Inland Waters

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/tinw20

Machine learning modeling of lake chlorophyll in a data-scarce region (Northern Patagonia, Chile): insights for environmental monitoring

Luciano Caputo, Cristian Rios Molina, Roxanna Ayllon-Arauco & Iván Felipe Benavides

To cite this article: Luciano Caputo, Cristian Rios Molina, Roxanna Ayllon-Arauco & Iván Felipe Benavides (2024) Machine learning modeling of lake chlorophyll in a data-scarce region (Northern Patagonia, Chile): insights for environmental monitoring, Inland Waters, 14:1-2, 83-96, DOI: 10.1080/20442041.2024.2359329

To link to this article: https://doi.org/10.1080/20442041.2024.2359329

INLAND WATERS
2024, VOL. 14, NOS. 1–2, 83–96
https://doi.org/10.1080/20442041.2024.2359329

Machine learning modeling of lake chlorophyll in a data-scarce region (Northern Patagonia, Chile): insights for environmental monitoring

Luciano Caputo,[a,b] Cristian Rios Molina ,[c,d] Roxanna Ayllon-Arauco,[d] and Iván Felipe Benavides [e]

[a]Instituto de Ciencias Marinas y Limnológicas (ICML), Facultad de Ciencias, Instituto de Ciencias Marinas y Limnológicas, Universidad Austral de Chile, Valdivia, Chile; [b]Núcleo GIC-ADAPRES, Centro Transdisciplinario de Estudios Ambientales y Desarrollo Humano Sostenible (CEAM), Universidad Austral de Chile, Valdivia, Chile; [c]Núcleo Milenio de Agronomía Marina de Algas (MASH), Facultad de Ciencias, Universidad Austral de Chile, Valdivia, Chile; [d]Programa de Doctorado en Biología Marina, Facultad de Ciencias, Universidad Austral de Chile, Valdivia, Chile; [e]Grupo de Investigación Recursos Hidrobiológicos, Universidad Nacional de Colombia, sede Palmira, Colombia

ABSTRACT
Among South American countries, Chile is highly susceptible to climate change impacts on water resources and ecosystems. Chilean lakes and rivers have been impacted by anthropogenic activities leading to chemical pollution and eutrophication. Concerns for conservation and management of water resources have led to the current development of regulations for environmental quality of Northern Patagonian lakes. In this context, we analyze historical limnological databases (1979–2022) for these lakes utilizing random forest (RF) models. After filtering, we retained data for 11 lakes including key variables of dissolved oxygen, conductivity, transparency, temperature, pH, total nitrogen, total phosphorus, and chlorophyll *a*. This dataset yielded robust results, accurately predicting chlorophyll *a* concentration. Furthermore, we added lake geomorphological parameters, enhancing the performance of the model. Our study demonstrates the need to improve long-term monitoring programs, optimizing environmental data recording for efficient investment. We conclude that the studied lakes generally maintain their oligotrophic characteristics and are more sensitive to nitrogen than phosphorus loading. Our results highlight the need to implement adaptative management plans at the watershed level to regulate anthropogenic nitrogen contamination from agriculture, pisciculture, and urbanization. The features selected by RF, coupled with the assessment of historical trophic state variation, allow the establishment of permissible concentration thresholds for major nutrients and other sentinel variables, informing the development of regulations for environmental quality. Lastly, the enhanced performance of RF modeling that includes geographical variables demonstrates the need to standardize and integrate geographical data in monitoring practices.

ARTICLE HISTORY
Received 6 October 2023
Accepted 20 May 2024

KEYWORDS
environmental policy; oligotrophic patterns; optimized monitoring; random forest; regional limnology; water quality

Introduction

Global change promoters such as land use change, eutrophication, hydrological disturbance, chemical pollution, overexploitation, and invasive species are affecting the structure and functioning of freshwater ecosystems at the global level (Janse et al. 2015, Tickner et al. 2020, Woolway et al. 2020, Merz et al. 2023). Many empirical studies based on in situ measurements alert us to global trends of eutrophication (Wurtsbaugh et al. 2019), salinization (Hébert et al. 2023), and deoxygenation (Jane et al. 2021) in lakes, with a generalized increase of phytoplankton blooms since the 1980s (Taranu et al. 2015, Ho et al. 2019). Regional trends of lake eutrophication and increased frequency and intensity of cyanobacteria blooms are related with human impact on watersheds and increased water temperature in lakes due to global warming (Meerhoff et al. 2022). South American rivers and lakes are among the most vulnerable on the planet because of increased human pressure over hydrological resources for industrial applications, with negative impacts on water quality and biodiversity (Torremorell et al. 2021). Moreover, several South American countries are currently experiencing broad water shortages, such as Uruguay, Argentina, and Bolivia (Brêda et al. 2020). Consensus is that among South American countries, Chile is highly vulnerable to climate change impacts and hydrological stress in its watersheds because of the physiography and environmental gradients associated with the broad latitudinal variation and altitude along the Andes (IPCC 2023).

CONTACT Cristian Rios Molina cristian.ariosm@gmail.com Núcleo Milenio de Agronomía Marina de Algas (MASH), Facultad de Ciencias, Universidad Austral de Chile, 5090000 Valdivia, Chile; Programa de Doctorado en Biología Marina, Facultad de Ciencias, Universidad Austral de Chile, Campus Isla Teja, 5090000 Valdivia, Chile

Supplemental material for this article can be accessed online at https://doi.org/10.1080/20442041.2024.2359329.

Chilean freshwaters comprise 101 basins, including 1200 rivers and 15 000 lakes and lagoons (DGA 2014), increasingly threatened by the effects of climate change and the combination of rising human pressures on water resources (IPCC 2023), with several confirmed impacts in terms of severity and spatial representation across the country (Fierro et al. 2017, Fuentealba et al. 2021, Navedo and Vargas-Chacoff 2021, Alaniz et al. 2022, Gutiérrez et al. 2022, Hidalgo-Corrotea et al. 2023). The rising demand and reduced availability of freshwater combined with poor water management has induced problems of water scarcity and pollution (Pizarro et al. 2022, Herrera et al. 2023), leading the country into an unprecedented water crisis. Recent studies state that the greatest ecological risk for lakes and rivers in southern Chile is related to combined effects of land use change (Echeverría et al. 2012), mainly driven by deforestation and urban expansion, resulting in the discharge of residual water from populated areas (Fierro et al. 2017, Hidalgo-Corrotea et al. 2023), and by the massive expansion of freshwater and marine aquaculture (Nimptsch et al. 2015). Anthropogenic pressures on these lakes have led to changes in productivity and signs of eutrophication over the last decade (Pizarro et al. 2016) and might create suitable conditions for invasive species colonization (Caputo et al. 2018). For example, in Lake Rupanco (León-Muñoz et al. 2013) and Lake Villarrica (Nimptsch et al. 2016), chlorophyll *a* concentrations have increased (Chl-*a* ≥10 µg L^{-1}; Nimptsch et al. 2015). Chl-*a* has been widely used as an indicator of phytoplankton biomass and is a reliable proxy for trophic status and environmental change (Felip and Catalan 2000). Along with the trophic status increase, summer blooms of *Dolichospermum* sp. have been recorded more frequently (Nimptsch et al. 2016, Rodríguez-López et al. 2023). Lake Villarrica has shown a significant ecological degradation, where in <2 decades its trophic state changed from oligotrophic with high transparency and low phytoplankton productivity, to meso-eutrophic with high productivity, toxic algal blooms, and reduced transparency (Rodríguez-López et al. 2023). In response, this lake has become the first waterbody in the National Clean-up Plan.

Nonetheless, many lakes in southern Chile still maintain their oligotrophic condition with transparent waters, low mineral concentrations, and relatively neutral pH (Campos 1984, Woelfl 2007), thus making early action on risk mitigation a priority for government entities in charge of environmental monitoring and regulation (MMA 2020). Since the 1980s, the National Water Authority (DGA) through its Department of Conservation and Protection of Hydric Resources has developed the Lake Control Network monitoring program ("Red minima de Lagos"), which currently includes the seasonal monitoring of 20 lakes along Chile. The Ministry of the Environment (MMA) is the government institution responsible for the coordination to generate environmental quality and emission standards to "regulate pollutants, so as to prevent that their concentrations and residence time could risk environmental conservation" (Law 19.300). The last environmental regulation program of MMA (2020–2021) prioritizes secondary norms for environmental quality (NSCA) for the protection of Northern Patagonian lakes. These regulations mainly aim to decrease risks to nature conservation and hydric security by maintaining water quality and trophic status of lakes. The NSCA are fundamental environmental regulations for compliance to the recently enacted Climate Change Law Framework (LMCC; 13 June 2022, Law N° 21.455) that offer an opportunity to develop integrated watershed management programs (currently nonexistent), aligned with the international commitments ratified by the Chilean government during the last Conference of the Agreement of Parts in Biological Diversity (COP 15 Kunming-Montreal 2022).

In this context, the general objective of this study is to analyze historical limnological databases (1979–2022) from the DGA for Northern Patagonian lakes and identify environmental variables that better predict Chl-*a* as a proxy of productivity over a regional scale in Chile. Although the DGA is the main institution in charge of environmental monitoring of waterbodies, the databases are highly heterogeneous, discontinuous, and "noisy" (Supplemental Fig. S1–S2), needing appraisal. For this purpose, we implemented machine learning (ML) tools for data modeling, specifically random forest (RF), and different posterior techniques to interpret and explain results, such as a metric of variable importance ranking and partial dependence between predictor variables and response. RF-based models have been successfully applied to rivers, lakes, and reservoirs for different purposes, and proven to be useful to analyze the behavior of limnological variables describing water quality, trophic state evolution, and forecasting harmful algal blooms (e.g., Hollister et al. 2016, Derot et al. 2020, Virro et al. 2022). From our results, we propose an optimized monitoring design based on the selection of key parameters to improve sampling standardization, minimize costs, and maximize monitoring efficacy.

Material and methods

Data sources

In situ limnological data for these lakes were obtained from various databases maintained and curated by the

DGA, Chilean Government (public repository for DGA data: http://www.dga.cl/servicioshidrometeorologicos/). Monitoring campaigns to build these databases were conducted by the DGA between 1979 and 2021, generally 4 times per year (seasonally), usually with 3 sampling stations per lake at different depths. Additionally, part of the data are from the MMA, Chilean Government, acquired through direct request following the mechanisms outlined in the Chilean legislation, specifically the law for access to public information (Ley Sobre Acceso a la Información Pública N° 20285). The consolidated database for the current study encompasses 8371 observations for 18 Northern Patagonian lakes, with 18 limnological parameters recorded in situ: coordinates, water sample depth (WSD), temperature (Temp), electrical conductivity (EC), Chl-*a*, pH, dissolved oxygen (DO), oxygen saturation % (DO_sat), turbidity (NTU), transparency characterized as Secchi depth (hereafter Transp), nitrate (NO_3), nitrite (NO_2), ammonium (NH_4), Kjeldahl nitrogen (KN), total nitrogen (TN), total phosphorus (TP), phosphate (PO_4), silica (SiO_2), and chemical oxygen demand (COD). After subsequent filtering steps, we constructed the primary database for modeling (henceforth named DGA-only; summarized in Supplemental Table S1). Additionally, from relevant literature we obtained 6 different geomorphological and geographical lake parameters to construct an extended dataset for model testing (henceforth named Geographical-added): altitude (m a.s.l.), lake area (LA), lake volume (LV), maximum depth (Zmax), euphotic and maximum depth ratio (Zeuf/Zmax), and watershed area (WA) (geographical and morphometric characteristics and reference coordinates in Table 1).

Study area

The study lakes have geographical proximity within Chilean Northern Patagonia (ranging from 39°S to 43°S), mainly characterized as monomictic (with depths >100 m; Fig. 1). They are situated within well-preserved Andean watersheds that share common physiographic features and a glacial origin (Geller 1992, Pizarro et al. 2016). The climate of the study area is characterized as temperate rainy, experiencing precipitation throughout the year, with an annual average >3300 mm. The dominant vegetation in the region is the Andean-Patagonian temperate forest, composed primarily of evergreen species such as different *Nothofagus* spp. The rainy season is between May and August, whereas the dry season spans from December to April. The average surface annual water temperature (0–10 m) across the studied lakes is 15.6 °C (max = 19.4 °C, min = 10.3 °C), with a mean summer of 19.3 °C and winter of 10.9 °C (obtained from the database used in this study).

Data preparation

(a) *Dataset homogenization*: Given the significant heterogeneity in data quality among lake datasets, we conducted a rigorous data cleaning and homogenization process to create an appropriate final dataset for subsequent analyses. The following steps were employed to accomplish this process: (1) removal of non-numeric characters and values (e.g., comments, observations, signs, etc.), involving extracting the numeric values from cells whenever available or assigning the null value code (NA) when appropriate; (2) correction and standardization of variable labels, such as lake names, dates, and seasons; (3) revision and homogenization of the formats used to separate columns in the datasets (e.g., comma, tabulation or space); (4) unification of redundant variables, such as TN, which may have been separated into different columns based on the laboratory responsible for the measurement and data reporting.

(b) *Dataset filtering*: Our resulting dataset underwent further exploration and filtering based on heterogeneity criteria in the sampling, missing data, and potentially deviant data. The following steps were implemented to ensure the integrity and reliability of the final dataset for subsequent analyses. (1) Because of a significant bias in sampling towards shallower depths (~81% of the observations were within 50 m depth; see Supplemental Fig. S1), we generated 2 datasets with and without values deeper than 50 m, to test if samples of greater depths contribute to the performance of the model or are negligible for this purpose. (2) An examination of missing data revealed highly heterogeneous sampling patterns across years and lakes (Supplemental Fig. S2); consequently, only variables with reliable data integrity were selected for the final dataset (those retained in Supplemental Table S1). (3) Missing data and outliers were removed from the final dataset according to expert criteria. (4) We also incorporated geographical and morphometric parameters of these lakes, obtained from relevant literature (Supplemental Table S1). By applying these rigorous steps, we ensured that the final dataset used for subsequent analyses was of high quality and suitable for addressing our research objectives.

Table 1. Geographical and morphometric parameters of the study lakes.

Lake	Latitude (°S)	Longitude (°W)	Altitude (m a.s.l.)	Watershed area (km^2)	Lake area (km^2)	Lake volume (km^3)	*Zmax* (m)	*Zeuf/Zmax*
Caburga (CAB)	39°07′	71°46′	505	325	51.9	8.88	327	5
Calafquen (CAL)	39°30′	72°09′	203	733	120.6	13.91	212	7
Chapo (CHA)	41°25′	72°32′	241	323	45.3	8.296	298	6
Colico (COL)	39°04′	72°00′	473	432	60.9	9.5	374	13.3
Maihue (MAI)	40°40′	72°29′	90	1602	47.2	5.7	207	5
Neltume (NEL)	39°43′	72°12′	197	763	9.8	0.6	90	27
Panguipulli (PAN)	39°47′	71°58′	130	3811	116.9	14.7	268	7
Puyehue (PUY)	40°48′	72°33′	188	1510	165.4	12.6	123	18
Ranco (RAN)	40°14′	72°27′	64	3997	442.6	54.1	199	8
Rupanco (RUP)	40°48′	72°33′	123	909	247.5	38.0	273	14
Todos los Santos (TOD)	41°05′	72°15′	189	3036	178.5	34.4	337	6

(c) *Final preparation for modeling:* Chl-*a* was chosen as the response variable as an indicator of phytoplankton biomass present in water and proxy for trophic status. Following data pre-processing and filtering, we defined 2 sets of predictor variables. The first set comprised the in situ measurements retained from the DGA databases after filtering, including latitude, longitude, season of the year, WSD, Temp, EC, pH, DO, Transp, TN, and TP (DGA-only). The second dataset included the DGA-only data plus geographical and morphological variables obtained from the literature (Geographical-added). Using these 2 datasets, DGA-only and Geographical-added, we conducted a comparison of RF regression models' performance. The primary objective was to assess whether the inclusion of the supplementary geographical and morphological information enhanced the models' predictive capabilities. Through this analysis, we investigated whether the integration of geographical and morphological factors could lead to more accurate predictions of Chl-*a* concentrations, providing valuable insights into the influence of these variables on phytoplankton biomass and trophic status in the studied lakes.

Random forest training and validation

Random forest (RF) is an ensemble algorithm of several prediction trees, characterized for operating with an iteration process in which each tree is generated with randomly selected training data partitioned according to a random subset of predictor variables. In general, RF is a nonlinear, nonparametric method able to manage a large number of predictors, has demonstrated sturdiness for the analysis of noisy databases with outliers while analyzing significant relationships among variables, and is based on a consensus of multiple iterations of the ensemble (Breiman 2001). Additionally, compared to more sophisticated ML models such as artificial neural network, RF has lower computational requirements and is able to generate highly precise predictions in diverse case studies (e.g., Han et al. 2018, Yokoyama and Yamaguchi 2020, Watanabe et al. 2021). Hence, this ML algorithm is highly appropriate for analyzing and generating predictive models and gaining insights about challenging databases without a specific experimental design, such as the one in our study, where traditional statistical approaches such as linear or mixed models may not deliver reliable or robust results.

RF analysis was performed using the R package *randomForest* v4.7-1.1 (Liaw and Wiener 2002) under the R-base version 4.3.2 (R Core Team 2023). Because Chl-*a* is a continuous variable, a regression type RF was selected for the analysis. To prevent overfitting, the categorical variables of "Lake" and "Year" were excluded, and only the limnological and geographical data specified earlier were used as predictors. Additionally, the sampling season was transformed into a dummy variable. Both the DGA-only and Geographical-added datasets were randomly partitioned into training and testing in a 70:30 ratio. For the training dataset, the hyper parameters (number of variables randomly sampled at each split [mtry] and number of trees to grow [ntree]) were fine-tuned using the 10-fold cross-validation method with a grid search ranging from 1 to the total number of variables for each dataset. This process was conducted using the R package *caret* v6.0-94 (Kuhn 2008). By following this methodology, we aimed to ensure robust model training and hyper parameter optimization for the RF regression analysis, ultimately leading to accurate predictions of Chl-*a* concentrations in the studied lakes.

Model performance was assessed by comparing the observed and predicted Chl-*a* values using the R package *stats* v4.3.2 (R Core Team 2023) with the test dataset. The Spearman correlation coefficient (Spearman's *R*) was employed as an indicator of RF performance to predict Chl-*a* concentrations. To compare

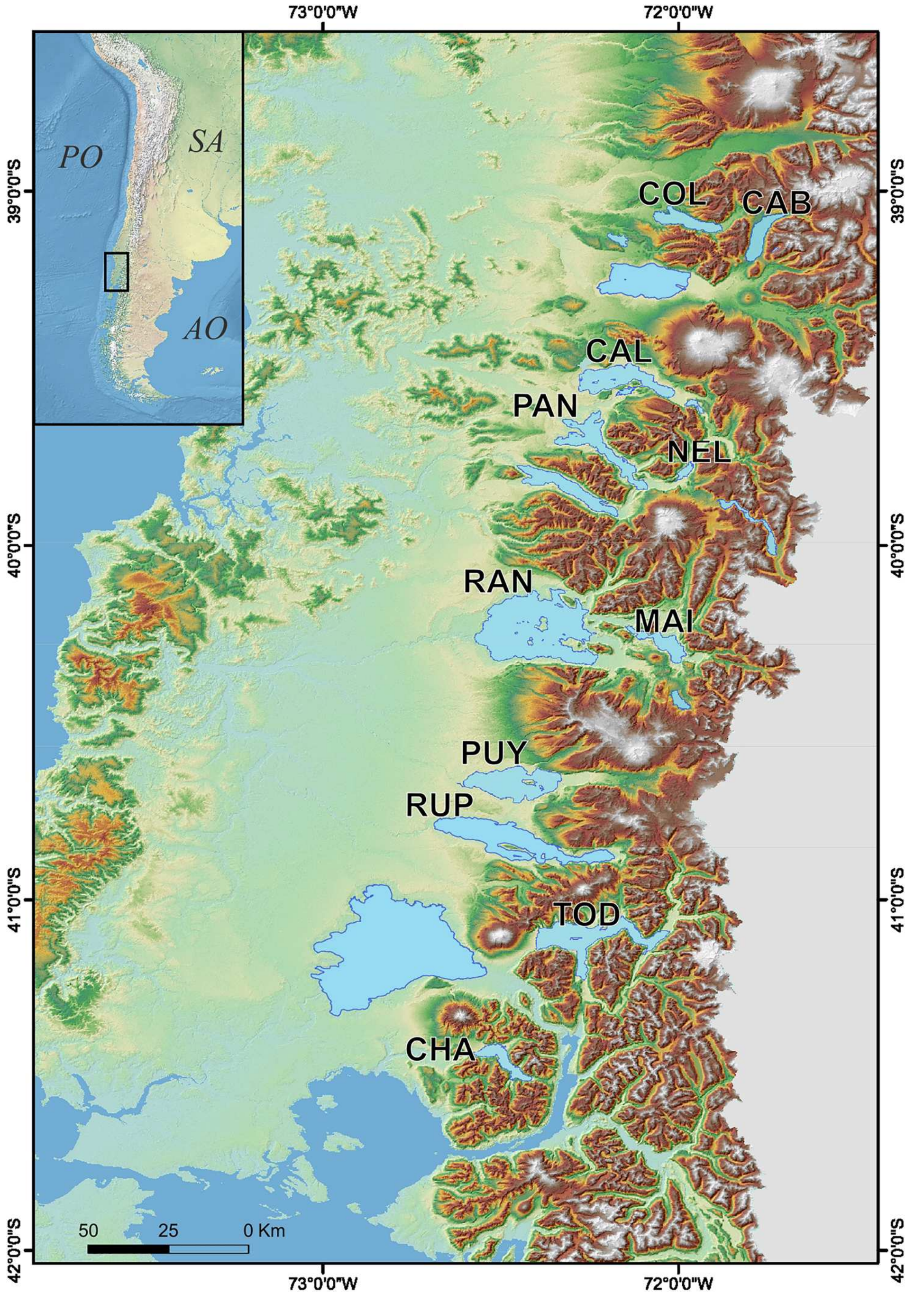

Figure 1. Study area with the geographical location of the North Patagonian lakes. The 11 lakes included in the random forest models are named by acronym: CAB (Caburga), CAL (Calafquen), CHA (Chapo), COL (Colico), MAI (Maihue), NEL (Neltume), PAN (Panguipulli), PUY (Puyehue), RAN (Ranco), RUP (Rupanco), and TOD (Todos los Santos). Displayed at the left uppermost corner is the broad geographical location of the study area, naming SA (South America), PO (Pacific Ocean) and AO (Atlantic Ocean).

model performance with and without added geographical variables, we conducted 500 iterations of random subsampling for the partition scheme on each dataset. For each iteration of subsampling, we trained separate RF models and subsequently calculated the Spearman's *R* between the predictions and the observations for each test data subsample. This approach allowed us to obtain a robust evaluation of the RF models' performance, considering the impact of random selection of observations for training and validation. By conducting multiple iterations of random subsampling and calculating Spearman's *R* for each iteration, we could draw meaningful conclusions about the efficacy of incorporating geographical information in the models' predictions of Chl-*a* concentrations.

Similarly, mean performance was calculated for the datasets including samples at WSD of 0 to 50 m (cut off at approximate euphotic zone) and 0 to 250 (*Z*max), also conducting 500 iterations of random subsampling for each dataset, allowing us to precisely conclude if samples from greater depths contribute to model performance and specially to the overall objective of efficient environmental monitoring for government institutions. To account for the reduction of samples available for modeling when cutting WSD to 50 m as maximum, which could potentially decrease performance, a new dataset was constructed retaining the few samples available with WSD >50 and only reducing the number of the broadly available samples with WSD <50 m, thus equaling the number of samples for both datasets. To further explore the decrease in performance by reduced sampling, subsequent reduced datasets (with 90%, 80%, 70%, 60%, and 50% of data points) were produced by randomly subsampling the full dataset; the performance of the RF modeling was then analyzed in the same manner as the previous tests.

For RF interpretation and explanation, we employed as a measure of variable importance the percentage increase in mean square error (%incMSE) using package *randomForest* v4.7-1.1. The %incMSE is the increase in the mean squared error in the model as the values of the observations for a particular variable are randomly permuted and changed, essentially destroying the association between predictor and response (Breiman 2001). Additionally, we estimated the marginal effects of each predictor on Chl-*a* using R package *pdp* v0.8.1 and plotted them as partial dependence plots (PDP; Greenwell 2017). Briefly, PDP are low dimensional graphics (a selected predictor against the response variable) rendered by fixing all other variables and gradually changing the targeted predictor so that its marginal effect over the response variable emerges (Greenwell 2017). A potential pitfall for this method might occur if predictors are correlated to some degree, so that a single variable PDP may not show insightful results. In our study, marginal effect plots were produced to observe the functional relationship between each predictor and Chl-*a*. Throughout the analyses, data management tasks were conducted using the functionalities provided by R package *dplyr* v1.1.3 (Wickham et al. 2023), and plots were generated using R package *ggplot2* v3.4.3 (Wickham 2009).

Additionally, in this study we evaluated lake trophic status based on the classification and thresholds proposed by Smith et al. (1999), a method that combines parameters such as concentrations of TN, TP, Chl-*a*, and Transp separately for lakes.

Results

After filtering, the original dataset of 8371 data points, including 18 lakes, was reduced to 1174 reliable observations for 11 lakes, which we used for RF modeling, specifically centered on 11 lakes and 22 years of data (from 1997 to 2019; for mean seasonal values of the main limnological variables see Table 1). RF modeling considered the analysis of 2 databases: DGA-only (only physicochemical data of 11 lakes) and Geographical-added (physicochemical data plus geographical and morphometric variables). For both datasets, a 10-fold cross-validation consistently showed that using 4 randomly selected predictors (mtry = 4) was the best hyperparameter for the modeling process (Fig. 2a). Additionally, the model convergence was achieved with 2000 trees (ntree = 2000; Fig. 2b).

Model performance

After 500 validation iterations of subsampling, the DGA-only RF model achieved a mean Spearman's *R* of 0.752 between observed and predicted Chl-*a* values (Fig. 3a). The model revealed high prediction performance for Chl-*a* values at middle and higher values of concentration but tended to underestimate predictions towards lower Chl-*a* values (Fig. 3b and 4). By contrast, the RF model produced with the Geographical-added dataset exhibited a slight but noticeable improvement in Chl-*a* prediction performance, with a Spearman's *R* of 0.775 (Fig. 3). In the same sense, this model displayed a better fit of Chl-*a* predictions throughout all lakes for lower and higher values (Fig. 3b and 4) than the DGA-only model, further proving that geographical attributes are relevant for modeling.

The test of performance on different WSD cuts (i.e., the RF model with the full range of WSD against a model excluding samples with WSD >50 m) revealed that overall performance was not impacted when excluding samples from greater depths. Specifically, the performance of the RF model including samples across all WSD showed a mean Spearman's *R* of 0.766 while the model excluding samples with WSD >50 m displayed a mean Spearman's *R* of 0.763 (Supplemental Fig. S3). When additionally testing for the effect of reduction in sample availability for modeling, results showed a clear reduction of performance with fewer available

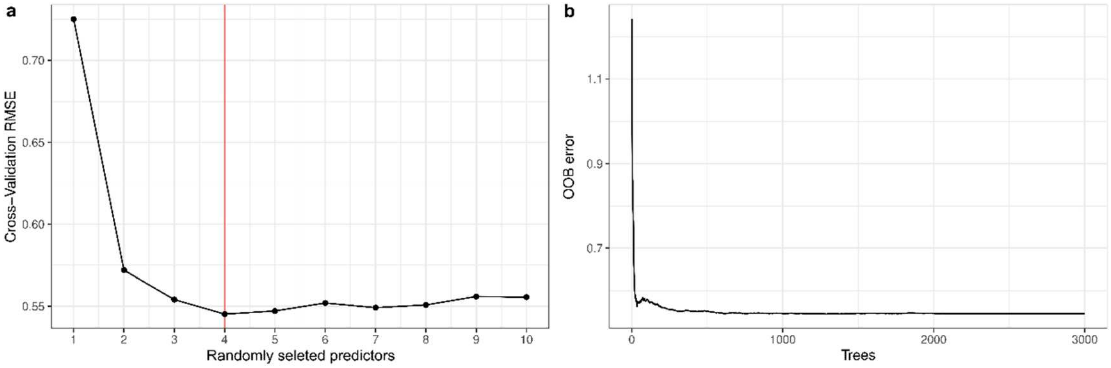

Figure 2. (a) Hyperparameter tuning for the number of randomly selected predictors (mtry) using 10-fold cross-validation, the red vertical line indicates the selected value for mtry. (b) Convergence analysis of the number of trees through the method of minimizing out-of-bag (OOB) error.

samples, revealing that the full model would be further improved with the addition of more quality samples (Supplemental Fig. S4).

Variable importance ranking

First, the importance of variables as measured in %IncMSE showed that the addition of variables in the Geographical-added model didn't modify the relative importance of the first model with the DGA-only variables, but instead were intertwined to these with a varying impact on Chl-*a* prediction; therefore, we show the results of importance for all variables in the Geographical-added RF model (Fig. 5). We were able to identify a set of variables with higher Chl-*a* predictive power through %incMSE, in descending order (see Fig. 5): DO, Temp, Transp, EC, and pH. The location of the measuring stations (latitude and longitude) proved relevant in training the model. Additionally, TN, TP, and WSD had comparatively medium importance. Some geographical and morphometric variables (LA, WA, coordinates, and LV) were also of medium importance, with values near those for TN and TP. The categories of sampling season, together with some geographical variables such as *Zmax* altitude and *Zeuf*/*Zmax*, had the lowest importance for prediction.

Partial dependence plot

By parsimony, PDP results (Fig. 6) depict selected variables of high to medium importance (according to %IncMSE): DO, Temp, Transp, EC, pH, TN, area, TP, and watershed. Chl-*a* prediction displayed a broad diversity of functional responses across variables. The most relevant functional responses are described as follows. (1) The isolated effect of DO shows no increase in Chl-*a* at lower concentrations and then a logistic growth from a concentration of 10 to 13 mg L^{-1}, related to an increment of Chl-*a* concentration from 1.0 to 1.6 µg L^{-1}. (2) Using Temp as a Chl-*a* predictor, the maximum values occur at Temp ~10 °C, the typical temperature for lakes completely mixed during winter, coinciding with the lowest Transp. When Temp increases above 10 °C, the predicted Chl-*a* sharply declines, with minimum Chl-*a* values up to 20 °C. The observed relation of Transp follows a negative, smoother trend, with an important decrease of Chl-*a* from 1.4 to 1.0 µg L^{-1} after an increment of Transp from 5 to 25 m. Regarding EC, the model highlights higher Chl-*a* values (20–40 µg L^{-1}) predicted at the lowest range of EC, displaying more homogeneous predicted values of Chl-*a* (1.2 µg L^{-1}) along with increasing EC up to 80 µS/cm. pH displays higher concentration values towards more alkaline values. Note also the effect of TN and TP, relevant trophic variables that limit phytoplankton production of lakes. In the case of TN, PDP exhibits a logistic increasing trend in a predicted Chl-*a* in a TN range of 0.2 to 0.6 mg L^{-1}; for TP, the Chl-*a* increase starts with minimal concentrations of almost 0.01 mg L^{-1}, with high increments of Chl-*a* from 0.03 to 0.04 mg L^{-1}. (3) Chl-*a* variation in relation to WSD shows a pattern of minimal Chl-*a* concentration at minimal depth, increasing with depth until reaching a maximum of 1.17 µg L^{-1} between 10 to 20 m WSD, decreasing at 30 m WSD, and then increasing again at higher depths.

Discussion

From a comprehensive dataset spanning 43 years across 18 lakes ($n = 8374$), a limited 14% subset (1174 observations) proved viable to include in the RF modeling. The marked disparity in the quality of the data across lakes and years of sampling highlights an urgent need to overhaul the limnological monitoring protocols for Chilean

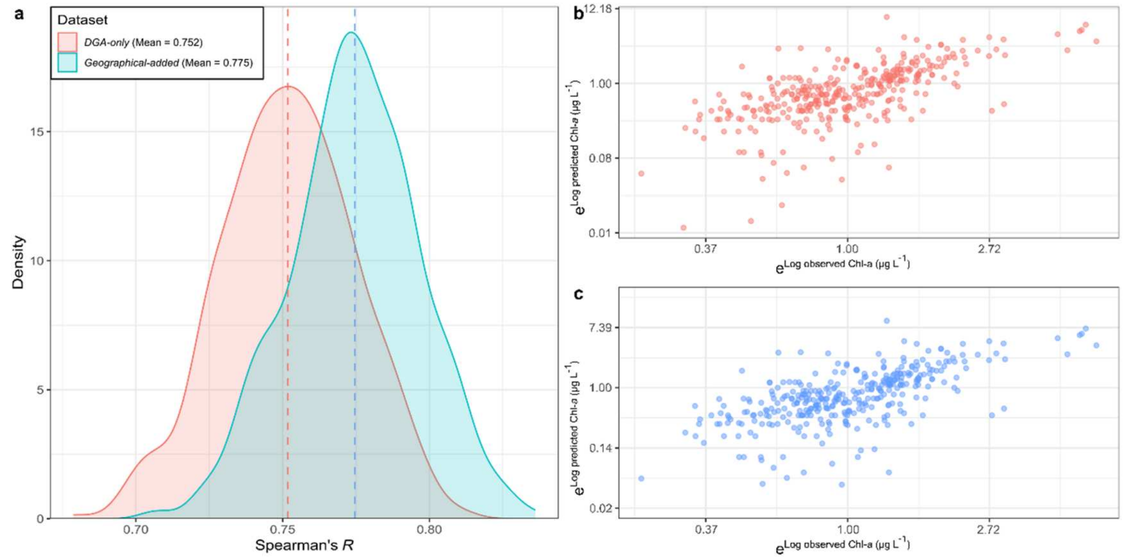

Figure 3. (a) Density plot for the distribution of Spearman's R coefficient values obtained through 500 random forest (RF) iterations of subsampling and modeling on each dataset: in red "DGA-only" and in blue "Geographical-added." Vertical dashed lines represent the mean value for each distribution (red and blue, respectively). (b–c) Scatter plots of observed versus predicted values for each subset of test data representing the mean performance of RF models for each model: (b) DGA-only and (c) Geographical-added. The density plot shows a slight but noticeable increase in performance with the geographical and morphological variables added to the model, further assessed on a more compact distribution of observed versus predicted data points.

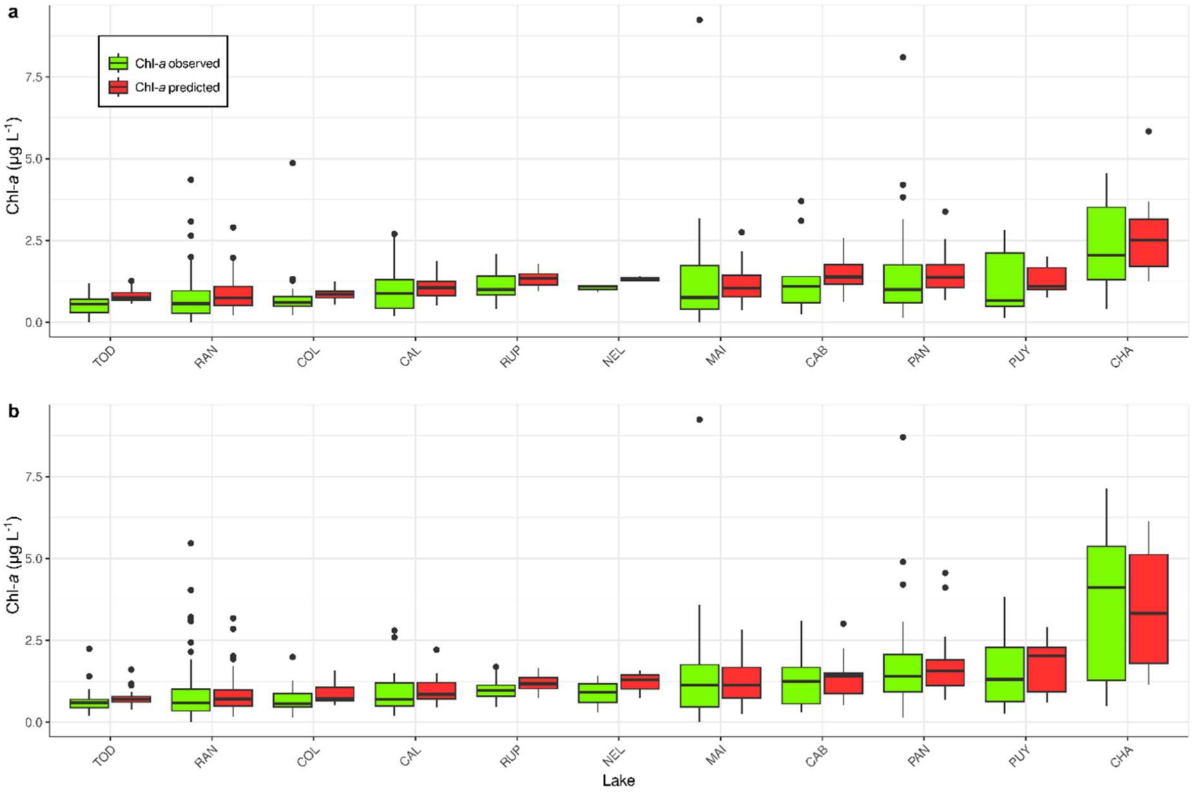

Figure 4. Boxplots of observed and predicted Chl-*a* resulting from both models: (a) DGA-only dataset model and (b) Geographical-added dataset model. The studied lakes are ordered according to increasing mean Chl-*a* content as retrieved from the predictions on the Geographical-added model. TOD (Todos los Santos), RAN (Ranco), COL (Colico), CAL (Calafquen), RUP (Rupanco), NEL (Neltume), MAI (Maihue), CAB (Caburga), PAN (Panguipulli), PUY (Puyehue), CHA (Chapo).

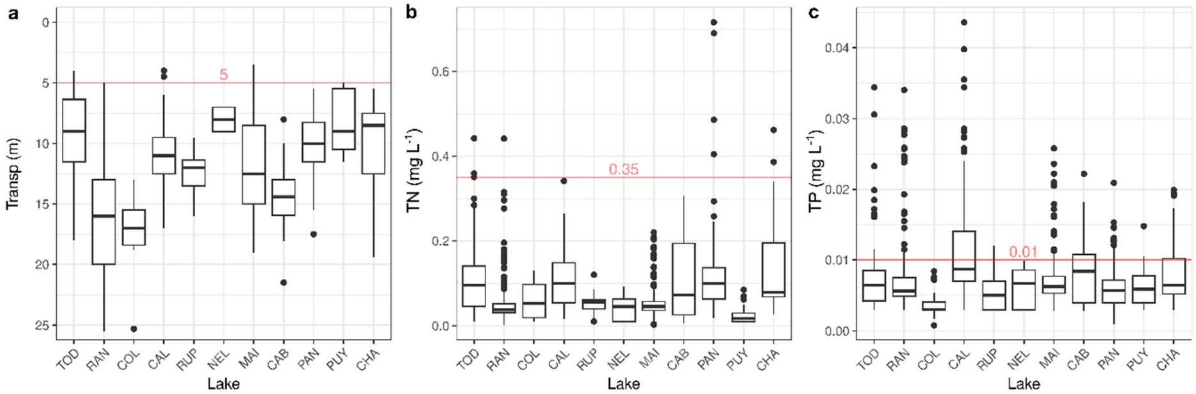

Figure 5. Boxplots of variation of trophic state parameters: (a) Transp (transparency as Secchi depth), (b) TN (total nitrogen), and (c) TP (total phosphorus) according to Smith et al. (1999). The studied lakes are ordered according to increasing mean Chl-*a* content as retrieved from the predictions on the Geographical-added model: TOD (Todos los Santos), RAN (Ranco), COL (Colico), CAL (Calafquen), RUP (Rupanco), NEL (Neltume), MAI (Maihue), CAB (Caburga), PAN (Panguipulli), PUY (Puyehue), CHA (Chapo). Red horizontal lines depict the thresholds for each parameter characterizing oligotrophic conditions.

lakes, embracing standardized criteria to optimize the acquisition of data for enduring environmental monitoring. An extensive scrutiny of data processing revealed substantial information loss stemming from a dual source: the pronounced heterogeneity (e.g., station numbers, WSD, and sampling frequency) and the incomplete nature concerning database variables historically endorsed by the Dirección General de Aguas (DGA; Supplemental Fig. S2). Notably, a considerable fraction of the excluded data pertained to trophic state parameters encompassing both total and dissolved nutrient concentrations (TN, TP, NO_2, PO_4) within the lakes. For instance, numerous observations were plagued by a lack of data for PO_4 and dissolved nitrogen fractions such as NH_4, NO_2, and NO_3 (Supplemental Fig. S2). Alarmingly, in several cases these fractions exhibited values surpassing aggregate concentrations, thereby unmasking potential analytical laboratory inconsistencies. This data deficit is a pivotal loss for comprehending eutrophication patterns within lakes on a regional scale subjected to the pressures of anthropogenic intervention and climate change repercussions.

Despite the constrained historical dataset for the studied lakes, the implemented ML algorithm proficiently captured the overarching patterns in water quality variation and trophic state across lakes and at a regional scale. Particularly, RF modeled Chl-*a* concentrations with an improved performance when combining limnological data with added geographical and morphological parameters, coinciding with RF results for Northern Hemisphere lakes, where incorporating geographical variables greatly improved performance (see Hollister et al. 2016). Additionally, recently Huovinen et al. (2019) demonstrated in Lake Panguipulli that remote sensing data (from Landsat and Sentinel products) show a close similarity within in situ temperature, turbidity, and chlorophyll. Hence, the pursuit of a deep assessment, standardization, and integration of geographical and spatial variables in the environmental monitoring of Northern Patagonian lakes is crucial for improving the precision of diagnosis and enhancing adaptive strategies for water resource management, taking advantage of modern remote sensing products such as the Harmonized Landsat and Sentinel-2 (Claverie et al. 2018).

Our results emphasize the contribution of specific limnological variables as good predictors of Chl-*a*, necessitating their rigorous measurement to optimize environmental monitoring of lake trophic state in the long term. These variables include the physicochemical variables DO, Transp, Temp, EC, and pH, which have high predictive power for Chl-*a* concentrations. Nevertheless, major nutrients (TN and TP) also play a significant role in predicting Chl-*a*. Our study mainly included oligotrophic lakes characterized by low N concentration, mainly as organic N, with a minimal deposition of atmospheric N compared with Northern Hemisphere lakes (Rogora et al. 2008). A comparative revision of trophic status characterization, using the Smith et al. (1999) framework encompassing parameters such as Transp, TN, and TP, reaffirmed the prevalence of oligotrophic patterns across most lakes (Fig. 7). According to Rogora et al. (2008) and Diaz et al. (2007), N deficiency in Northern Patagonian lakes is the main factor explaining their oligotrophic conditions where high transparency of lakes is characteristic (Woelfl 2007). Finally, as shown in the partial dependence plots (PDP), these lakes respond to small changes in TN and TP at low concentrations, which resonates with the acknowledged co-limitation of primary productivity by TN and TP (Steinhart et al. 1999, Diaz et al. 2007).

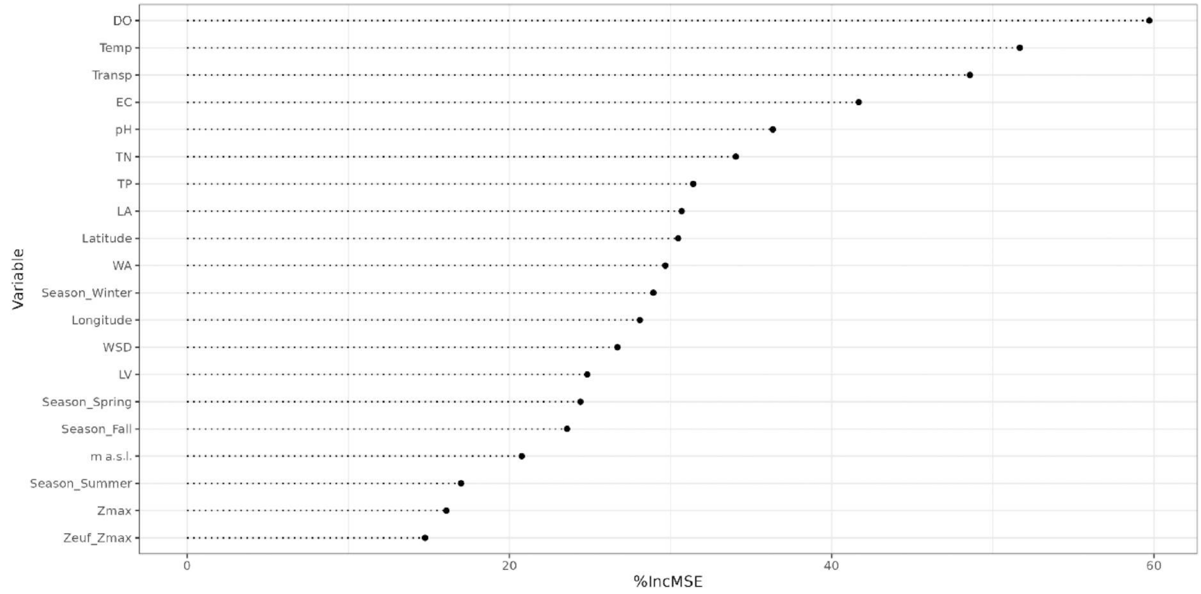

Figure 6. Ranking of relative importance of all variables (Geographical-added model) according to %IncMSE. Variables arranged by their importance score: dissolved oxygen (DO), water temperature (Temp), Secchi disk transparency (Transp), electric conductivity (EC), hydrogen potential (pH), total nitrogen (TN), total phosphorus (TP), lake area (LA), latitude (Latitude), watershed area (WA), winter (Season_Winter), longitude (Longitude), water sample depth (WSD), lake volume (LV), spring (Season_Spring), fall (Season_Fall), altitude (m a.s.l.), summer (Season_Summer), maximum depth (Zmax) and euphotic/maximum depth ratio (Zeuf_Zmax).

Moreover, the PDP analysis aptly delineated the regional dynamics of Nord Patagonian lakes, capturing the functional relationship of DO, Temp, and Transp with Chl-*a*. These lakes showcase recurrent surges in Chl-*a* biomass, chiefly attributed to diatom taxa such as *Melosira* and *Aulacoseira* during the cold mixing phases of winter (Campos 1984, Reynolds 2006, Diaz et al. 2007), which could explain the increment of Chl-*a* at lower Temp showcased in the PDP. This ecological pattern thrives under optimal conditions of

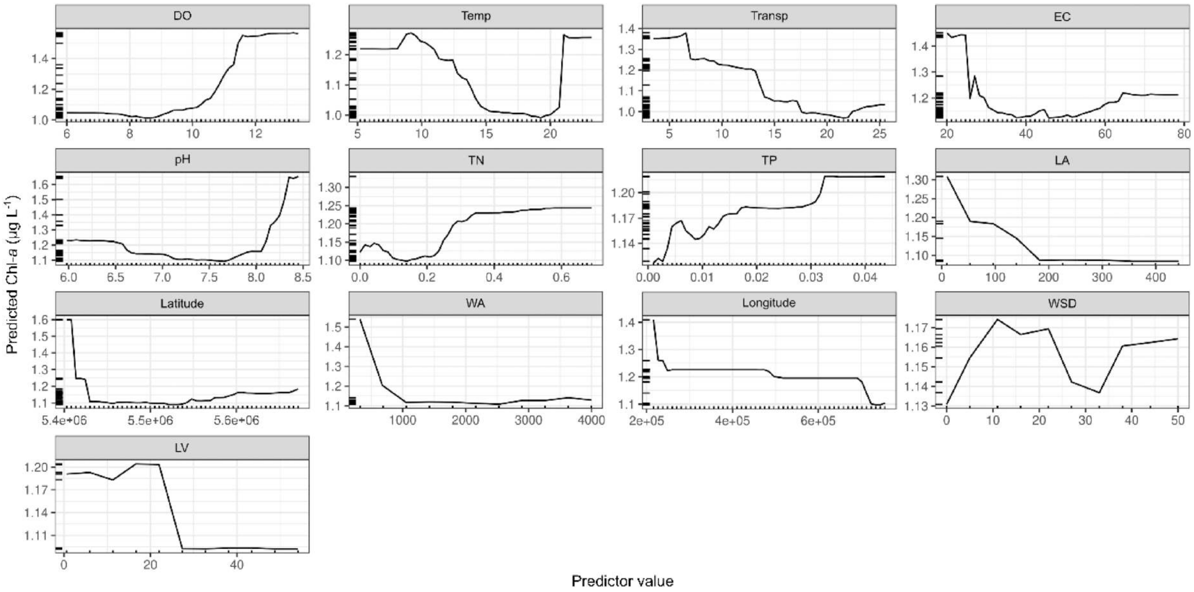

Figure 7. Partial dependence plots (PDP) based on the Geographical-added model displaying the highest 13 ranked variables from the %IncMSE (>25) (see Fig. 6). In decreasing order from top left to bottom right, calculated from the Geographical-added model. Tick marks in the y-axis represent individual values for observed Chl-*a* in the test dataset used to construct PDP. Variables arranged by their importance score: dissolved oxygen (DO), water temperature (Temp), Secchi disk transparency (Transp), electrical conductivity (EC), hydrogen potential (pH), total nitrogen (TN), total phosphorus (TP), lake area (LA), latitude (Latitude), watershed area (WA), longitude (Longitude), water sample depth (WSD), lake volume (LV).

lower temperatures and heightened turbidity during mixing events (Queimaliños and Diaz 2014). Furthermore, the analysis of Chl-*a* against WSD showed the most substantial Chl-*a* increments occurring at depths between 0 and 40 m. This Chl-*a* distribution pattern seems to correspond to the establishment of the deep chlorophyll maximum, extensively recognized in the stratified Nord Patagonic lakes during summer (Perez 2002, Modenutti et al. 2013, Queimaliños and Diaz 2014). Southern Hemisphere lakes, particularly those located in Northern Patagonia, are well known for low phytoplankton biomass production and the development of considerably deeper epilimnetic strata compared to lakes in the Northern hemisphere (Soto 2002). This characteristic acts as an essential physical driver, exerting selective pressure on the morphofunctional attributes of phytoplankton such as pennates and central chain-forming diatoms whose development is favored in times of thermal circulation (Reynolds et al. 1986, Reynolds 2006). Thus, these lakes usually exhibit an increasing trend of phytoplankton biomass during winter periods when transparency usually decreases, in contrast to warm summer periods when small centric diatoms such as *Cyclotella* sp. usually dominate phytoplankton biomass (Balseiro et al. 1997).

The results of this study highlight that the predictive power of Chl-*a* is essentially the same whether we consider all water samples by WSD (0–250 m) or restrict sampling to the euphotic layer (for these lakes in the 0–50 m range; Van de Vyver et al. 2016). The minimal change in performance (a Spearman's *R* of 0.763 for WSD <50 m in contrast to 0.766 for all depths) indicates that samples from greater depths are not particularly informative when modeling Chl-*a* (Supplemental Fig. S3). Moreover, analysis of RF performance with subsequent reduced datasets highlights that prediction precision is limited by the number of samples retained (Supplemental Fig. S4). This statistical evidence suggests that lake monitoring could be optimized by focusing on the long-term monitoring of the euphotic layer (0–50 m). Collecting water samples at greater depths (>50 m) has logistic and practical difficulties, especially in these large and deep lakes, considering frequent heavy rain and strong winds (e.g., "Puelche"), implying increased costs of operation.

In summary, our modeling of Chl-*a* concentration with RF shows the overall good predictive performance of a restricted set of physicochemical variables (DO, Transp, Temp, EC, pH, TN, and TP). Based on our results, monitoring should favor their rigorous and consistent measurement over laboratory analysis of multiple but inconsistent sets of variables found over the databases explored, thus requiring less government funding and optimizing the availability of quality samples. This recommendation is particularly relevant considering that results show the performance of prediction could be improved just with the addition of more data points. Furthermore, our results show that monitoring could be optimized by consistently targeting the euphotic zone (0–50 m), thus reducing costs and complexity of sampling campaigns. Finally, the incorporation of geographical variables and remote sensing emerges as an ideal trajectory to increment monitoring capacity in terms of precision and coverage. This integration holds the promise of bolstering the accuracy and comprehensiveness of lake monitoring endeavors, enhancing our ability to decipher the intricate dynamics of these critical aquatic ecosystems and develop coherent environmental policies to secure water quality.

Conclusions

The demonstrated predictive capacity of the RF model for Chl-*a* concentrations, primarily reliant on the variables of DO, transparency/Secchi depth, temperature, and electrical conductivity (EC), followed by pH and nutrient concentrations like TN and TP, is presented as a benchmark for the optimization of Northern Patagonian lakes monitoring strategies. By adhering to this essential repertoire of key variables during Chl-*a* predictions and the insights presented in our study, as in the relation of WSD and number of observations with the predictive performance, the efficacy of regional lake monitoring could be significantly fortified. Decreasing costs and sampling efforts has the potential benefit of freeing resources to increase the number and density of observations, thus greatly increasing predicting capability with future monitoring efforts. Our study underscores a compelling directive to focus efforts of monitoring towards an in-depth exploration of the euphotic zone in lakes (0–50 m), aiming to attain a consistent dataset across all lakes, especially for this upper layer of biological importance. Furthermore, we provide evidence that supports the incorporation of geographical and remote sensing information as an ideal trajectory for future lake monitoring initiatives and overall improvement. This approach, locally implemented effectively in previous research (e.g., Huovinen et al. 2019), not only facilitates model calibration but also confers higher levels of precision and greatly increases the spatial and temporal coverage of the monitoring effort. This integration holds the promise of bolstering the accuracy and comprehensiveness of lake monitoring endeavors, enhancing our ability to decipher the intricate dynamics of these critical aquatic ecosystems and develop coherent environmental policies to warranted water security.

Acknowledgements

The authors thank Dr Carla Olmo for her constructive criticism in the initial version of the paper and the two anonymus reviewers. Cristian Ríos Molina acknowledges support from BECA DE DOCTORADO NACIONAL (Año 2022 – Folio 21221749), ANID – Subdirección de Capital Humano (Chile). Luciano Caputo acknowledges the proyecto InES I+D INID210009. The authors dedicate this study to Marine Biologist Ana Ester Lorca Torres (QEPD), a highly esteemed colleague and friend whose career in freshwater quality analysis made significant contributions to research at the Universidad Austral de Chile and to Chilean Limnology in general.

Disclosure statement

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sector.

Author contribution statement

LC: conceptualization, investigation, data curation, methodology, writing. CRM: data curation, formal analysis, methodology, software, validation, visualization, writing. RAA: data curation, investigation, writing. IFB: methodology, supervision, writing.

ORCID

Cristian Rios Molina http://orcid.org/0000-0002-5641-6836
Iván Felipe Benavides http://orcid.org/0000-0002-1139-3909

References

Alaniz AJ, Smith-Ramírez C, Rendón-Funes A, Hidalgo-Corrotea C, Carvajal MA, Vergara PM, Fuentes N. 2022. Multiscale spatial analysis of headwater vulnerability in South-Central Chile reveals a high threat due to deforestation and climate change. Sci Total Environ. 849:157930.

Balseiro EG, Modenutti BE, Queimaliños CP. 1997. Nutrient recycling and shifts in N:P ratio by different zooplankton structures in a south Andes lake. J Plankton Res. 19:805–817.

Brêda JPLF, De Paiva RCD, Collischon W, Bravo JM, Siqueira VA, Steinke EB. 2020. Climate change impacts on South American water balance from a continental-scale hydrological model driven by CMIP5 projections. Clim Change. 159(4):503–522.

Breiman L. 2001. Random forests. Mach Learn. 45(1):5–32.

Campos H. 1984. Limnological study of Araucanian lakes (Chile). SIL Proceedings, 1922–2010. 22(2):1319–1327.

Caputo L, Huovinen P, Sommaruga R, Gómez I. 2018. Water transparency affects the survival of the medusa stage of the invasive freshwater jellyfish *Craspedacusta sowerbii*. Hydrobiologia. 817(1):179–191.

Claverie M, Ju J, Masek JG, Dungan JL, Vermote EF, Roger JC, Skakun SV, Justice C. 2018. The harmonized Landsat and Sentinel-2 surface reflectance data set. Remote Sens Environ. 219:145–161.

Derot J, Yajima H, Jacquet S. 2020. Advances in forecasting harmful algal blooms using machine learning models: a case study with *Planktothrix rubescens* in Lake Geneva. Harmful Algae. 99:101906.

Diaz M, Pedrozo F, Reynolds C, Temporetti P. 2007. Chemical composition and the nitrogen-regulated trophic state of Patagonian lakes. Limnologica. 37(1):17–27.

Dirección General de Aguas (DGA). 2014. Evaluación de la condición trófica de la red de control de lagos de la DGA [Evaluation of the condition of the DGA lake control network]. Santiago: Gobierno de Chile. Spanish.

Echeverría C, Newton A, Nahuelhual L, Coomes D, Rey-Benayas JM. 2012. How landscapes change: integration of spatial patterns and human processes in temperate landscapes of southern Chile. Appl Geogr. 32(2):822–831.

Felip M, Catalan J. 2000. The relationship between phytoplankton biovolume and chlorophyll in a deep oligotrophic lake: decoupling in their spatial and temporal maxima. J Plankton Res. 22(1):91–106.

Fierro P, Bertrán C, Tapia J, Hauenstein E, Peña-Cortés F, Vergara C, Cerna C, Vargas-Chacoff L. 2017. Effects of local land-use on riparian vegetation, water quality, and the functional organization of macroinvertebrate assemblages. Sci Total Environ. 609:724–734.

Fuentealba M, Latorre C, Frugone-Álvarez M, Sarricolea P, Godoy-Aguirre C, Armesto J, Villacís LA, Laura Carrevedo M, Meseguer-Ruiz O, Valero-Garcés B. 2021. Crossing a critical threshold: accelerated and widespread land use changes drive recent carbon and nitrogen dynamics in Vichuquén Lake (35°s) in central Chile. Sci Total Environ. 791:148209.

Geller W. 1992. The temperature stratification and related characteristics of Chilean lakes in midsummer. Aquat Sci. 54(1):37–57.

Greenwell BM. 2017. pdp: an r package for constructing partial dependence plots. R J. 9(1):421.

Gutiérrez JS, Moore JN, Donnelly JP, Dorador C, Navedo JG, Senner NR. 2022. Climate change and lithium mining influence flamingo abundance in the Lithium Triangle. Proc R Soc B. 289(1970):2021–2388.

Han T, Jiang D, Zhao Q, Wang L, Yin K. 2018. Comparison of random forest, artificial neural networks and support vector machine for intelligent diagnosis of rotating machinery. Trans Inst Meas Control. 40(8):2681–2693.

Hébert M, Symons CC, Cañedo-Argüelles M, Arnott SE, Derry AM, Fugère V, Hintz WD, Melles SJ, Astorg L, Baker HK, et al. 2023. Lake salinization drives consistent losses of zooplankton abundance and diversity across coordinated mesocosm experiments. Limnol Oceanogr Lett. 8(1):19–29.

Herrera C, Urrutia J, Gamboa C, Salgado X, Godfrey L, Rivas A, Jódar J, Custodio E, León C, Sigl V, et al. 2023. Evaluation of the impact of the intensive exploitation of groundwater and the mega-drought based on the hydrochemical and isotopic composition of the waters of the Chacabuco-Polpaico basin in central Chile. Sci Total Environ. 895:165055.

Hidalgo-Corrotea C, Alaniz AJ, Vergara PM, Moreira-Arce D, Carvajal MA, Pacheco-Cancino P, Espinosa A. 2023. High vulnerability of coastal wetlands in Chile at multiple scales derived from climate change, urbanization, and exotic forest plantations. Sci Total Environ. 903:166130.

Ho JC, Michalak AM, Pahlevan N. 2019. Widespread global increase in intense lake phytoplankton blooms since the 1980s. Nature. 574(7780):667–670.

Hollister JW, Milstead WB, Kreakie BJ. 2016. Modeling lake trophic state: a random forest approach. Ecosphere. 7(3): e01321.

Huovinen P, Ramírez J, Caputo L, Gómez I. 2019. Mapping of spatial and temporal variation of water characteristics through satellite remote sensing in Lake Panguipulli, Chile. Sci Total Environ. 679:196–208.

[IPCC] Intergovernmental Panel on Climate Change. 2023. Climate change 2021 – the physical science basis: Working Group I contribution to the sixth assessment report of the Intergovernmental Panel on Climate Change. 1st ed. Cambridge (UK): Cambridge University Press.

Jane SF, Hansen GJA, Kraemer BM, Leavitt PR, Mincer JL, North RL, Pilla RM, Stetler JT, Williamson CE, Woolway RI, et al. 2021. Widespread deoxygenation of temperate lakes. Nature. 594(7861):66–70.

Janse JH, Kuiper JJ, Weijters MJ, Westerbeek EP, Jeuken MHJL, Bakkenes M, Alkemade R, Mooij WM, Verhoeven JTA. 2015. GLOBIO-Aquatic, a global model of human impact on the biodiversity of inland aquatic ecosystems. Environ Sci Policy. 48:99–114.

Kuhn M. 2008. Building predictive models in R using the caret package. J Stat Soft. 28(5).

León-Muñoz J, Echeverría C, Marcé R, Riss W, Sherman B, Iriarte JL. 2013. The combined impact of land use change and aquaculture on sediment and water quality in oligotrophic Lake Rupanco (North Patagonia, Chile, 40.8°S). J Environ Manag. 128:283–291.

Liaw A, Wiener M. Classification and regression by randomForest. R News. 2(3):18–22.

Meerhoff M, Audet J, Davidson TA, De Meester L, Hilt S, Kosten S, Liu Z, Mazzeo N, Paerl H, Scheffer M, Jeppesen E. 2022. Feedback between climate change and eutrophication: revisiting the allied attack concept and how to strike back. Inland Waters. 12(2):187–204.

Merz E, Saberski E, Gilarranz LJ, Isles PDF, Sugihara G, Berger C, Pomati F. 2023. Disruption of ecological networks in lakes by climate change and nutrient fluctuations. Nat Clim Chang. 13(4):389–396.

[MMA] Ministerio del Medio Ambiente. 2020. Sexto Informe Nacional de Biodiversidad de Chile [Sixth National Biodiversity Report of Chile]. Santiago (Chile): Ministerio del Medio Ambiente. Informe elaborado en el marco del Convenio sobre la Diversidad Biológica. Ministerio del Medio Ambiente de Chile. Spanish.

Modenutti B, Balseiro E, Bastidas Navarro M, Laspoumaderes C, Souza MS, Cuassolo F. 2013a. Environmental changes affecting light climate in oligotrophic mountain lakes: the deep chlorophyll maxima as a sensitive variable. Aquat Sci. 75(3):361–371.

Navedo JG, Vargas-Chacoff L. 2021. Salmon aquaculture threatens Patagonia. Science. 372(6543):695–696.

Nimptsch J, Woelfl S, Osorio S, Valenzuela J, Ebersbach P, Von Tuempling W, Palma R, Encina F, Figueroa D, Kamjunke N, Graeber D. 2015. Tracing dissolved organic matter (DOM) from land-based aquaculture systems in North Patagonian streams. Sci Total Environ. 537:129–138.

Nimptsch J, Woelfl S, Osorio S, Valenzuela J, Moreira C, Ramos V, Castelo-Branco R, Leão PN, Vasconcelos V. 2016. First record of toxins associated with cyanobacterial blooms in oligotrophic North Patagonian lakes of Chile – a genomic approach: cyanotoxins in North Patagonian lakes. Int Rev Hydrobiol. 101(1–2):57–68.

Perez GL. 2002. Light climate and plankton in the deep chlorophyll maxima in North Patagonian Andean lakes. J Plankton Res. 24(6):591–599.

Pizarro J, Vergara PM, Cerda S, Briones D. 2016. Cooling and eutrophication of southern Chilean lakes. Sci Total Environ. 541:683–691.

Pizarro R, Garcia-Chevesich PA, McCray JE, Sharp JO, Valdés-Pineda R, Sangüesa C, Jaque-Becerra D, Álvarez P, Norambuena S, Ibáñez A, et al. 2022. Climate change and overuse: water resource challenges during economic growth in Coquimbo, Chile. Sustainability. 14(6):3440.

Queimaliños C, Diaz M. 2014. Phytoplankton of Andean Patagonian lakes. Adv Limnol. 65:235–256.

R Core Team. 2023. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing

Reynolds CS. 2006. Ecology of phytoplankton. Cambridge (UK): Cambridge University Press.

Reynolds CS, Montecino V, Graf ME, Cabrera S. 1986. Short-term dynamics of a *Melosira* population in the plankton of an impoundment in central Chile. J Plankton Res. 8(4): 715–740.

Rodríguez-López L, Duran-Llacer I, Bravo Alvarez L, Lami A, Urrutia R. 2023. Recovery of water quality and detection of algal blooms in Lake Villarrica through Landsat satellite images and monitoring data. Remote Sens. 15(7): 1929.

Rogora M, Arese C, Balestrini R, Marchetto A. 2008. Climate control on sulphate and nitrate concentrations in alpine streams of Northern Italy along a nitrogen saturation gradient. Hydrol Earth Syst Sci. 12:371–381.

Smith V, Tilman GD, Nekola JC. 1999. Eutrophication: impacts of excess nutrient inputs on freshwater, marine, and terrestrial ecosystems. Environ Pollut. 100:179–196.

Soto D. 2002. Oligotrophic patterns in southern Chilean lakes: the relevance of nutrients and mixing depth. Rev Chil Hist Nat. 75(2):377–393.

Steinhart G, Likens GE, Soto D. 1999. Nutrient limitation in Lago Chaiquenes (Parque Nacional Alerce Andino, Chile): evidence from nutrient enrichment experiments and physiological assays. Rev Chil Hist Nat. 72:559–568.

Taranu ZE, Gregory-Eaves I, Leavitt PR, Bunting L, Buchaca T, Catalan J, Domaizon I, Guilizzoni P, Lami A, McGowan S, et al. 2015. Acceleration of cyanobacterial dominance in north temperate-subarctic lakes during the Anthropocene. Ecol. Lett. 18(4):375–384.

Tickner D, Opperman JJ, Abell R, Acreman M, Arthington AH, Bunn SE, Cooke SJ, Dalton J, Darwall W, Edwards G, et al. 2020. Bending the curve of global freshwater biodiversity loss: an emergency recovery plan. BioScience. 70(4):330–342.

Torremorell A, Hegoburu C, Brandimarte AL, Rodrigues EHC, Pompêo M, Da Silva SC, Moschini-Carlos V, Caputo L, Fierro P, Mojica JI, et al. 2021. Current and future threats for ecological quality management of South

American freshwater ecosystems. Inland Waters. 11(2):125–140.

Van de Vyver E, Van Wichelen J, Vanormelingen P, Vannieuwenhuyze W, Daveloose I, de Jong R, de Blok R, Urrutia R, Tytgat B, Verleyen E, Vyverman W. 2016. Variation in phytoplankton pigment composition in relation to mixing conditions in temperate South-Central Chilean lakes. Limnologica. 79:125715.

Virro H, Kmoch A, Vainu M, Uuemaa E. 2022. Random forest-based modeling of stream nutrients at national level in a data-scarce region. Sci Total Environ. 840:156613.

Watanabe E, Noyama S, Kiyono K, Inoue H, Atarashi H, Okumura K, Yamashita T, Lip GYH, Kodani E, Origasa H. 2021. Comparison among random forest, logistic regression, and existing clinical risk scores for predicting outcomes in patients with atrial fibrillation: a report from the J-RHYTHM registry. Clin Cardiol. 44(9):1305–1315.

Wickham H. 2009. Ggplot2: elegant graphics for data analysis. New York (NY): Springer New York.

Wickham H, François R, Henry L, Müller K, Vaughan D. 2023. dplyr: a grammar of data manipulation. R package version 1.1.4. https://github.com/tidyverse/dplyr, https://dplyr.tidyverse.org

Woelfl S. 2007. The distribution of large mixotrophic ciliates (Stentor) in deep North Patagonian lakes (Chile): first results. Limnologica. 37:28–36. do

Woolway RI, Kraemer BM, Lenters JD, Merchant CJ, O'Reilly CM, Sharma S. 2020. Global lake responses to climate change. Nat Rev Earth Environ. 1(8):388–403.

Wurtsbaugh WA, Paerl HW, Dodds WK. 2019. Nutrients, eutrophication and harmful algal blooms along the freshwater to marine continuum. WIREs Water. 6(5):6:e1373.

Yokoyama A, Yamaguchi N. 2020. Comparison between ANN and random forest for leakage current alarm prediction. Energy Reports. 6:150–157.