water

MDPI

Article

Machine Learning Algorithms for the Estimation of Water Quality Parameters in Lake Llanquihue in Southern Chile

Lien Rodríguez-López [1,*], David Bustos Usta [2], Lisandra Bravo Alvarez [3], Iongel Duran-Llacer [4], Andrea Lami [5], Rebeca Martínez-Retureta [6] and Roberto Urrutia [6]

[1] Facultad de Ingeniería, Arquitectura y Diseño, Universidad San Sebastián, Lientur 1457, Concepción 4030000, Chile
[2] Facultad de Oceanografía, Univesidad 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] Hémera Centro de Observación de la Tierra, Facultad de Ciencias, Ingeniería y Tecnología, Universidad Mayor, Camino La Pirámide 5750, Huechuraba, Santiago 8580745, Chile; iongel.duran@umayor.cl
[5] Institute of Water Research IRSA, Sezione di Verbania, 1000015 Verbania, CP, Italy; andrea.lami@cnr.it
[6] Facultad de Ciencias Ambientales, Universidad de Concepción, Concepción 4030000, Chile; rebecmartinez@udec.cl (R.M.-R.); rurrutia@udec.cl (R.U.)
* Correspondence: lien.rodriguez@uss.cl; Tel.: +56-999-168-115

Abstract: The world's water ecosystems have been affected by various human activities. Artificial intelligence techniques, especially machine learning, have become an important tool for predicting the water quality of inland aquatic ecosystems. As an excellent biological indicator, chlorophyll-a was studied to determine the state of water quality in Lake Llanquihue, located in southern Chile. A 31-year time series (1989 to 2020) of data collected in situ was used to determine the evolution of limnological parameters at eight spaced stations covering all of the main points of the lake, and the year, month, day, and hour time intervals were selected. Using machine learning techniques, out of eight estimation algorithms that were applied with real data to estimate chlorophyll-a, three models showed better performance (XGBoost, LightGBM, and AdaBoost). The results for the best models show excellent performance, with a coefficient of determination between 0.81 and 0.99, a root-mean-square error of between 0.03 ug/L and 0.46 ug/L, and a mean bias error of between 0.01 and 0.27 ug/L. These models are scalable and applicable to other lake systems of interest that present similar conditions and can support decision making related to water resources.

Keywords: machine learning algorithms; chlorophyll-a; lake

Citation: Rodríguez-López, L.; Bustos Usta, D.; Bravo Alvarez, L.; Duran-Llacer, I.; Lami, A.; Martínez-Retureta, R.; Urrutia, R. Machine Learning Algorithms for the Estimation of Water Quality Parameters in Lake Llanquihue in Southern Chile. *Water* **2023**, *15*, 1994. https://doi.org/10.3390/w15111994

Academic Editor: Il-Moon Chung

Received: 14 April 2023
Revised: 16 May 2023
Accepted: 20 May 2023
Published: 24 May 2023

1. Introduction

Most of the world's inland aquatic ecosystems have been affected in some way by various human activities [1,2]. Population growth and the exponentially increasing use of freshwater are indispensable components of any analysis of inland aquatic systems [3–5]. Among the main impacts of anthropogenic activity on the biodiversity of inland aquatic ecosystems, including those in Chile, are habitat loss and degradation [6], caused mainly by transformations by agriculture [7–9], afforestation [10], and direct transformations of aquatic systems caused by various types of civil projects, such as irrigation [8,11,12], hydroelectricity [13], tourism [14,15], transportation infrastructure, land-use changes [9,16], the introduction of invasive and exotic species [17,18], the consequences of cascading processes of native species losses [19], and point and diffuse chemical pollution (from industrial projects and livestock activity, among other sources) [20]. The characterization of lakes in southern Chile has focused mainly on their trophic states. One of the main objectives of limnological research is to help preserve trophic states. All efforts are aimed at minimizing, as much as possible, the eutrophication processes that some of the lakes are suffering

recently due to aquaculture activities [21]. The lack of management of production and residential activities in lake drainage basins requires special attention from the scientific community, public services, and local communities. It is necessary to generate basic knowledge and decide on environmental protection actions that will allow the control of the trophic states of lakes and their sustainable use [15,22]. Due to the gradual accumulation of nutrients in aquatic systems, degradation gradually occurs, characterized by increases in algal biomass, loss of biodiversity, algal blooms, and the generation or establishment of hypoxic and/or anoxic conditions in the water column and sediments, which are all consequences of the eutrophication process [23]. Many lakes in the world are suffering degradation due to this phenomenon. In [24,25], studies were carried out using remote sensing of the pollution of the Caspian Sea due to anthropogenic pressures. In [26], the occurrence of algal blooms in the Laurentian Great Lakes, a phenomenon characteristic of eutrophic lakes, is investigated. On the other hand, ref. [27] focuses on Lake Murten in Switzerland and the process of cultural eutrophication that it undergoes. All of the above research shows that large lakes in the world are suffering from effects due to the eutrophication process.

Artificial intelligence techniques, especially machine learning, have been increasingly used in recent environmental research in both oceanic and inland aquatic ecosystems [28]. The prediction of water quality is important for the preparation and regulation of water quality, and different artificial intelligence models can contribute to this purpose [29]. They provide tools for solving supervised, unsupervised, and semi-supervised learning problems in water quality sensing, specifically for lakes [30]. In the case of supervised learning, there are two types of problems: regression and classification. There are different tools for regression, from the simplest, such as basic linear models [31], to the most sophisticated, such as ensemble models or neural networks [32–34]. Different modeling strategies are used because there is no single solution for highly complex problems such as chlorophyll prediction; therefore, having a wide range of available modeling tools can provide greater clarity about which strategy works best. Machine learning has primarily been used in studies on chlorophyll-a [35,36], suspended sediments [37,38], and light attenuation [39,40]. Although there have been numerous investigations that relate deep learning algorithms to water quality parameter detection, the vast majority are focused on lakes in the northern hemisphere, leaving southern lake systems less studied [41–43].

In Chile, the Dirección General de Aguas (DGA), the agency in charge of monitoring the country's lake ecosystems, includes, in its network, only 20 of the country's 175 lakes, indicating an insufficient effort to monitor these ecosystems and the evolution of their trophic states [15]. Thus, the monitoring and study of lakes in Chile pose a challenge, and the need has arisen to implement, for the first time in a Chilean lake, artificial intelligence algorithms as an alternative to the detection of biomass increases using the chlorophyll-a bioindicator. Therefore, the objectives of this research are (i) to describe the behavior of limnological variables in the last thirty years at the monitoring points of Lake Llanquihue, (ii) to predict the chlorophyll-a variable for the studied lake system using machine learning models, and (iii) to describe the main algal groups of Lake Llanquihue for the period (1989–2020).

2. Materials and Methods

2.1. Study Area

Lake Llanquihue is part of an important tourist lake district area called "Lagos Araucanos", which includes 12 glacial lakes between 30°–42° S and 71°–72° W, at altitudes between 117 to 590 m above sea level [44]. Lake Llanquihue is the region's largest lake and the country's second-largest lake [45]. It has a shoreline of 196.5 km and a surface area of 870.5 km^2. Its mean depth has been estimated to be 187 m, and it stores a volume of 158.6 km^3 [46]. The watershed area is very small compared to the lake's surface area, which directly affects the water renewal period by 74 years [47].

2.2. In Situ Monitoring Data

Water quality parameters were selected through monitoring campaigns carried out by the Dirección General de Aguas (DGA) from 1989 to 2020 (31 years). These campaigns are carried out four times a year, in each season, with samples taken at different sites (see Table 1), covering a large spatial area of the lake. The locations were georeferenced in the field considering the following criteria: lake or river morphology, the presence of tributaries (the distance from their influence), the presence of industrial effluents or urban discharges, depth, and accessibility (in the case of tributaries). The in situ parameters selected for this study were Secchi disk depth (SD), chlorophyll-a (Chl-a) (Standard Methods N°10200 H DGALGOCL1/2009), turbidity (NTU) (Standard Methods N°2130 B), total nitrogen (Nt) (Standard Methods N°4500-N C), and total phosphorus (Pt) (Standard Methods N°4500-P E). Field data were collected at eight sampling stations (Ll1-Ll8), (see Figure 1) at depths of 0, 15, and 30 m.

Table 1. Sampling stations at Lake Llanquihue.

N°	COD_BNA	STATION	CODE	Latitude	Longitude	Samples	Train	Test
1	10410006-6	PUERTO OCTAY 1	Ll-1	−40.9765244	−72.8631503	139	111	28
2	10410012-0	PUERTO OCTAY 2	Ll-2	−41.0137713	−72.8482236	20	16	4
3	10410007-4	FRUTILLAR 1	Ll-3	−41.1318026	−72.9892806	135	108	27
4	10410013-9	FRUTILLAR 2	Ll-4	−41.1304389	−72.9482228	30	24	6
5	10410008-2	PUERTO VARAS 1	Ll-5	−41.3115347	−72.9623349	134	107	27
6	10410014-7	PUERTO VARAS 2	Ll-6	−41.2637797	−72.9315548	56	44	12
7	10410009-0	ENSENADA	Ll-7	−41.1962615	−72.593688	178	142	36
8	10410011-2	Z MAX	Ll-8	−41.1009927	−72.6648749	25	20	5

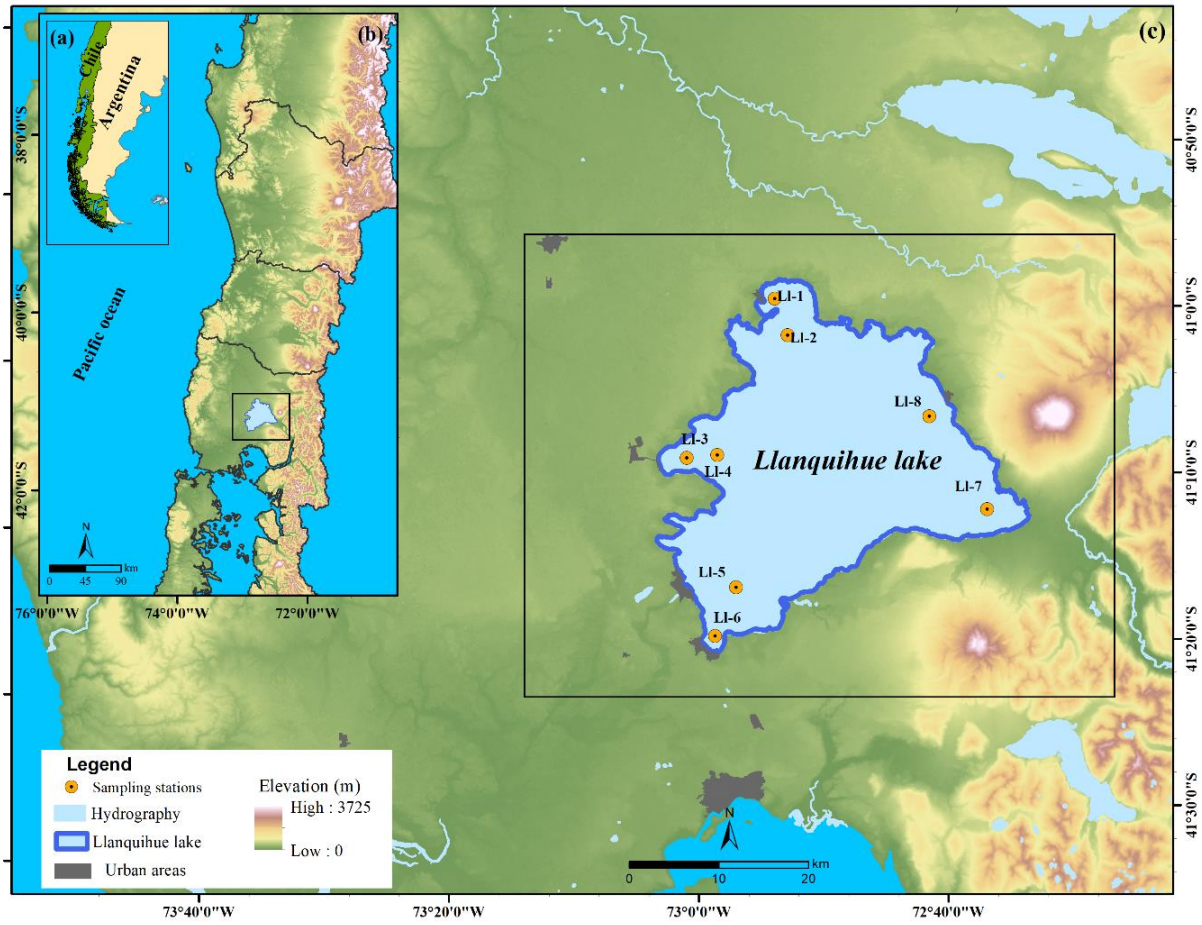

Figure 1. The study area: (**a**) South America, (**b**) South-Central Chile, and (**c**) Lake Llanquihue.

2.3. Data Wrangling and Features Engineering

Rows with more than 80% of features with null values were removed from the database. Therefore, the data wrangling process started with outlier treatment, data integrity analysis, and the imputation of missing values. Several biogeochemical and physical variables including chlorophyll-a (Chl), total nitrogen (N), total phosphorus (P), Silica (si), DQO, dissolved oxygen (O_D), oxygen saturation % (O_D_sat), PH, temperature (Temp), relative humidity (Hum), wind velocity (Wind), conductivity (Conduct), and transparency Secchi depth (Trans) were selected for the analysis.

Additionally, location variables, latitude and longitude, and times in terms of year, month, day, and hour were selected. Finally, dummy variables were created to associate the respective measurement with its sampling station. In total, 26 covariates (independent variables) were used for the prediction of chlorophyll (dependent variable).

Data cleaning was carried out according to each sampling station (Ensenada, Frutillar, Puerto Octay, Puerto Varas, Frutillar 2, Puerto Varas 2, Puerto Octay 2, and Zmax).

The cleaning steps were as follows:

1. Remove non-numerical values from each of the selected variables and replace them with null values.
2. Extract the year, month, and day for each measurement, verifying consistency and integrity.
3. Apply sensible imputation for the null values of each column using a robust central tendency measurement, the median.
4. Split data for training and test validation. In total, for all measurements, the first 80% collected at each sampling station over time were selected for training, and the remaining 20% were used for testing (Table 1).
5. Standardize numerical variables (N, P, Si, DQO, O_D, O_D_sat, PH, Temp, Wind, Hum, Conduct, Trans, and Chl) using the PowerTransformer method, a technique for transforming numerical input or output variables to have a uniform or a Gaussian probability distribution. A power transform will make the probability distribution of a variable more Gaussian [48].

2.4. Machine and Deep Learning Algorithms

This section describes the different modeling methodologies used for chlorophyll prediction. It is important to bear in mind that there is no perfect model, but having different modeling perspectives allows provides a better idea of how feasible learning is for a given task, which is why different models are selected to identify which ones perform better at forecasting chlorophyll values. The analysis covers ensemble methods, including bagging (i.e., random forest) and boosting (XGBoost, AdaBoost, GradientBoosting, and LightGBM) strategies, support vector machines (SVMs), and neural networks (i.e., MLP and ANN). For each algorithm, a brief description is provided.

Random Forest

Random forest is a bagging algorithm introduced by [49] as an adaptation of the algorithm proposed by [50]. The mathematical foundations of random forest were described by Breiman at the end of the 20th century [51], and it has been among the most innovative machine learning techniques. Along with the boosting technique, random forests can be used for either classification (categorical response) or regression problems (continuous response) for supervised learning [49].

The random forest regressor alternative was selected to predict chlorophyll-a using a different number of decision trees in various subsamples from bootstrapped datasets constructed from the original dataset to improve the predictive accuracy by avoiding overfitting [49,51].

The idea in bagging is to reduce variance by constructing many noisy, approximate, unbiased models (i.e., decision trees). Trees are ideal candidates for bagging since they are designed to provide an understanding of complex interaction structures from data, and if they are grown with enough depth, bias can be reduced. The algorithm is described in the following steps:

Random Forest Algorithm

1. For b = 1 to B:
   (a) Draw a bootstrap sample **Z*** of size N from the training data.
   (b) Grow a random forest tree T_b to the bootstrapped data by recursively repeating the following steps for each terminal node of the tree until the minimum node size n_{min} is reached:
      i. Select m variables at random from the p variables,
      ii. Pick the best variable/split point among the m,
      iii. Split the node into two daughter nodes.
2. Output the ensemble of trees $\{T_b\}_1^B$.
3. To predict at a new point x:
   Regression: $\hat{f}_{rf}(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x)$.
   Classification: Let $\hat{C}_b(x)$ be the class prediction of the bth random forest tree. Then $\hat{C}_{rf}(x) = majority\ vote\{\hat{C}_b(x)\}_1^B$.

Multiple parameters were evaluated to identify the best configuration [52], including the maximum depth (15, 20, 25, 30), the number of trees (100, 120, 140, 150), the number of features considered for the best split (square root, log2), and the minimum number of samples needed to split an internal node (2, 3, 4, 5). In addition, default parameters such as the function to measure the quality of a split (squared error) were selected.

AdaBoost

AdaBoost is a machine learning algorithm created from the boosting technique using several weak estimators to reduce bias. The idea was proposed by Freund and Schapire and is one of the most common algorithms with applications in numerous fields [53].

The AdaBoost regressor alternative (Ying et al., 2013) was selected to predict chlorophyll-a. The algorithm is described by the following steps:

AdaBoost Algorithm

1. Initialization: Given training data from the instance space $S = \{(x_1, y_1), \ldots, (x_m, y_m)\}$ where $x_i \epsilon \chi$ and $y_i \epsilon$ Y.
2. Initialize the distribution $D_1(i) = \frac{1}{m}$.
3. For $t = 1, \ldots, T$:
   (a) Train a weak learner $h_t : \chi \to \mathbb{R}$ using the distribution D_t,
   (b) Determine weight α_t of h_t,
   (c) Update the distribution throughout the training set:

$$D_{t+1}(i) = \frac{D_t(i)e^{-\alpha_t y_i h_t(x_i)}}{Z_t}$$

   where Z_t is a normalization factor chosen so that D_{t+1} will be a distribution.
4. Calculate the final score:

$$f(x) = \sum_{t=0}^{T} \alpha_t h_t(x).$$

Several parameters were assessed to identify the best configuration [53,54] including several estimators (120, 140, 160, 180) and learning rates to control overfitting, with a higher learning rate increasing the contribution of each regressor (0.01, 0.1, 0.5) and the loss function to update weights (linear, square, exponential).

Gradient Boosting

The gradient boosting regressor alternative was selected to predict chlorophyll-a. This is an ensemble learning technique constructed from the boosting methodology [55,56]. The algorithm was created with the idea of learning from a functional mapping function defined as y = F(x,B), where B is the set of parameters of F such that some cost function C is minimized [57].

Gradient Boosting Algorithm

1. Define Input: Dataset D, loss function L, base learner $\mathcal{L}_{\Phi}$, number of iterations M, and the learning rate η.
2. Initialize $\hat{f}^{(0)}(x) = \hat{f}_0(x) = \hat{\theta_0} = arg\min_{\theta} \sum_{i=1}^{n} L(y_i, \theta)$.
3. For $m = 1, 2, \ldots, M$:
   (a) $\widehat{g_m}(x_i) = \left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]$,
   (b) $\hat{\theta}_m = arg \min_{\phi \epsilon \Phi, \beta} \sum_{i=1}^{n} [(-\hat{g}_m(x_i)) - \beta\phi(x_i)]^2$,
   (c) $\hat{\rho}_m = arg\min_{\rho} \sum_{i=1}^{n} L(y_i, \hat{f}^{(m-1)}(x_i) + \rho\hat{\phi}_m(x_i))$,
   (d) $\hat{f}_m(x) = \eta\hat{\rho}_m\hat{\phi}_m(x)$,
   (e) $\hat{f}^{(m)}(x) = \hat{f}^{(m-1)}(x) + \hat{f}^{(m)}(x)$.
4. Calculate the output:
$$\hat{f}(x) \equiv \hat{f}^{(M)}(x) = \sum_{m=0}^{M} \hat{f}_m(x).$$

Several parameters were assessed to identify the best configuration [53,54], including learning rate (0.01, 0.1), maximum depth (25, 30, 35), number of trees (70, 100, 120, 140), and the number of features for the best split (square root and log2). Furthermore, other hyperparameters were set, such as the loss function to be optimized (squared error) and the function to measure the quality of a split (friedman_mse).

XGBoost

XGBoost is a scalable machine learning algorithm based on the boosting methodology, which in recent decades has been widely recognized as a very proficient approximation in several machine learning and data mining challenges with some advantages such as training execution time, scalability, error reduction, simplified calculations, and lower computational cost [58] for the prediction of chlorophyll-a.

The XGBoost regression alternative was selected for the regression task. XGBoost minimizes an objective function using regularization (L1 and L2) to penalize unnecessary complexity in the model. The training task is an iterative process, with new trees added and the error reduced using a serial process to improve the final prediction. It is similar to the gradient boosting algorithm since the gradient descent algorithm is implemented to minimize the loss when the complexity increases [58,59]. A description of the algorithm is presented in Figure 2.

For the XGBoost algorithm, four types of parameters were tuned. General parameters are related to the booster type, which is gbtree (gradient boosting). Booster parameters such as maximum depth (15, 20, 25, 30), learning rate (0.01, 0.1), and an L1 regularization term on weight (0, 0.3, 0.5) were considered. Finally, the learning task parameters and the command line parameters were set to default.

LightGBM

A LightGBM regression alternative was selected for predicting chlorophyll-a. This alternative is known as gradient boosting decision tree (GBDT), which is a popular machine learning algorithm such as XGBoost. It is a novel technique to address the multidimensionality problem with high efficiency and scalability using two techniques: gradient-based one-sided sampling (GOSS), which excludes a significant proportion of data instances using information gain criteria and exclusive feature bundling (EFB) as an effective method to reduce the number of features.

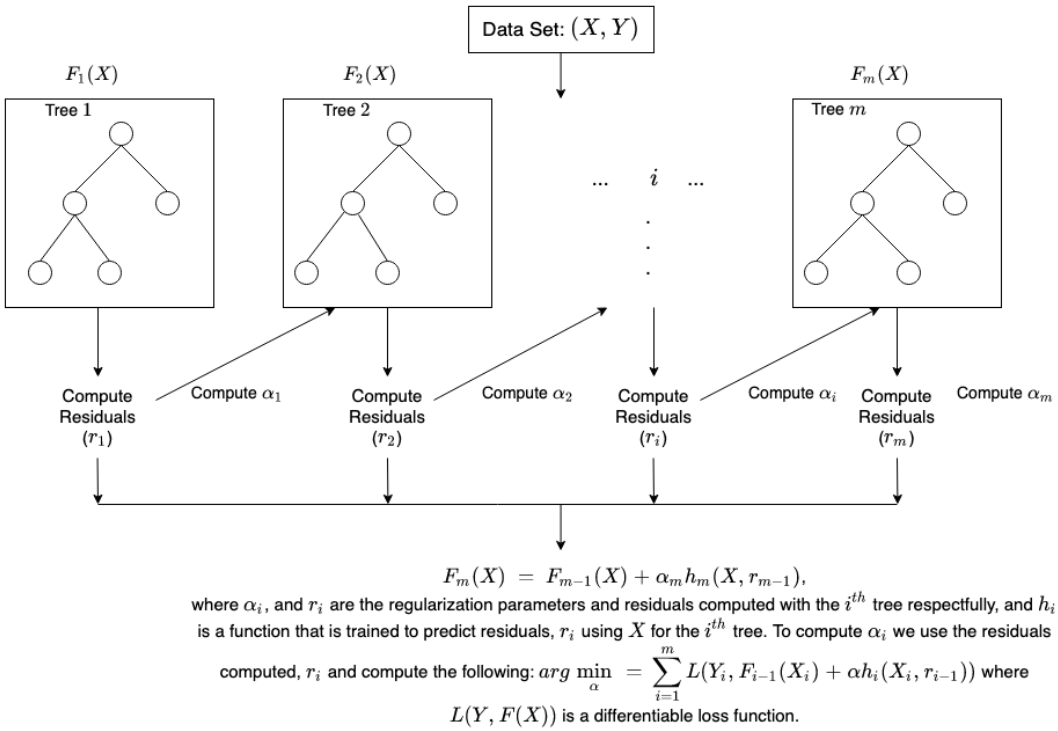

Figure 2. Description of the XGBoost algorithm, adapted from (Chen and Guestrin 2016).

GBDT has been shown to speed up the training process of conventional gradient boosting decision trees by more than 20 times while achieving almost the same accuracy. It can be considered a similar alternative to the XGBoost algorithm [60]. The algorithm is described in the following steps:

LightGBM Algorithm

1. Define Input: Dataset D, loss function L, base learner $\mathcal{L}_\Phi$, number of iterations M, the sampling ratio of large gradient data (a), and the sampling ratio of small gradient data (b).
2. Merge mutually exclusive features using the exclusive feature bundling (EFB) method.
3. Initialize $\hat{f}^{(0)}(x) = \hat{f}_0(x) = \hat{\theta_0} = arg\min_{\theta} \sum_{i=1}^{n} L(y_i, \theta)$.
4. For $m = 1, 2, \ldots, M$
   (a) Compute the absolute values of gradients:
   $$r_i = [\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}],$$
   (b) Resample dataset using Gradient-based One-side Sampling (GOSS) method,
   (c) Compute the information gains,
   (d) Get a new decision tree $\theta_m(X) = \theta_{m-1}(X) + \theta_m(X)'$.
5. Return: $\theta_m(X)$.

Several parameters were assessed to identify the best configuration [60], including boosting type (gbdt = traditional gradient boosting decision tree, dart = dropouts meet multiple additive regression trees, goss = gradient-based one-sided sampling), learning rate (0.01, 0.1), maximum depth (20, 30, 35), and several estimators (70, 100, 120, 140). Support vector machine (SVM), the support vector regression (SVR) algorithm, was developed by [61]. SVR finds a function that estimates the difference between the input and output variable [62] using the following equation:

$$S_i = s(Chl_i) = \sum_{i=1}^{T} w_i \Phi(Z_i) + b \quad (1)$$

where S_i is the network output, Z_i is the input data, which is diagramed into a higher-dimensional feature using a nonlinear mapping function $\Phi(Z_i)$, and w_i and b are coefficients determined by minimizing the regularized risk function based on the network output and real value [63].

We evaluated several kernel functions to select the optimum performance, including the linear function given by <x,x^'>, the polynomial one (degrees 3 and 4), which is represented by the similarity of the vectors in the training dataset in a feature space over polynomials of the original variables involved in the kernel defined by (γ<x, x^'> + r)^d, where d denotes the degree and r a constant, and RBF (radial basis function), which adds a radial basis method to improve the transformation given by exp(⟦−γ|(|x − x^′ |)| ⟧ ^2), where γ is a parameter defined by the algorithm that sets the "sparsity" of the kernel, usually under a scaling pattern.

Multilayer Perceptron (MLP)

The multilayer perceptron (MLP) is an artificial neural network created by Frank Rosenblatt in 1957 that generates a set of outputs from the inputs using multiple hidden layers of connected nodes as a directed graph between input and output layers using the backpropagation algorithm for training.

A few combinations of hidden layers (32,16,8), (32,16), and (16,8,4) were considered. Furthermore, we evaluated several activation functions for the hidden layers (tanh and relu), different types of solvers (sgd and Adam) for weight optimization, an L2 regularization term (0.001, 0.05), and two different learning rates (constant and adaptive).

Artificial Neural Network (ANN)

This is a deep learning algorithm for classification and regression tasks from multiple inputs with the ability to handle complex environmental interactions between variables [64–66]. ANNs are usually designed with more than one input layer, several hidden layers, and an output layer [67]. The general formula is given by: $Y = f(X, W) + \epsilon$, where Y is the vector of model outputs, X is the vector of inputs, W represents the weights, and the function f represents the relationship between outputs, inputs, and parameters of the model [68]. In this case, the relu activation function was selected, and two hidden layers, the first with 32 neurons and the second with 16 neurons, define the geometric configuration to avoid bottlenecks in the learning task. The Adam optimizer and 100 epochs were selected for the training process.

2.5. K-Fold Cross-Validation

After creating the models described in Section 2.3, we had to determine the performance of each model. To this end, we applied the k-fold cross-validation (CV) method described by [69], which is considered a statistical method to evaluate and compare algorithm performance by dividing data into two segments: one used for learning and the other to validate results without overlapping data instances between groups. We used the most applied alternative, known as k-fold, using cross-validation with k = 5 folds [70,71].

2.6. Hyperparameter Tuning

There are different alternatives to determine the best hyperparameter configuration in each model including GridSearchCV and RandomizedSearchCV as defined by [72]. However, for this study, the GridSearchCV alternative was selected because it is more exhaustive and provides good results when the hyperparameter space is correctly defined.

2.7. Performance Metrics

2.7.1. Mean Absolute Error

Mean absolute error (MAE) is defined as the average of the difference between the observed and predicted values. The lower the MAE, the better the model [63,73]. The MAE can be calculated using the following equation:

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \quad (2)$$

2.7.2. Root-Mean-Square Error

The root-mean-square error (RMSE) is often used as a measure of statistical error. The lower its value, the better the model [63]. The RMSE is defined by the following equation:

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \quad (3)$$

2.7.3. Coefficient of Determination

The coefficient of determination is defined as the proportion of the variation in the dependent variable explained by the independent variables. It represents the squared correlation between the observed and predicted values [73]. The higher the R^2, the better the model. It is defined by the following equation:

$$R^2 = 1 - \frac{SSR}{SST} \quad (4)$$

where SSR is defined as the residual sum of squares and SST is the total sum of squares, given by:

$$\text{SSR} = \sum_{i=1}^{n}(y_i - f(x_i))^2 \quad (5)$$

$$\text{SST} = \sum_{i=1}^{n}(y_i - \bar{y})^2 \quad (6)$$

The SSR is defined as the sum of deviations and is a measure of discrepancy between the data and the estimations from the model. A small SSR indicates worse performance. In addition, the SST represents the total error concerning the mean value of the target variable.

2.8. Collection and Treatment of Samples for the Identification of Algal Groups

La Dirección General de Aguas carried out phytoplankton monitoring in the field for which quantitative phytoplankton samples were collected in 500 mL Van Dorn bottles at different depths from the surface in some of the system's effluents and tributaries at the Lake Llanquihue stations [72]. Qualitative sampling was performed, which consisted of a 50-micron phytoplankton net drag at each sampling station in the lake. Each sample was stored in 500 mL plastic bottles, duly labeled, and kept fresh at 4 °C, then, preserved with 1% Lugol's solution until observation. A total of 25 samples were taken.

3. Results

3.1. Water Quality Parameters Summary

Figure 3 shows the in situ limnological parameters chosen for this study for which a time series from 1989 to 2020 was used.

The Chl-a values ranged between maximum values in winter of 1.60 ug/L at the Ll-5 sampling point and 0.99 ug/L at the Ll-2 station, having an average of 1.23 ug/L and 0.486 ug/L, respectively. However, Chl-a presents high values for all sampling stations and seasons of the year. The turbidity variable presents its highest values in autumn and winter, an expected situation due to the higher precipitation during these periods, with values between 7.5–5.4 NTU (stations Ll-1 and Ll-3). As for the nutrients nitrogen and phosphorus, which have been reported in previous studies as limiting the productivity of Chl-a in Lake Llanquihue [47], their values are highest in summer and spring, which coincide with the high Chl-a values attributed to the increase in productivity in these seasons. Total phosphorus reached a maximum value of 35 mg/L during spring at station Ll-4, and total nitrogen reached a maximum of 0.200 mg/L in summer at Ll-5. The Secchi disk measurements, meanwhile, as a parameter indicative of transparency, reflect the high transparency of Lake Llanquihue in its oligotrophic state, despite the impacts and multiple uses in its surrounding basin. The values ranged from an average of 16 m in summer to 8.7 m in winter.

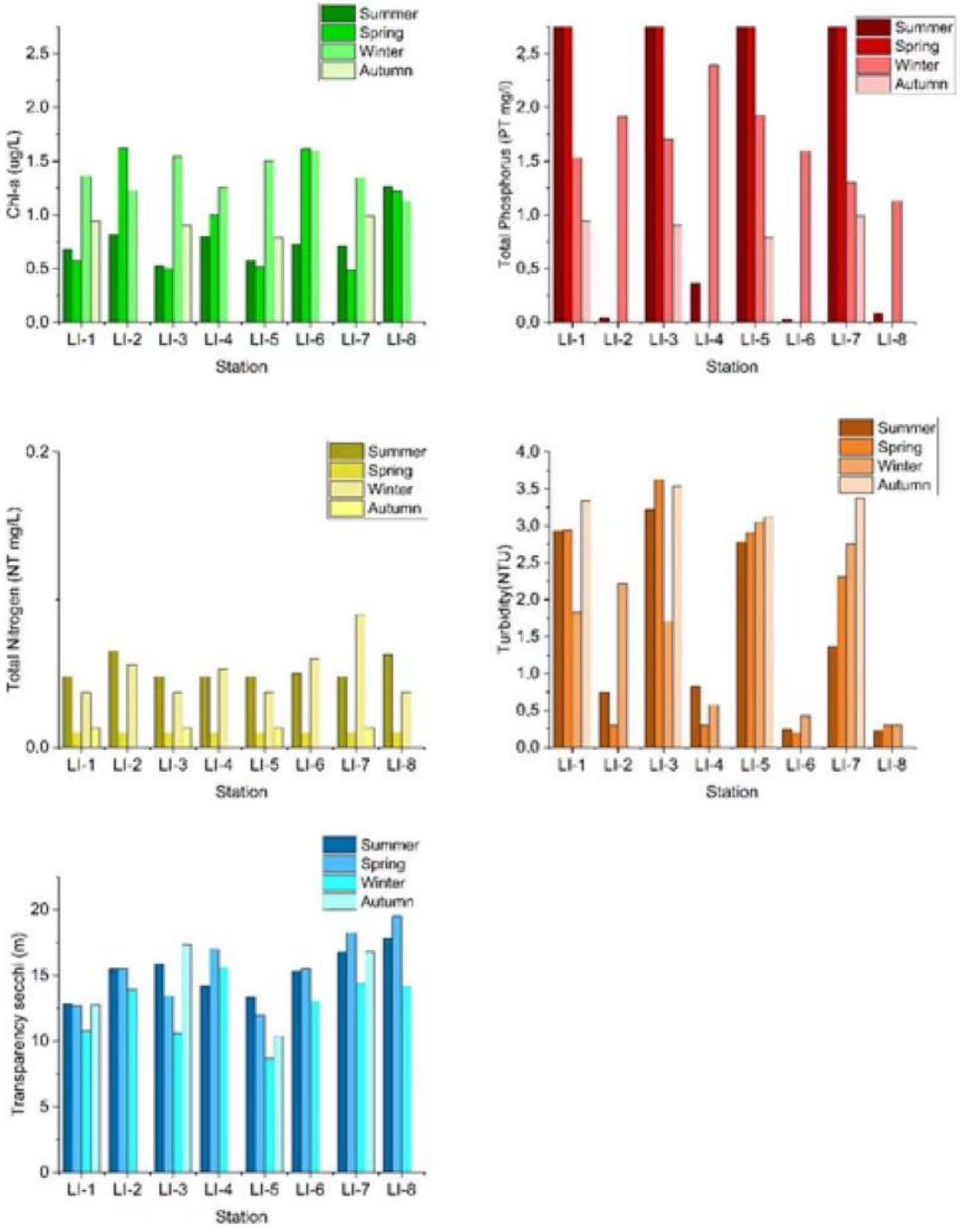

Figure 3. The behavior of limnological parameters (average values).

3.2. Chlorophyll-a Prediction

The best hyperparameters for each model after the cross-validation method was applied are presented below:

Random forest: criterion = squared error, max depth = 25, max features = square root, min_samples_split = 2, and the number of estimators = 150. The results for each sampling station are presented in Figure 4.

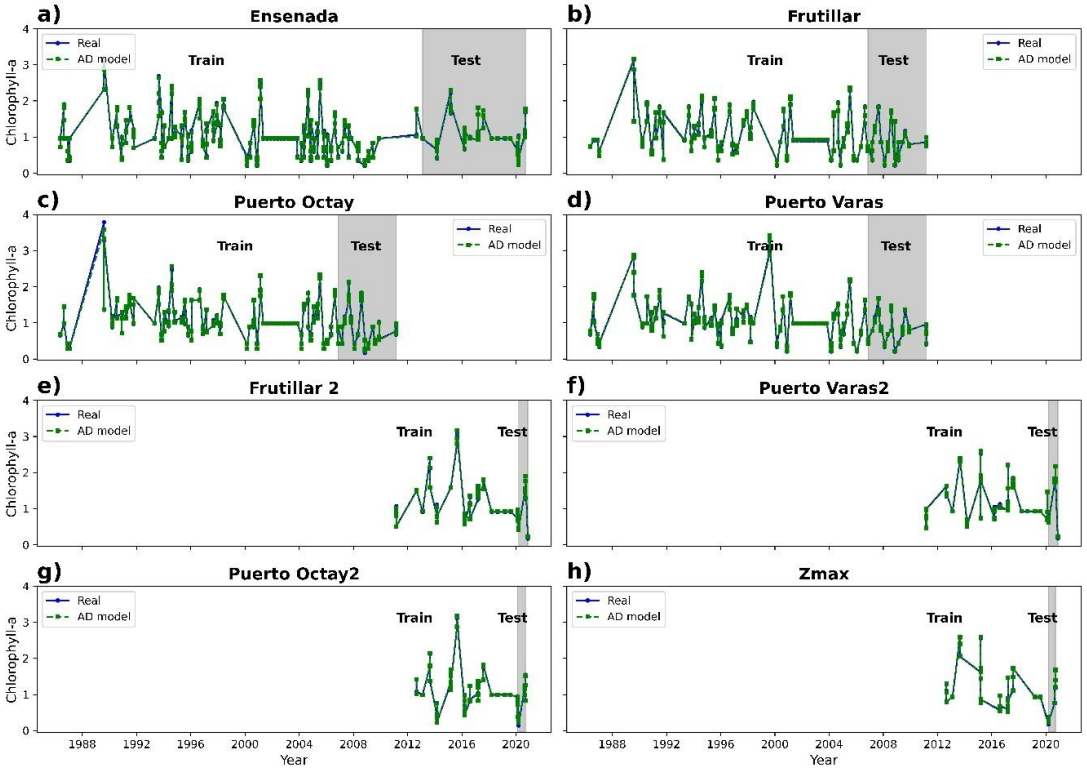

Figure 4. Chlorophyll-a time series plot based on observed data for (**a**) Ensenada, (**b**) Frutillar, (**c**) Puerto Octay, (**d**) Puerto Varas, (**e**) Frutillar 2, (**f**) Puerto Varas 2, (**g**) Puerto Octay 2, and (**h**) Zmax sampling stations using the AdaBoost model (AD model) in green. Real data are represented by the blue lines. Regions with gray backgrounds represent periods for validation purposes.

AdaBoost regressor: number of estimators = 180, learning rate = 0.5, and loss method = square. The results for each sampling station are presented in Figure 5.

XGBoost: alpha = 0.01, learning rate = 0.1, max depth = 15, and the number of estimators = 140. The results for each sampling station are presented in Figure 6.

Gradient boosting: learning rate = 0.1, loss = squared error, max depth =30, max features = sqrt, and the number of estimators = 140. The results for each sampling station are presented in Figure 7.

LightGBM: boosting type = gradient boosting decision trees (gbdt), learning rate = 0.1, max depth = 30, and the number of estimators = 100. The results for each sampling station are presented in Figure 8.

SVM regressor: kernel = linear, gamma = scale method. The results for each sampling station are presented in Figure 9.

MLP regressor: activation function = rectified linear unit (relu), alpha = 0.0001, hidden layer sizes (32, 16), learning rate = constant, and solver = Adam. The results for each sampling station are presented in Figure 10.

Artificial neural network (ANN): The input layer consists of 29 variables; two hidden layers with 32 and 16 neurons and then an output layer was selected. The results for each sampling station (not shown) are similar to those of the MLP regressor.

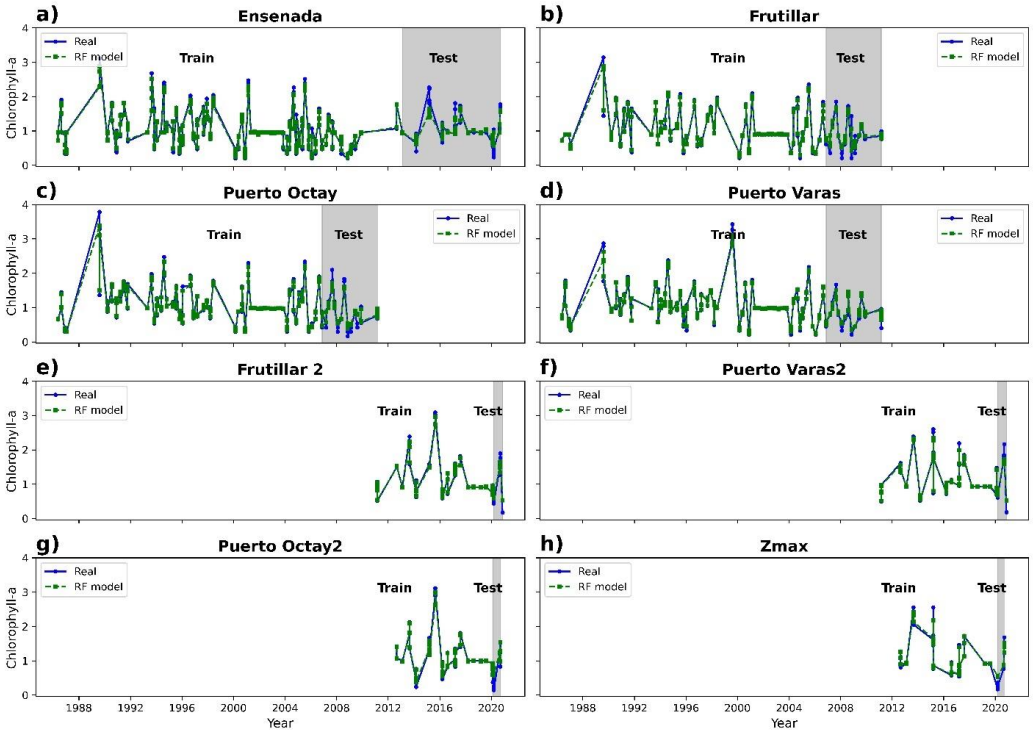

Figure 5. Chlorophyll-a time series plot based on observed data for (**a**) Ensenada, (**b**) Frutillar, (**c**) Puerto Octay, (**d**) Puerto Varas, (**e**) Frutillar 2, (**f**) Puerto Varas 2, (**g**) Puerto Octay 2, and (**h**) Zmax sampling stations using the random forest regressor model (RF model) in green. Real data are represented by the blue lines. Regions with gray backgrounds represent periods for validation purposes.

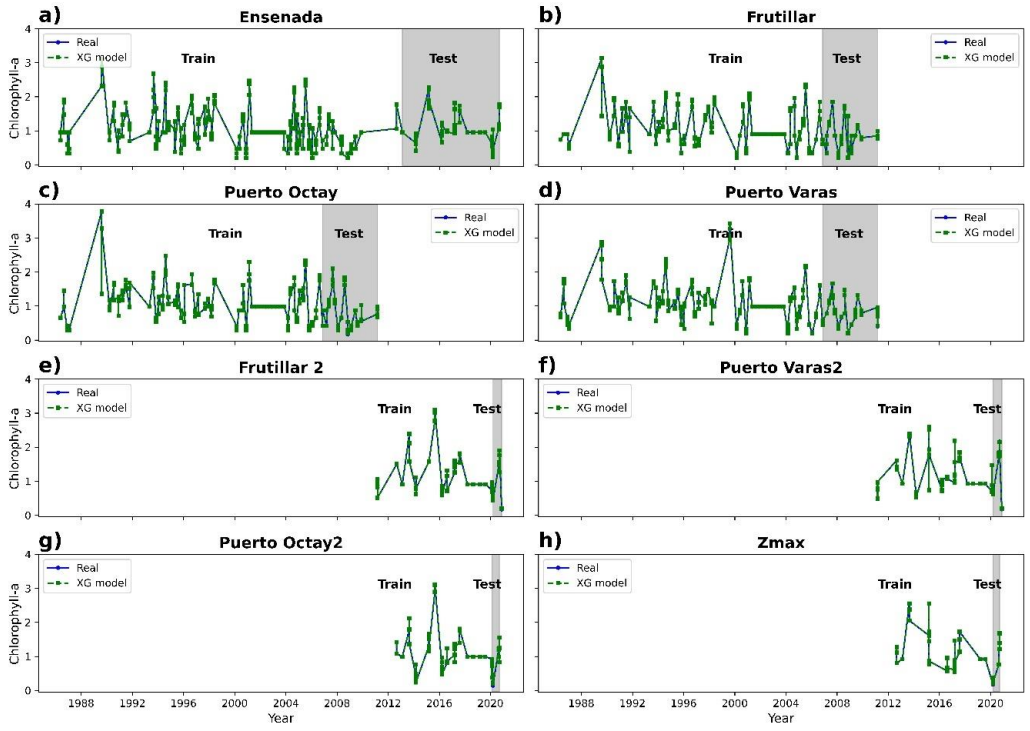

Figure 6. Chlorophyll-a time series plot based on observed data for (**a**) Ensenada, (**b**) Frutillar, (**c**) Puerto Octay, (**d**) Puerto Varas, (**e**) Frutillar 2, (**f**) Puerto Varas 2, (**g**) Puerto Octay 2, and (**h**) Zmax sampling stations using the XGBoost regressor model (XG model) in green. Real data are represented by the blue lines. Regions with gray backgrounds represent periods for validation purposes.

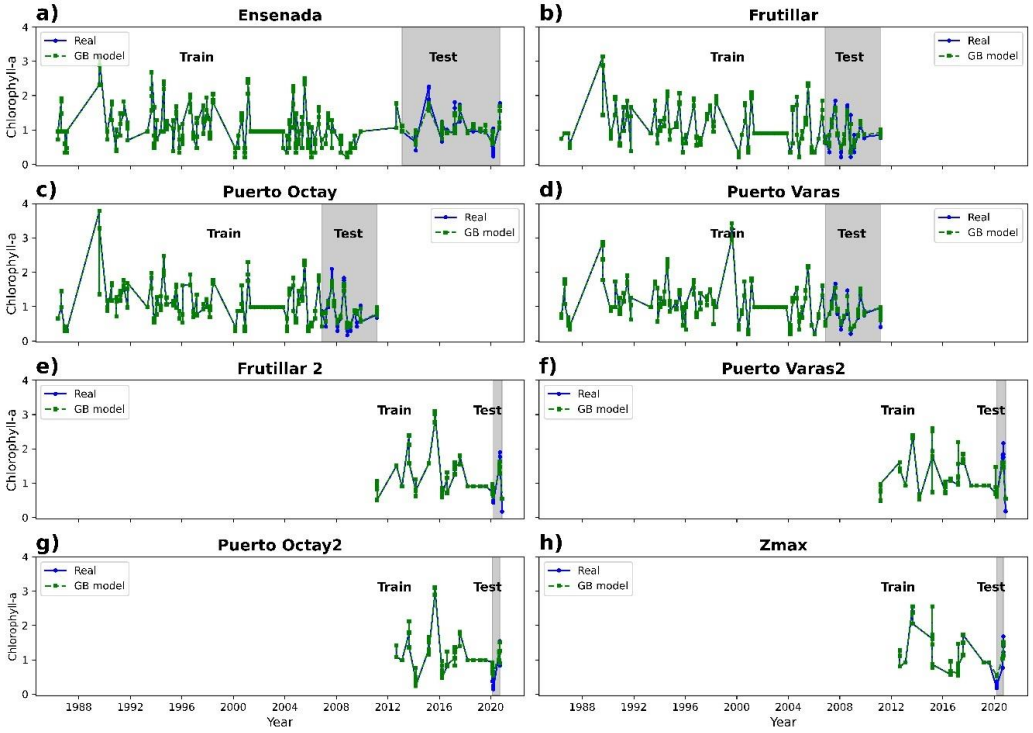

Figure 7. Chlorophyll-a time series plot based on observed data for (**a**) Ensenada, (**b**) Frutillar, (**c**) Puerto Octay, (**d**) Puerto Varas, (**e**) Frutillar 2, (**f**) Puerto Varas 2, (**g**) Puerto Octay 2, and (**h**) Zmax sampling stations using the Gradient boosting model (GB model) in green. Real data are represented by the blue lines. Regions with gray backgrounds represent periods for validation purposes.

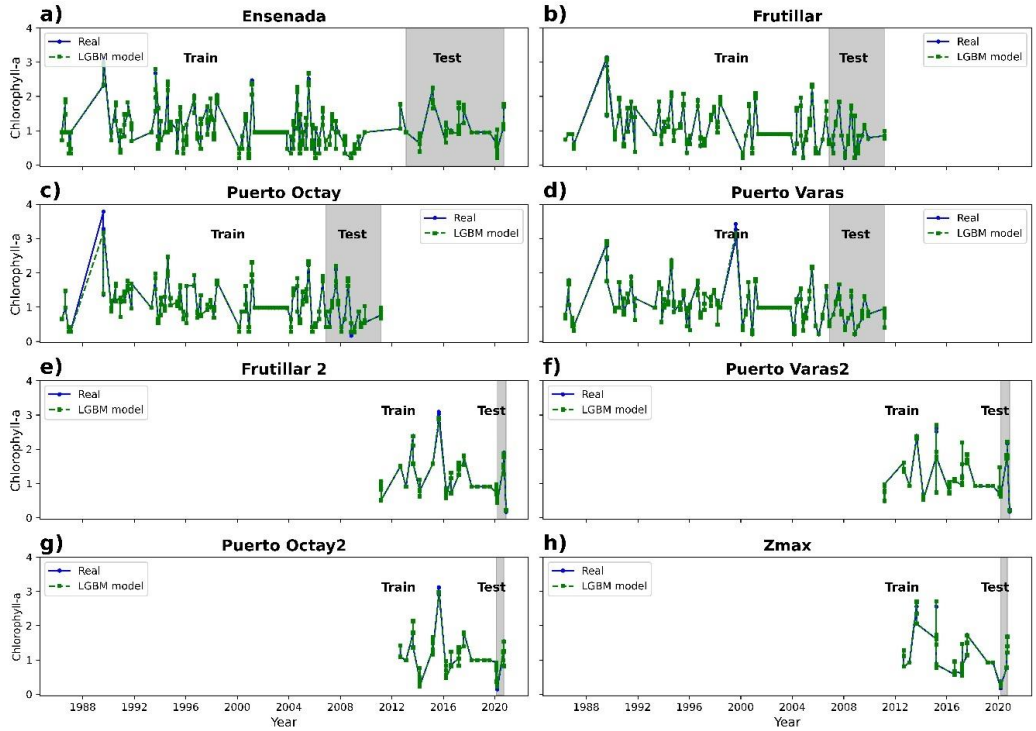

Figure 8. Chlorophyll-a time series plot based on observed data for (**a**) Ensenada, (**b**) Frutillar, (**c**) Puerto Octay, (**d**) Puerto Varas, (**e**) Frutillar 2, (**f**) Puerto Varas 2, (**g**) Puerto Octay 2, and (**h**) Zmax sampling stations using the LightGBM model (LGBM model) in green. Real data are represented by the blue lines. Regions with gray backgrounds represent periods for validation purposes.

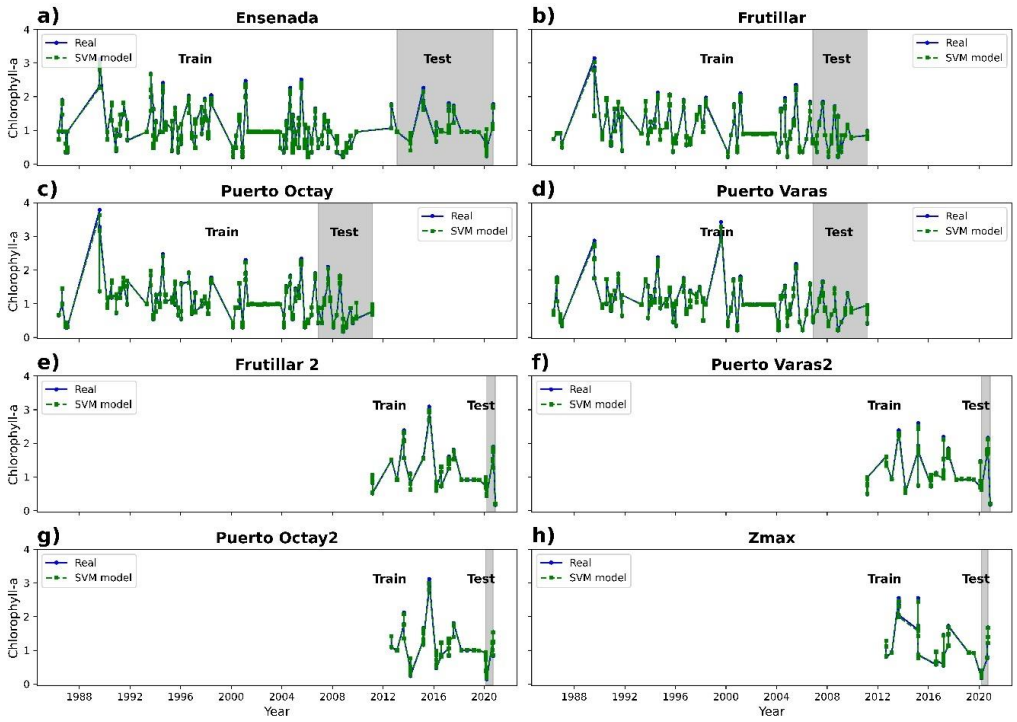

Figure 9. Chlorophyll-a time series plot based on observed data for (**a**) Ensenada, (**b**) Frutillar, (**c**) Puerto Octay, (**d**) Puerto Varas, (**e**) Frutillar 2, (**f**) Puerto Varas 2, (**g**) Puerto Octay 2, and (**h**) Zmax sampling stations using the SVM regressor model (SVM model) in green. Real data are represented by the blue lines. Regions with gray backgrounds represent periods for validation purposes.

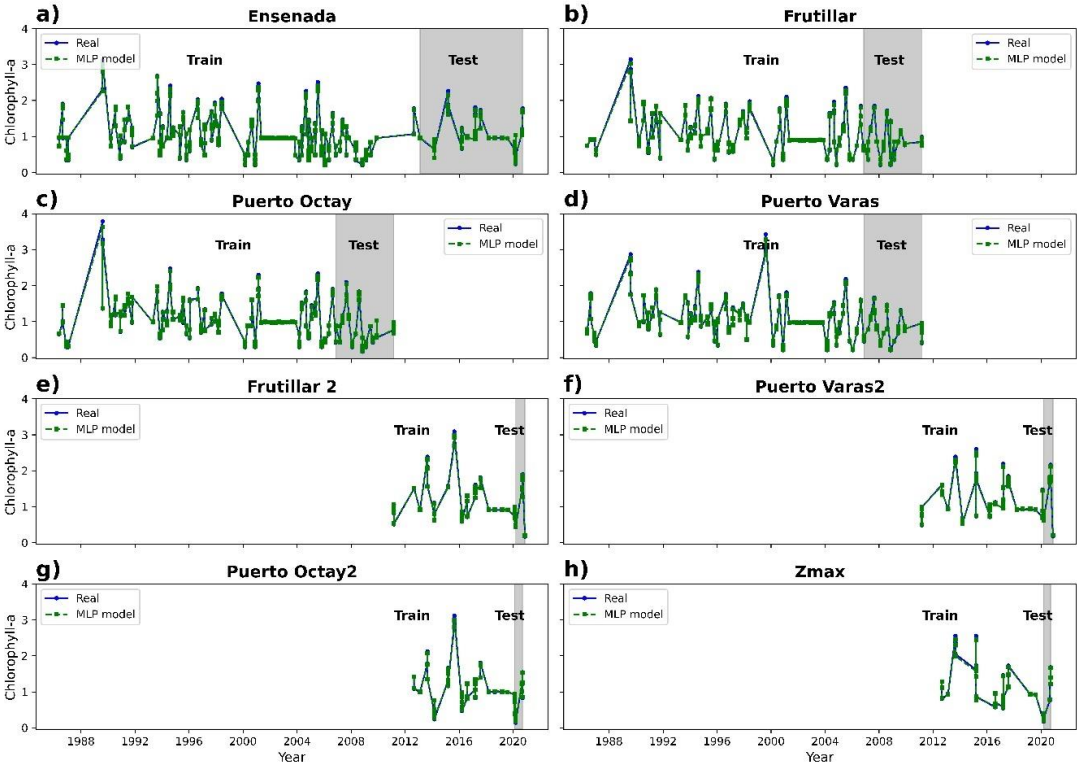

Figure 10. Chlorophyll-a time series plot based on observed data for (**a**) Ensenada, (**b**) Frutillar, (**c**) Puerto Octay, (**d**) Puerto Varas, (**e**) Frutillar 2, (**f**) Puerto Varas 2, (**g**) Puerto Octay 2, and (**h**) Zmax sampling stations using the MLP regressor model (MLP model) in green. Real data are represented by the blue lines. Regions with gray backgrounds represent periods for validation purposes.

3.3. Statistical Analysis

The results show very good performance, with a coefficient of determination between 0.81 and 0.99, a root-mean-square error between 0.03 ug/L and 0.46 ug/L, and a mean absolute error between 0.01 and 0.27 (see Figure S1 and Table 2). The worst performance is observed for the ANN, probably because the number of layers chosen is not complex enough to establish the appropriate relationships between the inputs and the output. Meanwhile, the best-performing models are XGBoost, LightGBM, and AdaBoost, demonstrating that in this case, the boosting technique gives better results compared to the bagging technique (random forest).

Table 2. Model performance metrics for chlorophyll-a prediction based on RF, AD, XG, GB, LGBM, SVM, MLP, and ANN.

Model	R^2	RMSE (ug/L)	MAE
Random forest	0.81	0.46	0.14
AdaBoost regressor	0.99	0.07	0.03
XGBoost regressor	0.99	0.03	0.01
Gradient boosting	0.81	0.46	0.16
LightGBM	0.99	0.06	0.01
SVM regressor	0.99	0.05	0.03
MLP regressor	0.97	0.19	0.10
ANN	0.85	0.41	0.27

When evaluating machine learning models, it is important to analyze the tradeoff between bias and variance to avoid overfitting and underfitting (see Figure 11). A lack of high variance in each of the algorithms is guaranteed as cross-validation strategies (K-fold) and hypertuning methodologies (GridSearCV) have been implemented to find optimal, robust, and scalable solutions, ensuring equilibrium points in the predictions, with low bias and variance.

3.4. Specific Composition and Relative Abundance

The phytoplankton community in Lake Llanquihue is composed of 12 classes, 36 genera, and 52 species, including 30 diatoms (22 *Bacillariophyceae*, 5 *Coscinodiscophyceae* and 3 *Mediophyceae*), 10 green algae (5 *Chlorophyceae*, 2 *Klebsormidiophyceae*, 2 *Trebouxiophyceae*, and *1 Conjugatophyceae*), 5 *dinoflagellates*, 3 *cyanobacteria*, 2 *Cryptophyceae*, 1 *Chrysophyceae*, and 1 *Xanthophyceae*. Centric diatoms of the class *Coscinodiscophyceae* accounted for 72% of the abundance of taxa in the community (Figure 12), mainly the diatom species *Aulacoseira granulata*, *Aulacoseira distans*, and *Fragilaria crotonensis*, as well as the *Chlorophyceae Westella botryoides* [72].

The phytoplankton community presented similar taxa richness (52 in 2018), with the diatoms *Alaucoseira granulata* and *Alaucoseira distans* predominating. It should be noted that the chrysophycean D. divergens decreased in abundance in 2018 and the chlorophycean W. botryoides increased. That same year, the dinoflagellate *Ceratium hirundinella* was recorded in 12 of the 25 analyzed samples, with a maximum abundance of 1570 cel/L at station Ll-3 during the summer at a depth of 30 m. This dinoflagellate was also reported in 2017 in similar abundances [72].

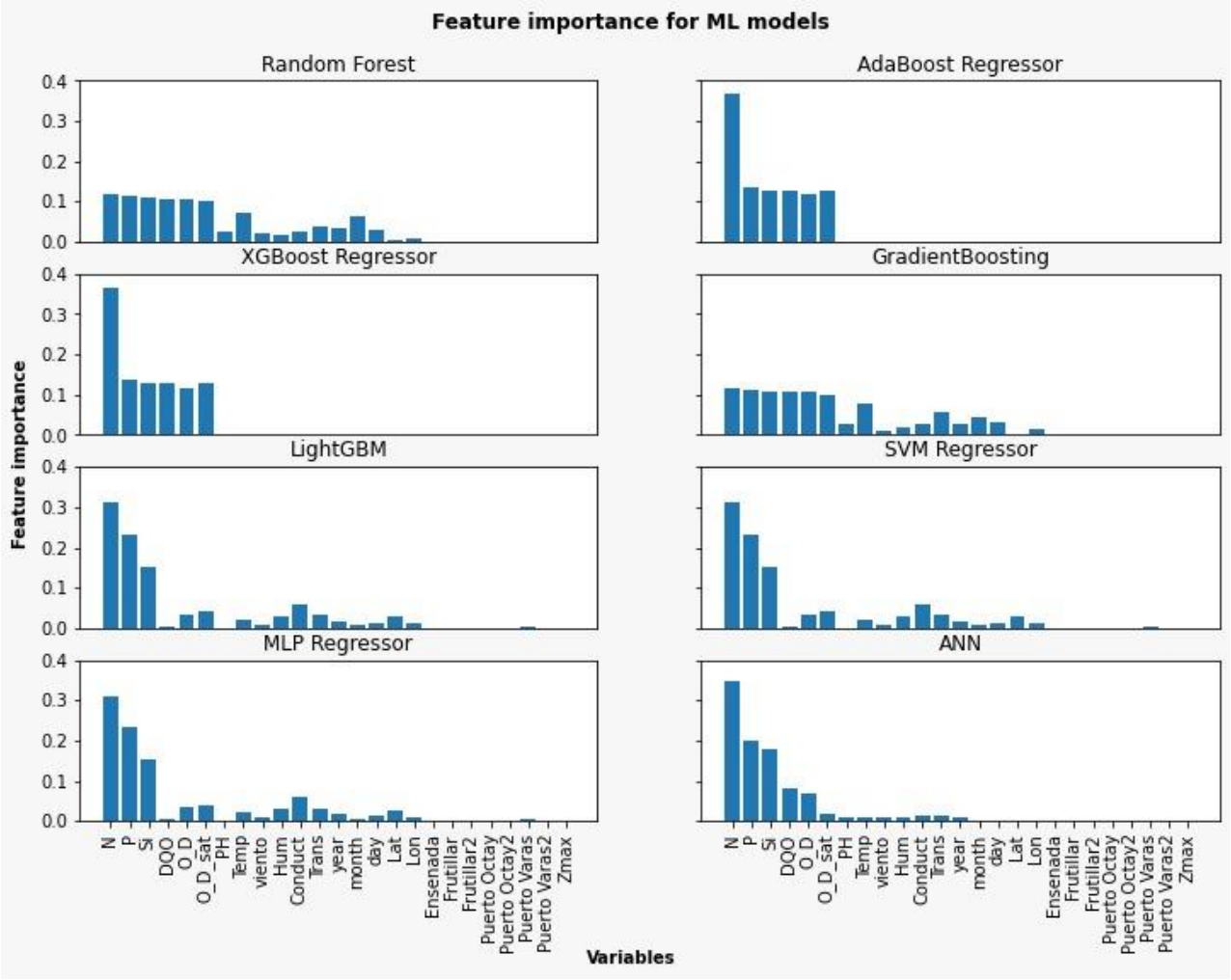

Figure 11. Feature importance in linear models.

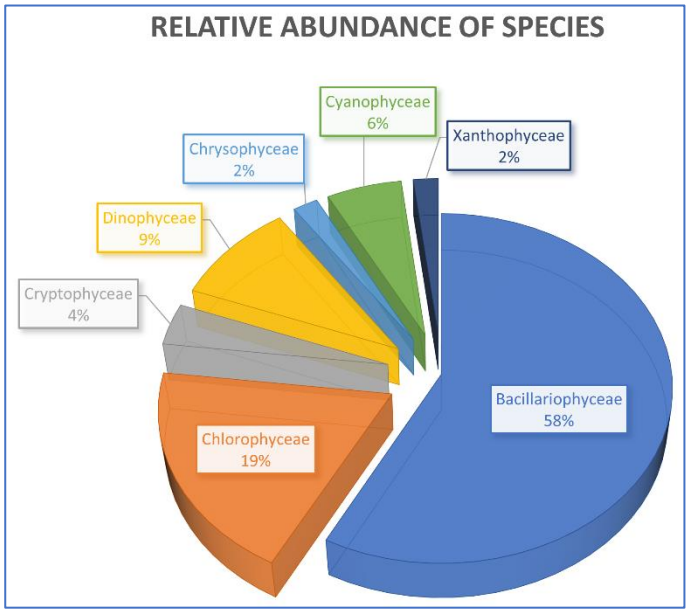

Figure 12. Relative abundance (%) of the number of species from different algae classes in Lake Llanquihue.

4. Discussion

Chlorophyll-a is an important water quality parameter for characterizing the eutrophication status of lake water bodies [73]. Our results indicate that the physical and biological characteristics of the aquatic system of Lake Llanquihue vary seasonally and that these changes influence chlorophyll-a. It is important to study the behavior of physical, chemical, and biological parameters and how they vary spatially according to the distinctive characteristics of freshwater aquatic systems to understand their structure and functioning. The behavior of the Chl-a variable in Lake Llanquihue can be derived from different parameters that determine algal productivity, such as temperature, transparency, and radiation, and chemical factors such as the nutrients nitrogen and total phosphorus, which is why these model input variables, as well as other lake system characteristics, were obtained over a 30-year period. On the other hand, the amount of inland water is very small relative to oceanic water, but it has much faster renewal periods [3]. Lake Llanquihue, due to its morphological and physical-geographical characteristics, has a prolonged renewal period for this aquatic ecosystem, where this stage of water renewal occurs every 74 years. Any disturbance in its quality would have ecological consequences for the organisms that inhabit the system, as well as economic and social consequences because it is a lake with multiple uses, including aquaculture and tourism.

In Chile, the traditional in situ monitoring carried out by the Dirección General de Aguas (DGA) is performed twice a year, in the summer and spring; however, it is not enough to monitor an aquatic system with random monitoring days to maintain a continuous system that reflects the real functioning of the physical–chemical or biological phenomena. In this work, for the first time in a Chilean lake, models based on deep learning were used to determine environmental variables according to an extensive input dataset from 1989 to 2020. A comparison of nine different machine learning algorithms was performed to estimate the levels of chlorophyll-a to obtain the best performance for the chlorophyll-a model for Lake Llanquihue during the study period.

It is important to consider that there is no perfect model, but having different modeling perspectives allows a wider scope in the feasibility of machine learning for a given task, that is why different models were selected to identify which performs better at predicting chlorophyll values. The analysis covered ensemble methods, including bagging (e.g., random forest) and boosting (XGBoost, AdaBoost, GradientBoosting, and LightGBM) strategies, support vector machines (SVM), and neural networks (e.g., MLP and ANN). Other works have used these algorithms in the detection of chlorophyll-a. For example, [73] used the XGBoost algorithm and obtained coefficient of determination values lower than those obtained in our study (0.88–0.90). This result may be a consequence of the smaller amount of total data used as a training set. The authors of [41] used a larger training set (n = 225) and a larger number of lakes, along with the boosting tree algorithm. They obtained an R^2 of 0.79, which was exceeded by that obtained using our algorithm (AdaBoost, R^2 = 0.99). Other authors have used various satellite sources such as multispectral images from the Landsat and Sentinel missions [34,36,43]. These investigations have used the neural networks model to estimate Chl-a and turbidity, obtaining coefficients of 0.89 and 0.71 for Landsat-9 and Sentinel-2, respectively. When the same model (ANN) was used in this study, the coefficient values were lower compared to the other models used; however, it exceeds them with respect to other metrics. In this first stage of research, it was decided to omit remote sensing data to evaluate the performance of the algorithms for the studied lake. In future work, the intention is to conduct near-real-time monitoring by combining these algorithms with data from different bands of satellite images.

5. Conclusions

Chlorophyll is a biological variable that has been reported as a good indicator of the algal communities in different aquatic systems. In this study, a series of in situ data from 1989 to 2020 recorded at eight monitoring stations spatially distributed in Lake Llanquihue was used to study the behavior of limnological variables at different points in the lake.

Nine artificial intelligence models were used combining the most important parameters in the determination of chlorophyll concentration, which determined this variable with high accuracy, with gradient boosting, LightGBM, support vector machine, and MLP obtaining the best results with respect to the real variable. Deep learning techniques were used for the first time to monitor the evolution of algal groups in Lake Llanquihue, located in southern Chile, with Chl-a as a bioindicator.

In future work, we intend to use the models that proved to be the best predictors of Chl-a for this lake system and include other satellite parameters to remotely detect events such as algal blooms. Due to climate conditions, it is difficult to monitor inland aquatic systems such as lakes in southern Chile in autumn and winter, making the use of estimation algorithms for this part of the year a novel and effective tool for monitoring events such as algal blooms identifiable by the chlorophyll variable.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/w15111994/s1, Figure S1: Correlation matrix between features.

Author Contributions: Conceptualization, L.R.-L.; methodology, L.R.-L.; software, I.D.-L.; validation, L.B.A. and L.R.-L.; formal analysis, L.R.-L.; investigation, L.R.-L.; resources, R.U.; data curation, L.R.-L.; writing—original draft preparation, L.R.-L. and R.M.-R.; writing—review and editing, L.R.-L., D.B.U. and I.D.-L.; visualization, L.B.A. and I.D.-L.; supervision, A.L.; project administration, R.U.; funding acquisition, L.R.-L. and R.U. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by VRID Universidad San Sebastián and CRHIAM (ANID/FONDAP/15130015).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available upon request from the corresponding author.

Acknowledgments: L.R.-L. is grateful to the VRIDFAI21/10 project of the Universidad San Sebastian. Special thanks are also given to the Centro de Recursos Hídricos para la Agricultura y la Minería (CRHIAM) (Project ANID/FONDAP/15130015). This publication was supported by the Vicerrectoría de Investigación y Doctorados de la Universidad San Sebastián—Fondo VRID_APC23/06.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Prakash, S. Impact of climate change on aquatic ecosystem and its biodiversity: An overview. *Int. J. Biol. Innov.* **2021**, *3*, 312–317. [CrossRef]
2. Yin, S.; Yi, Y.; Liu, Q.; Luo, Q.; Chen, K. A Review on Effects of Human Activities on Aquatic Organisms in the Yangtze River Basin since the 1950s. *River* **2022**, *1*, 104–119. [CrossRef]
3. Wetzel, R. *Limnology: Lake and River Ecosystems*; Gulf Professional Publishing: Oxford, UK, 2001.
4. Okello, C.; Tomasello, B.; Greggio, N.; Wambiji, N.; Antonellini, M. Impact of Population Growth and Climate Change on the Freshwater Resources of Lamu Island, Kenya. *Water* **2015**, *7*, 1264–1290. [CrossRef]
5. Boretti, A.; Rosa, L. Reassessing the Projections of the World Water Development Report. *NPJ Clean Water* **2019**, *2*, 15. [CrossRef]
6. Arthington, A.H.; Dulvy, N.K.; Gladstone, W.; Winfield, I.J. Fish Conservation in Freshwater and Marine Realms: Status, Threats and Management. *Aquat. Conserv. Mar. Freshw. Ecosyst.* **2016**, *26*, 838–857. [CrossRef]
7. Decaëns, T.; Martins, M.B.; Feijoo, A.; Oszwald, J.; Dolédec, S.; Mathieu, J.; Arnaud de Sartre, X.; Bonilla, D.; Brown, G.G.; Cuellar Criollo, Y.A.; et al. Biodiversity Loss along a Gradient of Deforestation in Amazonian Agricultural Landscapes. *Conserv. Biol.* **2018**, *32*, 1380–1391. [CrossRef] [PubMed]
8. Valdés-Pineda, R.; García-Chevesich, P.; Valdés, J.B.; Pizarro-Tapia, R. The First Drying Lake in Chile: Causes and Recovery Options. *Water* **2020**, *12*, 290. [CrossRef]
9. Ocampo-Melgar, A.; Barria, P.; Chadwick, C.; Diaz-Vasconcellos, R. Rural Transformation and Differential Vulnerability: Exploring Adaptation Strategies to Water Scarcity in the Aculeo Lake Basin (Chile). *Front. Environ. Sci.* **2022**, *10*, 955023. [CrossRef]
10. Palmeirim, A.F.; Santos-Filho, M.; Peres, C.A. Marked Decline in Forest-Dependent Small Mammals Following Habitat Loss and Fragmentation in an Amazonian Deforestation Frontier. *PLoS ONE* **2020**, *15*, e0230209. [CrossRef]

11. 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]
12. Giralt, D.; Pantoja, J.; Morales, M.B.; Traba, J.; Bota, G. Landscape-Scale Effects of Irrigation on a Dry Cereal Farmland Bird Community. *Front. Ecol. Evol.* **2021**, *9*, 611563. [CrossRef]
13. Sun, Y.; Mao, X.; Gao, T.; Liu, H.; Zhao, Y. Potential Water Withdrawal Reduction to Mitigate Riverine Ecosystem Degradation under Hydropower Development: A Computable General Equilibrium Model Analysis. *River Res. Appl.* **2021**, *37*, 1223–1230. [CrossRef]
14. Newsome, D. The Collapse of Tourism and Its Impact on Wildlife Tourism Destinations. *J. Tour. Futur.* **2020**, *7*, 295–302. [CrossRef]
15. Rodríguez-López, L.; González-Rodríguez, L.; Duran-Llacer, I.; García, W.; Cardenas, R.; Urrutia, R. Assessment of the Diffuse Attenuation Coefficient of Photosynthetically Active Radiation in a Chilean Lake. *Remote Sens.* **2022**, *14*, 4568. [CrossRef]
16. Bai, Y.; Ochuodho, T.O.; Yang, J. Impact of Land Use and Climate Change on Water-Related Ecosystem Services in Kentucky, USA. *Ecol. Indic.* **2019**, *102*, 51–64. [CrossRef]
17. Rodríguez-López, L.; González-Rodríguez, L.; Duran-Llacer, I.; Cardenas, R.; Urrutia, R. Spatio-Temporal Analysis of Chlorophyll in Six Araucanian Lakes of Central-South Chile from Landsat Imagery. *Ecol. Inf.* **2021**, *65*, 101431. [CrossRef]
18. Sánchez, O.; Robla, J.; Arias, A. Annotated and Updated Checklist of Land and Freshwater Molluscs from Asturias (Northern Spain) with Emphasis on Parasite Transmitters and Exotic Species. *Diversity* **2021**, *13*, 415. [CrossRef]
19. Liu, X.; Wang, H. Effects of Loss of Lateral Hydrological Connectivity on Fish Functional Diversity. *Conserv. Biol.* **2018**, *32*, 1336–1345. [CrossRef]
20. Doucet, C.V.; Labaj, A.L.; Kurek, J. Microfiber Content in Freshwater Mussels from Rural Tributaries of the Saint John River, Canada. *Water Air Soil Pollut.* **2021**, *232*, 32. [CrossRef]
21. Amtmann, C.A.; Blanco, G. Efectos de la Salmonicultura en las Economías Campesinas de la Región de Los Lagos, Chile. *Rev. Austral Cienc. Soc.* **2001**, 93–106. Available online: https://scholar.google.com.hk/scholar?hl=zh-CN&as_sdt=0%2C5&q=Efectos+de+la+Salmonicultura+en+las+Econom%C3%ADas+Campesinas+de+la+Regi%C3%B3n+de+Los+Lagos&btnG= (accessed on 13 April 2023). [CrossRef]
22. Feng, J.; Zhao, Z.; Wen, Y.; Hou, Y. Organically Linking Green Development and Ecological Environment Protection in Poyang Lake, China Using a Social-Ecological System (Ses) Framework. *Int. J. Environ. Res. Public Health* **2021**, *18*, 2572. [CrossRef] [PubMed]
23. Waters, S.; Hamilton, D.; Pan, G.; Michener, S.; Ogilvie, S. Oxygen Nanobubbles for Lake Restoration—Where Are We at? A Review of a New-Generation Approach to Managing Lake Eutrophication. *Water* **2022**, *14*, 1989. [CrossRef]
24. Modabberi, A.; Noori, R.; Madani, K.; Ehsani, A.H.; Danandeh Mehr, A.; Hooshyaripor, F.; Kløve, B. Caspian Sea Is Eutrophying: The Alarming Message of Satellite Data. *Environ. Res. Lett.* **2019**, *15*, 124047. [CrossRef]
25. Mozafari, Z.; Noori, R.; Siadatmousavi, S.M.; Afzalimehr, H.; Azizpour, J. Satellite-Based Monitoring of Eutrophication in the Earth's Largest Transboundary Lake. *GeoHealth* **2023**, *7*, e2022GH000770. [CrossRef]
26. McKindles, K.; Frenken, T.; McKay, R.M.L.; Bullerjahn, G.S. Binational Efforts Addressing Cyanobacterial Harmful Algal Blooms in the Great Lakes. In *Handbook of Environmental Chemistry*; Springer Science and Business Media Deutschland GmbH: Berlin, Germany, 2020; Volume 101, pp. 109–133.
27. Haas, M.; Baumann, F.; Castella, D.; Haghipour, N.; Reusch, A.; Strasser, M.; Eglinton, T.I.; Dubois, N. Roman-Driven Cultural Eutrophication of Lake Murten, Switzerland. *Earth Planet. Sci. Lett.* **2019**, *505*, 110–117. [CrossRef]
28. Zhao, W.L.; Gentine, P.; Reichstein, M.; Zhang, Y.; Zhou, S.; Wen, Y.; Lin, C.; Li, X.; Qiu, G.Y. Physics-Constrained Machine Learning of Evapotranspiration. *Geophys. Res. Lett.* **2019**, *46*, 14496–14507. [CrossRef]
29. Al-Adhaileh, M.H.; Alsaade, F.W. Modelling and Prediction of Water Quality by Using Artificial Intelligence. *Sustainability* **2021**, *13*, 4259. [CrossRef]
30. Najah Ahmed, A.; Binti Othman, F.; Abdulmohsin Afan, H.; Khaleel Ibrahim, R.; Ming Fai, C.; Shabbir Hossain, M.; Ehteram, M.; Elshafie, A. Machine Learning Methods for Better Water Quality Prediction. *J. Hydrol.* **2019**, *578*, 124084. [CrossRef]
31. Li, X.; Sha, J.; Wang, Z.L. Chlorophyll-A Prediction of Lakes with Different Water Quality Patterns in China Based on Hybrid Neural Networks. *Water* **2017**, *9*, 524. [CrossRef]
32. Ubah, J.I.; Orakwe, L.C.; Ogbu, K.N.; Awu, J.I.; Ahaneku, I.E.; Chukwuma, E.C. Forecasting Water Quality Parameters Using Artificial Neural Network for Irrigation Purposes. *Sci. Rep.* **2021**, *11*, 24438. [CrossRef]
33. Zhang, J.; Fu, P.; Meng, F.; Yang, X.; Xu, J.; Cui, Y. Estimation Algorithm for Chlorophyll-a Concentrations in Water from Hyperspectral Images Based on Feature Derivation and Ensemble Learning. *Ecol. Inf.* **2022**, *71*, 101783. [CrossRef]
34. Pahlevan, N.; Smith, B.; Schalles, J.; Binding, C.; Cao, Z.; Ma, R.; Alikas, K.; Kangro, K.; Gurlin, D.; Hà, N.; et al. Seamless Retrievals of Chlorophyll-a from Sentinel-2 (MSI) and Sentinel-3 (OLCI) in Inland and Coastal Waters: A Machine-Learning Approach. *Remote Sens. Environ.* **2020**, *240*, 111604. [CrossRef]
35. Li, S.; Song, K.; Wang, S.; Liu, G.; Wen, Z.; Shang, Y.; Lyu, L.; Chen, F.; Xu, S.; Tao, H.; et al. Quantification of Chlorophyll-a in Typical Lakes across China Using Sentinel-2 MSI Imagery with Machine Learning Algorithm. *Sci. Total Environ.* **2021**, *778*, 146271. [CrossRef] [PubMed]

36. Kuhn, C.; de Matos Valerio, A.; Ward, N.; Loken, L.; Sawakuchi, H.O.; Kampel, M.; Richey, J.; Stadler, P.; Crawford, J.; Striegl, R.; et al. Performance of Landsat-8 and Sentinel-2 Surface Reflectance Products for River Remote Sensing Retrievals of Chlorophyll-a and Turbidity. *Remote Sens. Environ.* **2019**, *224*, 104–118. [CrossRef]
37. Essam, Y.; Huang, Y.F.; Birima, A.H.; Ahmed, A.N.; El-Shafie, A. Predicting Suspended Sediment Load in Peninsular Malaysia Using Support Vector Machine and Deep Learning Algorithms. *Sci. Rep.* **2022**, *12*, 302. [CrossRef]
38. Schendorf, T.M.; Del Vecchio, R.; Koech, K.; Blough, N.V. A Standard Protocol for NaBH4 Reduction of CDOM and HS. *Limnol. Ocean. Methods* **2016**, *14*, 414–423. [CrossRef]
39. Rubin, H.J.; Lutz, D.A.; Steele, B.G.; Cottingham, K.L.; Weathers, K.C.; Ducey, M.J.; Palace, M.; Johnson, K.M.; Chipman, J.W. Remote Sensing of Lake Water Clarity: Performance and Transferability of Both Historical Algorithms and Machine Learning. *Remote Sens.* **2021**, *13*, 1434. [CrossRef]
40. Cao, Z.; Ma, R.; Duan, H.; Pahlevan, N.; Melack, J.; Shen, M.; Xue, K. A Machine Learning Approach to Estimate Chlorophyll-a from Landsat-8 Measurements in Inland Lakes. *Remote Sens. Environ.* **2020**, *248*, 111974. [CrossRef]
41. Niroumand-Jadidi, M.; Bovolo, F.; Bresciani, M.; Gege, P.; Giardino, C. Water Quality Retrieval from Landsat-9 (OLI-2) Imagery and Comparison to Sentinel-2. *Remote Sens.* **2022**, *14*, 4596. [CrossRef]
42. Delorenzo, A.; Colibri, A. *Preservation of a Pristine Lake for Future Generations: Llanquihue Lake, X Region, Chile*; LA 222 Hydrology for Planners Preservation of a Pristine Lake for Future Generations Term Project Proposal 2 Llanquihue Lake Preserving Pristine Lake; UC Berkeley: Berkeley, CA, USA, 2012; Available online: https://escholarship.org/uc/item/8b5146kj (accessed on 8 January 2023).
43. Collado, G.A.; Vidal, M.A.; Torres-Díaz, C.; Cabrera, F.J.; Araya, J.F.; Darrigran, G. Morphological and Molecular Identification of the Invasive Freshwater Snail Physa Acuta (Gastropoda: Physidae) into Llanquihue Lake, Chilean Patagonia. *An. Acad. Bras. Cienc.* **2020**, *92*, e20181101. [CrossRef]
44. DGA Atlas Del Agua. *Atlas del Agua: Chile 2016*; DGA: Los Angeles, CA, USA, 2016; Volume 1, pp. 1–24.
45. Campos, H.; Steffen, W.; Aguero, G.; Parra, O.; Zúñiga, L. Limnological study of Lake Llanquihue (Chile): Morphometry, physics, chemistry, plankton and primary productivity. *Arch. Hydrobiol. Supplementband. Monogr. Beiträge* **1988**, *81*, 37–67.
46. Tax, D.; Duin, R.; Juszczak, P.; Tax, D.M.J.; Duin, R.P.W. Feature Scaling in Support Vector Data Description. In Proceedings of the 8th Annual Conference of the Advanced School for Computing and Imaging, Lochem, The Netherlands, 19–21 June 2002.
47. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
48. Amit, Y.G.D. Cuantificación y Reconocimiento de Formas Con Árboles Aleatorios. *Comput. Neuronal* **1997**, *9*, 1545–1588. [CrossRef]
49. Breiman, L. Bagging Predictors. *Mach. Learn.* **1996**, *24*, 123–140. [CrossRef]
50. Hastie, T.; Tibshirani, R.; Friedman, J. Random Forests. In *The Elements of Statistical Learning*; Springer: Berlin/Heidelberg, Germany, 2009; pp. 587–604.
51. Schapire, R.E. Explaining AdaBoost. *Kybernetes* **2013**, *42*, 164–166. [CrossRef]
52. Cao, Y.; Miao, Q.-G.; Liu, J.-C.; Gao, L. Advance and Prospects of AdaBoost Algorithm. *Acta Autom. Sin.* **2013**, *39*, 745–758. [CrossRef]
53. Friedman, J.H. Greedy Function Approximation: A Gradient Boosting Machine. *Ann. Stat.* **2001**, *29*, 1189–1232. [CrossRef]
54. Friedman, J.H. Stochastic Gradient Boosting. *Comput. Stat. Data Anal.* **2002**, *38*, 367–378. [CrossRef]
55. Chen, Y.; Jia, Z.; Mercola, D.; Xie, X. A Gradient Boosting Algorithm for Survival Analysis via Direct Optimization of Concordance Index. *Comput. Math. Methods Med.* **2013**, *2013*, 873595. [CrossRef]
56. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; Association for Computing Machinery, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794.
57. Mitchell, R.; Frank, E. Accelerating the XGBoost Algorithm Using GPU Computing. *PeerJ Comput. Sci.* **2017**, *3*, e127. [CrossRef]
58. Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.-Y. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; Volume 30.
59. Cortes, C.; Vapnik, V.; Saitta, L. *Support-Vector Networks Editor*; Kluwer Academic Publishers: Dordrecht, The Netherlands, 1995; Volume 20.
60. Smola, A.J.; Schölkopf, B.; Schölkopf, S. *A Tutorial on Support Vector Regression*; Kluwer Academic Publishers: Dordrecht, The Netherlands, 2004; Volume 14.
61. Park, Y.; Cho, K.H.; Park, J.; Cha, S.M.; Kim, J.H. Development of Early-Warning Protocol for Predicting Chlorophyll-a Concentration Using Machine Learning Models in Freshwater and Estuarine Reservoirs, Korea. *Sci. Total Environ.* **2015**, *502*, 31–41. [CrossRef] [PubMed]
62. Ramchoun, H.; Amine, M.; Idrissi, J.; Ghanou, Y.; Ettaouil, M. Multilayer Perceptron: Architecture Optimization and Training. *Int. J. Interact. Multimed. Artif. Intell.* **2016**, *4*, 26. [CrossRef]
63. Mamun, M.; Kim, J.J.; Alam, M.A.; An, K.G. Prediction of Algal Chlorophyll-a and Water Clarity in Monsoon-Region Reservoir Using Machine Learning Approaches. *Water* **2020**, *12*, 30. [CrossRef]
64. Cui, S.; Liu, Y.; Zhang, Y.; He, L.; Wu, X. Algae Biomass and Radius Prediction Based on ARMA-BP Neural Network Combination Model. In *ACM International Conference Proceeding Series, Proceedings of the 2020 12th International Conference on Computer and*

Automation Engineering, Sydney, NSW Australia, 14 February 2020; Association for Computing Machinery: New York, NY, USA, 2020; pp. 23–26.
65. Noori, R.; Karbassi, A.R.; Ashrafi, K.; Ardestani, M.; Mehrdadi, N. Development and Application of Reduced-Order Neural Network Model Based on Proper Orthogonal Decomposition for BOD5 Monitoring: Active and Online Prediction. *Enviorn. Prog. Sustain. Energy* **2013**, *32*, 120–127. [CrossRef]
66. Marinósdóttir, H.; Jóhannsdóttir, A. Applications of Different Machine Learning Methods for Water Level Predictions. Ph.D. Thesis, Reykjavík University, Reykjavík, Iceland, 2019.
67. Mosier, C. Problems and Designs of Cross-Validation. *Educ. Psychol. Meas.* **1951**, *11*, 5–11. [CrossRef]
68. Browne, M.W. Cross-Validation Methods. *J. Math. Psychol.* **2000**, *44*, 108–132. [CrossRef]
69. Refaeilzadeh, P.; Tang, L.; Liu, H. Cross-Validation. In *Encyclopedia of Database Systems*; Springer: New York, NY, USA, 2009; Volume 5, pp. 532–538.
70. Bergstra, J.; Bengio, Y. Random Search for Hyper-Parameter Optimization. *J. Mach. Learn. Res.* **2012**, *13*, 281–305.
71. Alam, M.A.; Fukumizu, K. Hyperparameter Selection in Kernel Principal Component Analysis. *J. Comput. Sci.* **2014**, *10*, 1139–1150. [CrossRef]
72. DGA. *Ministerio de Obras Públicas Nombre Consultores: Director del Proyecto Profesionales*; DGA: Santiago, Chile, 2018.
73. Cui, Y.; Meng, F.; Fu, P.; Yang, X.; Zhang, Y.; Liu, P. Application of Hyperspectral Analysis of Chlorophyll a Concentration Inversion in Nansi Lake. *Ecol. Inf.* **2021**, *64*, 101360. [CrossRef]