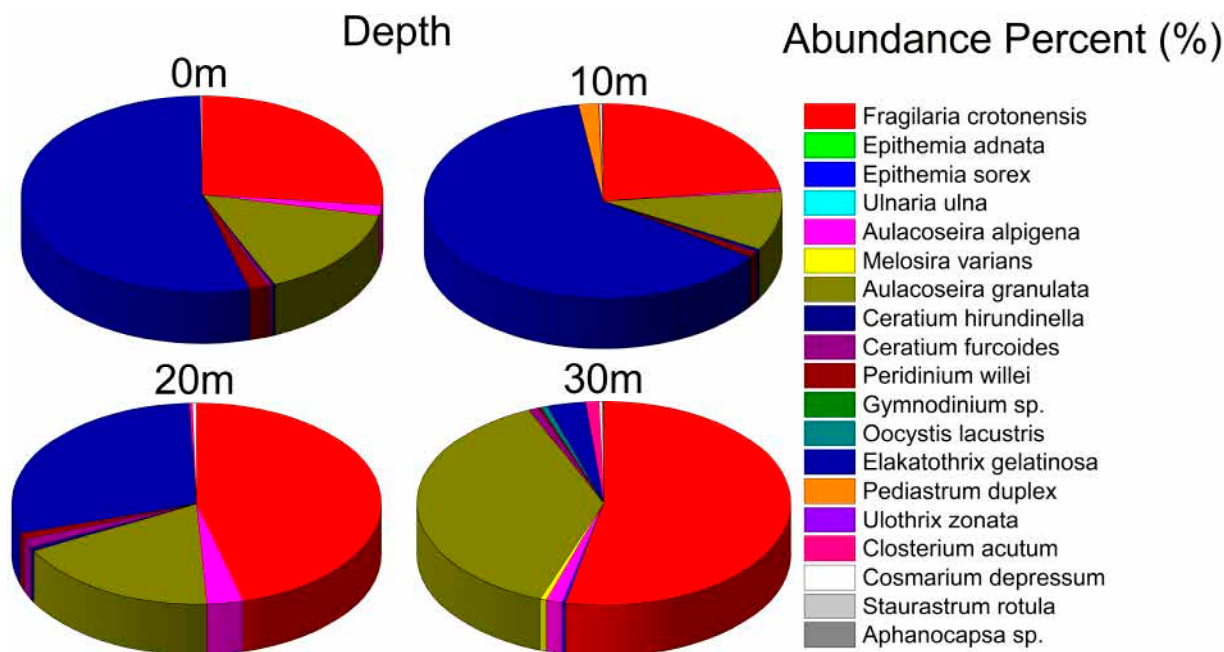
Figure 6. Abundance percent of algal species in Lake Villarica over 2021–2024.

Table 6. Comparison of the phytoplankton community in Lake Villarrica. Adapted from DGA 2023.

DGA MONITOR YEAR	SEASON	NUMBER OF SAMPLES	NUMBER OF CLASS (%)	NUMBER OF SPECIES	SPECIES MORE ABUNDANT	MAX. ABUNDANT CYANOPHYCEAN	MAX. ABUNDANT OF CERATIUM
2018–2020		100	Bacillariophyceae (40%)	63	Diatoms: *Fragilaria crotonensis.* Cyanobacteria: *Dolichospermum planctonicum*, Green algae: *Mucidosphaerium pulchellum.*	692.799 Cell/L *Aphanocapsa incerta*	129.495 Cell/L
2020–2021		72	Bacillariophyceae (93%)	101	Diatoms: *Fragilaria crotonensis* and *Aulacoseira granulata* Dinoflagellate: *Ceratium hirundinella*	219.912 Filaments/L *Dolichospermum lemmermanii*	72.517 Cell/L
2021–2022		120	Bacillariophyceae (54%)	54	Diatoms: *Fragilaria crotonensis* and *Aulacoseira granulata* Cyanobacteria: *Dolicospermum lemmermanii*	1.021.020 Cell/L *Dolicospermum lemmermanii*	50.266 Cell/L

4. Discussion

The development and evolution of algal blooms in lakes is determined by a complex combination of chemical, physical, and biological factors. Chemical factors include the concentration of nutrients such as nitrogen and phosphorus, which are essential for algal growth. Physical factors include aspects such as water temperature, available sunlight, and thermal stratification of the lake. Biological factors include the presence and abundance of other aquatic species that may compete with algae or, in some cases, favor their proliferation. Finally, meteorological factors, such as precipitation, winds, and seasonal variations, also play a crucial role in the dynamics of algal blooms. The interaction of all these elements determines the frequency, intensity, and duration of these phenomena, which can have important ecological and public health implications.

Not all algal blooms are toxic, so the study of algal pigments makes it possible to identify the species that perform the bloom and determine to which group of algae they belong. Algal pigments act as specific markers that facilitate this identification. The most frequent group that forms toxic blooms are cyanobacteria, also known as blue-green algae. These are characterized by pigments such as chlorophyll-a (Chl-a), common to all algae, and other characteristic pigments such as phycocyanin, which is unique to this group. The analysis of these pigments is essential for the management and monitoring of algal blooms, as it allows predicting and mitigating possible negative impacts on the environment and human health. In addition, understanding the conditions that favor the growth of cyanobacteria can help develop more effective prevention and control strategies.

For this reason, in this first preliminary work, we used four seasonal linear regression models developed by us, using a set of water quality variables that explain the behavior of phycocyanin pigment in Lake Villarrica. In the first case, we related chlorophyll-a (Chl-a) to phycocyanin, expecting to find a direct relationship between both variables, but this was not fulfilled for all seasons of the year. In Case 2, in addition to Chl-a, we included water temperature, since this parameter has a great influence on the algal photosynthesis process, and we obtained better results. We discovered a typical seasonal variability given by temperature fluctuations in Lake Villarrica, where in the spring, summer, and autumn seasons, conditions are favorable for algal blooms, while in winter, the natural seasonal conditions do not allow increases in algal productivity. For a third case, we included the turbidity variable along with the variables mentioned above and the statistical performance metrics of the models improved significantly, obtaining R^2 values up to 0.90 in the case of the fall model and a mean squared error (MSE) of 0.04 µg/L. In the last case used, we added the variable MOD, and the models showed a slight improvement in their performance. Our results are consistent with the work of [48], where through machine learning models they estimated phycocyanin in a reservoir in Brazil.

This methodological approach allowed us to better identify and quantify the factors influencing phycocyanin dynamics in the lake, which is crucial for effective management and monitoring of algal blooms. The results obtained indicate that the incorporation of multiple environmental variables and their seasonal analysis provide a more accurate and detailed understanding of the processes affecting cyanobacterial proliferation, thus allowing us to develop more informed and effective management strategies.

By analyzing the algal community of Lake Villarrica through samples taken in situ in the field, we were able to verify through the results of the identification of major algal groups and species that the most abundant group in the lake are the Bacillariophyceae (9 species) followed by the group of green algae including Charophyta and Cyanophycean. In addition, we found invasive species such as *Ceratium* and many bloom-forming cyanobacteria such as *Dolichospermum* sp.

In our future work, we plan to incorporate additional variables derived from satellite sources. This expansion will allow us to improve the interpretation of the temporal evolution of algae and more effectively analyze spatial differences within the lake. By integrating these supplemental data sources, we aim to achieve a more complete understanding of algal population dynamics over time and across regions of the lake. This approach will allow us to better monitor and manage the ecological health of the lake, providing valuable insights into the factors influencing algal growth and distribution.

5. Conclusions

The study of algal pigments is useful for understanding algal dynamics in aquatic systems. We build models for estimating algal pigments (phycocyanin) for different seasons of the year through a set of in situ data taken in Lake Villarrica in southern Chile for the period 2021–2024. We tested a total of $2^7 - 1 = 127$ different models for the lake based, firstly, on the stepwise regression results and, secondly, on the best correlation between phycocyanin (PC) and the other variables. The best correlations were tested using linear (Lin) and multiparametric regression models. We separated them into 4 cases according

to the importance weights of each variable and how, through their relationships with phycocyanin, the models performed best. We found that the model of Case 3 where the chlorophyll-a, temperature, and turbidity variables were incorporated presented better metrics and precision in the estimation of the study variable (R^2 = 0.90 and MSE = 0.04 μg/L). We will continue in future work to improve the accuracy of our models and understanding of algal blooms, in addition to incorporating other data sources such as meteorological and satellite data.

Supplementary Materials: The following supporting information can be downloaded at https://www.mdpi.com/article/10.3390/w16243708/s1, Table S1: Monthly average behavior of limnological variables for the year 2021, Table S2: Monthly average behavior of limnological variables for the year 2022, Table S3: Monthly average behavior of limnological variables for the year 2023, and Table S4: Monthly average behavior of limnological variables for the year 2024. Figure S1: Correlation matrix between the predictors (NTU, Temp, O_D, Chl-a, MOD) and the dependent variable (PC) (Sections 2.3 and 2.4) and *p*-values.

References

1. Pinto, R.; Vilarinho, R.; Carvalho, A.P.; Moreira, J.A.; Guimarães, L.; Oliva-Teles, L. Raman Spectroscopy Applied to Diatoms (Microalgae, Bacillariophyta): Prospective Use in the Environmental Diagnosis of Freshwater Ecosystems. *Water Res.* **2021**, *198*, 117102. [CrossRef] [PubMed]
2. Rodríguez-López, L.; Lami, A.; El Ouahabi, M.; Fagel, N.; Álvarez, D.; González-Rodríguez, L.; Schmidt, S.; Urrutia, R. Fossil Pigments and Environmental Conditions in the Oligotrophic Laja Lake in the Chilean Andes. *Anthropocene* **2022**, *37*, 100321. [CrossRef]
3. Patel, A.K.; Albarico, F.P.J.B.; Perumal, P.K.; Vadrale, A.P.; Ntan, C.T.; Chau, H.T.B.; Anwar, C.; ud dinWani, H.M.; Pal, A.; Saini, R.; et al. Algae as an Emerging Source of Bioactive Pigments. *Bioresour. Technol.* **2022**, *351*, 126910. [CrossRef] [PubMed]
4. Chini Zittelli, G.; Lauceri, R.; Faraloni, C.; Silva Benavides, A.M.; Torzillo, G. Valuable Pigments from Microalgae: Phycobiliproteins, Primary Carotenoids, and Fucoxanthin. *Photochem. Photobiol. Sci.* **2023**, *22*, 1733–1789. [CrossRef]
5. Elisabeth, B.; Rayen, F.; Behnam, T. Microalgae Culture Quality Indicators: A Review. *Crit. Rev. Biotechnol.* **2021**, *41*, 457–473. [CrossRef] [PubMed]
6. López-Sánchez, A.; Silva-Gálvez, A.L.; Aguilar-Juárez, Ó.; Senés-Guerrero, C.; Orozco-Nunnelly, D.A.; Carrillo-Nieves, D.; Gradilla-Hernández, M.S. Microalgae-Based Livestock Wastewater Treatment (MbWT) as a Circular Bioeconomy Approach: Enhancement of Biomass Productivity, Pollutant Removal and High-Value Compound Production. *J. Environ. Manag.* **2022**, *308*, 114612. [CrossRef]
7. de Oliveira Soares Silva Mizael, J.; Cardoso-Silva, S.; Frascareli, D.; Pompêo, M.L.M.; Moschini-Carlos, V. Ecosystem History of a Tropical Reservoir Revealed by Metals, Nutrients and Photosynthetic Pigments Preserved in Sediments. *Catena* **2020**, *184*, 104242. [CrossRef]
8. Belle, S.; Delcamp, E.; Nilsson, L.J.; Freiberg, R.; Appleby, G.P.; Piliposian, T.G.; Tõnno, I. Use of Sedimentary Algal Pigment Analyses to Infer Past Lake-Water Total Phosphorus Concentrations. *J. Paleolimnol.* **2022**, *68*, 415–426. [CrossRef]

9. Kakade, A.; Salama, E.S.; Han, H.; Zheng, Y.; Kulshrestha, S.; Jalalah, M.; Harraz, F.A.; Alsareii, S.A.; Li, X. World Eutrophic Pollution of Lake and River: Biotreatment Potential and Future Perspectives. *Environ. Technol. Innov.* **2021**, *23*, 101604. [CrossRef]

10. Pan, T.; Cui, C.; Qin, B.; Ding, K.; Zhou, J. Climate Change Intensifies Algal Biomass Resurgence in Eutrophic Lake Taihu, China. *Sci. Total Environ.* **2024**, *926*, 171934. [CrossRef]

11. Sukenik, A.; Kaplan, A. Cyanobacterial Harmful Algal Blooms in Aquatic Ecosystems: A Comprehensive Outlook on Current and Emerging Mitigation and Control Approaches. *Microorganisms* **2021**, *9*, 1472. [CrossRef]

12. Narayanan, M.; Devarayan, K.; Verma, M.; Selvaraj, M.; Ghramh, H.A.; Kandasamy, S. Assessing the Ecological Impact of Pesticides/Herbicides on Algal Communities: A Comprehensive Review. *Aquat. Toxicol.* **2024**, *268*, 106851. [CrossRef]

13. Ward, N.K.; Steele, B.G.; Weathers, K.C.; Cottingham, K.L.; Ewing, H.A.; Hanson, P.C.; Carey, C.C. Differential Responses of Maximum Versus Median Chlorophyll-a to Air Temperature and Nutrient Loads in an Oligotrophic Lake Over 31 Years. *Water Resour. Res.* **2020**, *56*, e2020WR027296. [CrossRef]

14. Rotta, L.; Alcântara, E.; Park, E.; Bernardo, N.; Watanabe, F. A Single Semi-Analytical Algorithm to Retrieve Chlorophyll-a Concentration in Oligo-to-Hypereutrophic Waters of a Tropical Reservoir Cascade. *Ecol. Indic.* **2021**, *120*, 106913. [CrossRef]

15. Chegoonian, A.M.; Zolfaghari, K.; Leavitt, P.R.; Baulch, H.M.; Duguay, C.R. Improvement of Field Fluorometry Estimates of Chlorophyll a Concentration in a Cyanobacteria-Rich Eutrophic Lake. *Limnol. Ocean. Methods* **2022**, *20*, 193–209. [CrossRef]

16. Rogora, M.; Austoni, M.; Caroni, R.; Giacomotti, P.; Kamburska, L.; Marchetto, A.; Mosello, R.; Orrù, A.; Tartari, G.; Dresti, C. Temporal Changes in Nutrients in a Deep Oligomictic Lake: The Role of External Loads versus Climate Change. *J. Limnol.* **2021**, *80*, 2051. [CrossRef]

17. Taipale, S.J.; Ventelä, A.M.; Litmanen, J.; Anttila, L. Poor Nutritional Quality of Primary Producers and Zooplankton Driven by Eutrophication Is Mitigated at Upper Trophic Levels. *Ecol. Evol.* **2022**, *12*, e8687. [CrossRef]

18. Zhao, L.; Zhu, R.; Zhou, Q.; Jeppesen, E.; Yang, K. Trophic Status and Lake Depth Play Important Roles in Determining the Nutrient-Chlorophyll a Relationship: Evidence from Thousands of Lakes Globally. *Water Res.* **2023**, *242*, 120182. [CrossRef]

19. Chatterjee, S.; More, M. Cyanobacterial Harmful Algal Bloom Toxin Microcystin and Increased Vibrio Occurrence as Climate-Change-Induced Biological Co-Stressors: Exposure and Disease Outcomes via Their Interaction with Gut–Liver–Brain Axis. *Toxins* **2023**, *15*, 289. [CrossRef] [PubMed]

20. Mishra, R.K. The Effect of Eutrophication on Drinking Water. *Br. J. Multidiscip. Adv. Stud.* **2023**, *4*, 7–20. [CrossRef]

21. Muruganandam, M.; Rajamanickam, S.; Sivarethinamohan, S.; Gaddam, M.K.R.; Velusamy, P.; Gomathi, R.; Ravindiran, G.; Gurugubelli, T.R.; Muniasamy, S.K. Impact of Climate Change and Anthropogenic Activities on Aquatic Ecosystem—A Review. *Environ. Res.* **2023**, *238*, 117233.

22. Glibert, P.M. Harmful Algae at the Complex Nexus of Eutrophication and Climate Change. *Harmful Algae* **2020**, *91*, 101583. [CrossRef]

23. Trottet, A.; George, C.; Drillet, G.; Lauro, F.M. Aquaculture in Coastal Urbanized Areas: A Comparative Review of the Challenges Posed by Harmful Algal Blooms. *Crit. Rev. Environ. Sci. Technol.* **2022**, *52*, 2888–2929. [CrossRef]

24. Summers, E.J.; Ryder, J.L. A Critical Review of Operational Strategies for the Management of Harmful Algal Blooms (HABs) in Inland Reservoirs. *J. Environ. Manag.* **2023**, *330*, 117141. [CrossRef]

25. Rodríguez-López, L.; Duran-Llacer, I.; González-Rodríguez, L.; Abarca-del-Rio, R.; Cárdenas, R.; Parra, O.; Martínez-Retureta, R.; Urrutia, R. Spectral Analysis Using LANDSAT Images to Monitor the Chlorophyll-a Concentration in Lake Laja in Chile. *Ecol. Inf.* **2020**, *60*, 101183. [CrossRef]

26. Vadeboncoeur, Y.; Moore, M.V.; Stewart, S.D.; Chandra, S.; Atkins, K.S.; Baron, J.S.; Bouma-Gregson, K.; Brothers, S.; Francoeur, S.N.; Genzoli, L.; et al. Blue Waters, Green Bottoms: Benthic Filamentous Algal Blooms Are an Emerging Threat to Clear Lakes Worldwide. *Bioscience* **2021**, *71*, 1011–1027. [CrossRef]

27. Rashidi, H.; Baulch, H.; Gill, A.; Bharadwaj, L.; Bradford, L. Monitoring, Managing, and Communicating Risk of Harmful Algal Blooms (HABs) in Recreational Resources across Canada. *Environ. Health Insights* **2021**, *15*, 11786302211014401. [CrossRef] [PubMed]

28. Song, B.; Park, K. Detection of Aquatic Plants Using Multispectral UAV Imagery and Vegetation Index. *Remote Sens.* **2020**, *12*, 387. [CrossRef]

29. Qing, S.; Runa, A.; Shun, B.; Zhao, W.; Bao, Y.; Hao, Y. Distinguishing and Mapping of Aquatic Vegetations and Yellow Algae Bloom with Landsat Satellite Data in a Complex Shallow Lake, China during 1986–2018. *Ecol. Indic.* **2020**, *112*, 106073. [CrossRef]

30. Camps-Valls, G.; Campos-Taberner, M.; Moreno-Martínez, Á.; Walther, S.; Duveiller, G.; Cescatti, A.; Mahecha, M.D.; Muñoz-Marí, J.; García-Haro, F.J.; Guanter, L.; et al. A Unified Vegetation Index for Quantifying the Terrestrial Biosphere. *Sci. Adv.* **2021**, *7*, eabc7447. [CrossRef]

31. Liang, S.; Gong, Z.; Wang, Y.; Zhao, J.; Zhao, W. Accurate Monitoring of Submerged Aquatic Vegetation in a Macrophytic Lake Using Time-Series Sentinel-2 Images. *Remote Sens.* **2022**, *14*, 640. [CrossRef]

32. Gao, H.; Li, R.; Shen, Q.; Yao, Y.; Shao, Y.; Zhou, Y.; Li, W.; Li, J.; Zhang, Y.; Liu, M. Deep-Learning-Based Automatic Extraction of Aquatic Vegetation from Sentinel-2 Images—A Case Study of Lake Honghu. *Remote Sens.* **2024**, *16*, 867. [CrossRef]

33. Tompoulidou, M.; Karadimou, E.; Apostolakis, A.; Tsiaoussi, V. A Geographic Object-Based Image Approach Based on the Sentinel-2 Multispectral Instrument for Lake Aquatic Vegetation Mapping: A Complementary Tool to In Situ Monitoring. *Remote Sens.* **2024**, *16*, 916. [CrossRef]

34. Ustin, S.L.; Middleton, E.M. Current and Near-Term Advances in Earth Observation for Ecological Applications. *Ecol. Process.* **2021**, *10*, 1. [CrossRef] [PubMed]

35. Yin, G.; Verger, A.; Descals, A.; Filella, I.; Peñuelas, J. A Broadband Green-Red Vegetation Index for Monitoring Gross Primary Production Phenology. *J. Remote Sens.* **2022**, *2022*, 9764982. [CrossRef]

36. Rodríguez-López, L.; Duran-Llacer, I.; Bravo Alvarez, L.; Lami, A.; Urrutia, R. Recovery of Water Quality and Detection of Algal Blooms in Lake Villarrica through Landsat Satellite Images and Monitoring Data. *Remote Sens.* **2023**, *15*, 1929. [CrossRef]

37. Chen, J.; Chen, S.; Fu, R.; Li, D.; Jiang, H.; Wang, C.; Peng, Y.; Jia, K.; Hicks, B.J. Remote Sensing Big Data for Water Environment Monitoring: Current Status, Challenges, and Future Prospects. *Earths Future* **2022**, *10*, e2021EF002289. [CrossRef]

38. Singh, J.K.; Kumar, P.; Vishwakarma, S. Multivariate and Statistical Evaluation of Coastal Water Quality and Seasonal Variation in the Physicochemical Properties of Gulf of Khambhat Region, Gujarat, India. *Water Air Soil Pollut.* **2022**, *233*, 358. [CrossRef]

39. Yépez, S.; Velásquez, G.; Torres, D.; Saavedra-Passache, R.; Pincheira, M.; Cid, H.; Rodríguez-López, L.; Contreras, A.; Frappart, F.; Cristóbal, J.; et al. Spatiotemporal Variations in Biophysical Water Quality Parameters: An Integrated In Situ and Remote Sensing Analysis of an Urban Lake in Chile. *Remote Sens.* **2024**, *16*, 427. [CrossRef]

40. Nimptsch, J.; Woelfl, S.; Osorio, S.; Valenzuela, J.; Moreira, C.; Ramos, V.; Castelo-Branco, R.; Leão, P.N.; Vasconcelos, V. First Record of Toxins Associated with Cyanobacterial Blooms in Oligotrophic North Patagonian Lakes of Chile—A Genomic Approach. *Int. Rev. Hydrobiol.* **2016**, *101*, 57–68. [CrossRef]

41. Huovinen, P.; Ramírez, J.; Caputo, L.; Gómez, I. Mapping of Spatial and Temporal Variation of Water Characteristics through Satellite Remote Sensing in Lake Panguipulli, Chile. *Sci. Total Environ.* **2019**, *679*, 196–208. [CrossRef] [PubMed]

42. Vo, T.M.; Tran, V.T.N.; Cuu, T.N.T.; Do, T.T.H.; Le, T.M. Domestic Violence and Its Association with Preterm or Low Birthweight Delivery in Vietnam. *Int. J. Womens Health* **2019**, *11*, 501–510. [CrossRef]

43. Weisberg, S. *Applied Linear Regression*; Wiley: Hoboken, NJ, USA, 2005.

44. Liu, F.T.; Ting, K.M.; Zhou, Z.H. Isolation Forest. In Proceedings of the IEEE International Conference on Data Mining, ICDM, Pisa, Italy, 15–19 December 2008; pp. 413–422.

45. Zhang, S. Nearest Neighbor Selection for Iteratively KNN Imputation. *J. Syst. Softw.* **2012**, *85*, 2541–2552. [CrossRef]

46. Yeo, I.; Johnson, R.A. A New Family of Power Transformations to Improve Normality or Symmetry. *Biometrika* **2000**, *87*, 954–959. [CrossRef]

47. Rodríguez-López, L.; Usta, D.B.; Duran-Llacer, I.; Alvarez, L.B.; Yépez, S.; Bourrel, L.; Frappart, F.; Urrutia, R. Estimation of Water Quality Parameters through a Combination of Deep Learning and Remote Sensing Techniques in a Lake in Southern Chile. *Remote Sens.* **2023**, *15*, 4157. [CrossRef]

48. Begliomini, F.N.; Barbosa, C.C.F.; Martins, V.S.; Novo, E.M.L.M.; Paulino, R.S.; Maciel, D.A.; Lima, T.M.A.; O'Shea, R.E.; Pahlevan, N.; Lamparelli, M.C. Machine Learning for Cyanobacteria Mapping on Tropical Urban Reservoirs Using PRISMA Hyperspectral Data. *ISPRS J. Photogramm. Remote Sens.* **2023**, *204*, 378–396. [CrossRef]

49. Mbachu, H.I.; Nduka, E.C.; Nja, M.E. Designing a Pseudo R-Squared Goodness-of-Fit Measure in Generalized Linear Models. *J. Math. Res.* **2012**, *4*, 148. [CrossRef]