Article

Not peer-reviewed version

Chlorophyll-a Estimation in 149 Tropical Semi-Arid Reservoirs Using Remote Sensing and Machine Learning

Victor Oliveira Santos * , Bruna Monallize Duarte Moura Guimarães , Iran Eduardo Lima Neto , Francisco de Assis Souza Filho , Paulo Alexandre Costa Rocha , Jesse Van Griensven Thé , Bahram Gharabaghi *

Posted Date: 25 December 2023

doi: 10.20944/preprints202312.1854.v1

Keywords: Chlorophyll-a; Sentinel-2 satellite; Machine learning; Freshwater Reservoirs; Eutrophication

Article

Chlorophyll-a Estimation in 149 Tropical Semi-Arid Reservoirs Using Remote Sensing and Machine Learning

Victor Oliveira Santos [1,*], Bruna Monallize Duarte Moura Guimarães [2], Iran Eduardo Lima Neto [2], Francisco de Assis Souza Filho [2], Paulo Alexandre Costa Rocha [3], Jesse Van Griensven Thé [4] and Bahram Gharabaghi [5,*]

[1] School of Engineering, University of Guelph, 50 Stone Rd E, Guelph, ON N1G 2W1, Canada; volive04@uoguelph.ca (V.O.S.);
[2] Department of Hydraulic and Environmental Engineering, Technology Center, Federal University of Ceará, Fortaleza 60020-181, CE, Brazil; bruna.monallize@gmail.com (B.M.D.M.G.); iran@deha.ufc.br (I.E.L.N.); assis@ufc.br (F.A.S.F.)
[3] Department of Mechanical Engineering, Technology Center, Federal University of Ceará, Fortaleza 60020-181, CE, Brazil; paulo.alexandre.costa.rocha@gmail.com (P.A.C.R.);
[4] Lakes Environmental, 170 Columbia St. W, Waterloo, ON N2L 3L3, Canada; jessethe@gmail.com (J.V.G.T.)
[5] School of Engineering, University of Guelph, 50 Stone Rd E, Guelph, ON N1G 2W1, Canada; bgharaba@uoguelph.ca (B.G.)
* Correspondence: volive04@uoguelph.ca (V.O.S.); bgharaba@uoguelph.ca (B.G.)

Abstract: Eutrophication, a global concern, impacts water quality, ecosystems, and human health. It's crucial to monitor algal blooms in freshwater reservoirs, as they indicate the trophic condition of a waterbody through Chlorophyll-a (Chla) concentration. Traditional monitoring methods, however, are expensive and time-consuming. Addressing this hindrance, we developed models using remotely sensed data from the Sentinel-2 satellite for large-scale coverage, including its bands and spectral indexes, to estimate the Chla concentration on 149 freshwater reservoirs in the state of Ceará, Brazil. Several machine learning models, including k-nearest neighbours, random forests, extreme gradient boosting, the least absolute shrinkage, group method of data handling (GMDH), and support vector machine models were trained and tested. A stepwise approach determined the best subset of input parameters. The best-performing model was the GMDH, achieving an R^2 of 0.91 and RMSE of 20.38 µg/L, which is a value consistent with the ones found in the literature. Nevertheless, the predicted Chla concentration values were most sensitive to the red, green, and near infra-red bands.

Keywords: Chlorophyll-a; Sentinel-2 satellite; machine learning; freshwater reservoirs; eutrophication

1. Introduction

Eutrophication, an important manifestation of water pollution in freshwater reservoirs, results from the excessive nutrient loads that cause algal blooms [1,2]. The excessive algae and aquatic plant growth result in oxygen depletion, ultimately causing severe impacts on aquatic ecosystems and considerably increasing the costs related to water treatment [3–6].

Chlorophyll-a (Chla), a photosynthetic pigment present in major algae groups, is widely used as a key indicator of phytoplankton presence [7,8]. Because the abundance of algae can reflect the state of eutrophication, Chla is one of the most important parameters for evaluating the trophic condition of water bodies [9,10]. While Chla concentration has long served as a key parameter in monitoring harmful blooms, accurately predicting Chla in reservoirs has proven to be a persistent challenge [11,12]. This difficulty is mainly due to the non-linear and non-stationary characteristics of Chla concentration, which are influenced by anthropogenic and hydrometeorological factors [2].

In regions characterized by semi-arid climates, such as the Northeastern Brazil, establishing an extensive network of multi-purpose artificial reservoirs has emerged as a reliable solution to confront the water scarcity challenges imposed by environmental constraints [13,14]. These reservoirs are notably susceptible to eutrophication due to hydroclimatic characteristics that combined favor photosynthesis and biodegradation [15,16], such as interannual variability of precipitation and stored volume [17], high temperatures and evaporation rates [18], and prolonged hydraulic retention time [19]. Moreover, this susceptibility is further aggravated by continued anthropogenic pressure on water bodies due to internal enrichment from aquaculture practice [3,20], inadequate coverage of sanitation systems [21], and a dense reservoir network [22,23].

Therefore, regularly monitoring Chla concentrations is crucial for implementing effective water quality management strategies to prevent further deterioration [24]. However, traditional sampling methods are expensive, time-consuming, and impractical for many reservoirs [25,26]. As an attractive option, satellite remote sensing and machine learning (ML) techniques offer a cost-effective approach for monitoring Chla concentrations and their spatiotemporal variations, providing data of complex environmental systems on larger scales [27]. The Sentinel-2 constellation has been proved to be a valuable asset in monitoring inland and coastal waters, once it has improved spatial resolution compared to other freely available sensing systems data, such as Landsat 8 [28,29].

The ML approach excels in retrieving complex non-linear relationships within satellite data, capturing the underlying structure bonding the satellite data and the desired target variable [30,31]. Combining ML architectures with remote sensing data have been able to provide top-notch results in a plethora of scientific fields, such as solar irradiance forecasting [32], mapping of mineral extraction sites [33], forest fire mapping [34], and crop water stress evaluation [35]. ML and satellite data performance has also been explored for Chla monitoring. A random forest (RF) based model was developed in [36]. In their study, the authors used Sentinel-2 imagery to retrieve Chla concentrations for lake Chagan, in China. Their proposed model provided good performance in determining the Chla concentrations, also complying to the biological mechanism in lakes, offering robust results to seasonal changes.

In Cao et al. [9], the authors used Landsat-8 remote sensing data together with extreme gradient boosting tree model (XGBoost) ML to determine Chla in lakes located in China. Their approach was implemented to analyze the spatiotemporal data from 2013 to 2018, and it demonstrated satisfactory performance in identifying the Chla behavior in the study location. In Hu et al. [37], the authors developed methodologies to mitigate spectral noise in satellite data to improve the performance of ML models to estimate Chla in global oceans using remote sensing from several satellites. Their results proved that the support vector regression (SVR) was the best-performing ML approach, surpassing the traditional band-ratio models and providing reduced image noise.

The Group Method of Data Handling (GMDH) has also been applied to hydrological scenarios, including Chla estimation [38] , water quality prediction [39], and image classification for plant diseases [40]. However, there exists a gap in the knowledge regarding the usage of this approach on modeling Chla concentration using satellite data, which the present study aims to fulfill. Additionally, the present study seeks to provide a deep insight into the performance of ML paradigms when applied to a vast area containing heterogeneous reservoirs.

Given this scenario, the potential to increase algae blooms and further degradation of these aquatic systems has increased the need to study often poorly monitored reservoirs in semi-arid regions. Therefore, this study aims to estimate Chla concentration through combined techniques of remote sensing and machine learning. For this, the following specific objectives were pursued in this research:

1. A comprehensive investigation on several input parameters for Chla modeling, including all the 13 bands of the MSI onboard of Sentinel-2 constellation and 16 different spectral indexes.
2. A comprehensive analysis and characterization of all the 149 tropical reservoirs extensively spread across the state of Ceará, located in the Brazilian semi-arid region.
3. The usage of stepwise approach on parameters selection.
4. The investigation of different machine learning paradigms for modeling Chla values.
5. To fulfill the gap in the knowledge regarding the usage of the GMDH ML model for Chla modeling using remote sensed data and spectral indexes.

6. To assess the performance of several machine learning paradigms applied to a vast region containing heterogeneous reservoirs.

2. Materials and Methods

2.1. Study site location

The Brazilian semi-arid region is one of the most populated semi-arid regions in the world, with approximately 28 million inhabitants and an area that occupies 12% of the national territory [41]. The state of Ceará, inhabited by around nine million people, covers an area comparable in size to England of 150,000 km^2, where 98.6% of its territory is in the semi-arid region [42]. The region is characterized by extreme climatic events such as recurrent droughts, sporadic flooding and high inter-annual variability [43,44]. According to the Köppen classification [45], the climate is 'BSh', characterized by a mean precipitation of 750 mm per year, a potential evaporation rate of 2000 mm per year, a mean annual temperature of 31°C and negative water balances for most of the year [13,46]. Moreover, Ceará state is influenced by two distinct seasons: a rainy season from January to April, when 80% of the total precipitation occurs, and a dry season, during the rest of the year [47].

Since Ceará's water supply relies predominantly on artificial reservoirs, the state has a dense reservoir network, serving as the main source for over 90% of the region's water resources [20] and with a storage capacity of approximately 18.6 billion cubic meters. Notably, the three largest reservoirs in the state, Castanhão (6,700 hm^3), Orós (1,940 hm^3), and Banabuiú (1,600 hm^3) collectively represent approximately 55% of the total storage capacity [48].

The study area includes data from 149 monitored reservoirs distributed across the state of Ceará in twelve watersheds. The longitude, latitude, and basic information of the reservoirs are listed in Table S1 in the supplementary material. These reservoirs are mainly used for human water supply, aquaculture, fish farming, and irrigation. As a result, pollution is predominantly attributed to nonpoint sources such as livestock, agriculture, and soil erosion, and point sources originating from sewage and fish farming [49]. The trophic condition of these reservoirs fluctuates with seasonal changes and flooding or drought events [50]. Figure 1 illustrates the geographical location of Ceará within the Brazilian territory and in the semi-arid region, as well as the location of the reservoirs distributed across its area.

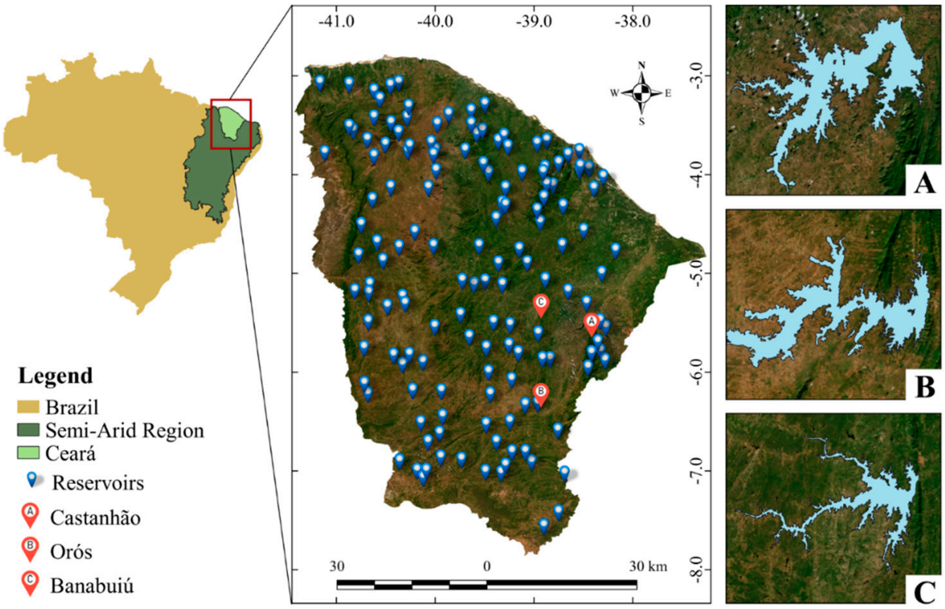

Figure 1. Geographical context of the study area and distribution of sampling points. Each blue marker represents a reservoir chlorophyll-a sampling point. Red markers indicate Ceara's largest reservoirs: A) Castanhão; B) Orós and C) Banabuiú.

2.2. Water quality data

This study used data from a total of 149 spatially distributed reservoirs (Figure 1), covering the years 2015 to 2021, including a total of 1,399 Chla samples. This information was collected from the database provided by the Portal Hidrológico do Ceará platform, a system in which monitoring has been carried out by the Water Resources Management Company (COGERH) [48], presenting consistent time series of hydrological and water quality parameters. The company carries out monitoring campaigns in all seasons of the year through quarterly sampling campaigns and in situ measurements of parameters. The Chla concentration data provided by the system was obtained through in-situ water samples collected at each reservoir sampling point at 0.3 meters from the surface and analyzed by accredited laboratories according to a standardized protocol [51].

2.3. Sentinel-2 satellite data

2.3.1. Band data

The Sentinel-2 mission is an effort of the European Space Agency (ESA) to monitor the Earth's environment. This mission is composed of a constellation of two polar-orbiting satellites, namely Sentinel-2A and Sentinel-2B. Both satellites are placed in the same sun-synchronous orbit, aiming to monitor the Earth's environmental changes [52,53]. The mission started with the launch of the first Sentinel-2A satellite in June 2015. The latter deployment of the second satellite, Sentinel-2B, in March 2017 reduced the revisiting time from 10 days to 5 days [54].

Besides the brief revisiting time, the Sentinel-2 mission is equipped with Multispectral Instruments (MSI), which provide information for different spatial resolutions, ranging from 10 m to 60 m, given the state-of-the-art anastigmatic telescope within it [53,55,56]. The MSI can register data from 13 spectral bands, varying from visible to near infrared (NIR) and short-wave infrared (SWIR), providing high-resolution data for both inland and coastal areas [29,53,57]. The combination of reduced revisiting time, high spatial resolution, and wide range of spectral bands deem the Sentinel-2 an important mission with agriculture applications and forest monitoring [28]. Table 1 compiles the information for each spectral band.

Table 1. Spectral band characteristics for MSI [53,58].

Band	Central wavelength (nm)	Band width (nm)	Spatial resolution (m)	Band spectral range
1	443	20	60	Coastal aerosol
2	490	65	10	Blue
3	560	35	10	Green
4	665	30	10	Red
5	705	15	20	Vegetation red edge 1
6	740	15	20	Vegetation red edge 2
7	783	20	20	Vegetation red edge 3
8	842	115	10	NIR
8A	865	20	20	Narrow NIR
9	945	20	60	Water vapor
10	1380	30	60	SWIR-Cirrus
11	1610	90	20	SWIR 1
12	2190	180	20	SWIR 2

Table 1 informs the wavelength for each band captured by the MSI. It is possible to notice that the wavelengths around 700 nm (NIR) suggest that the Sentinel-2 constellation is suited for capturing the phytoplankton spectral characteristics, including Chla, as the microscopic organisms cause a surge in spectral reflectance around the 700 nm mark [13,29,59].

2.3.2. Satellite spectral indexes

Satellite spectral indexes are derived from mathematical equations combining two or more spectra of the satellite bands. These indices are a helpful approach to extracting information from the spectral bands in a pixelwise fashion to model terrestrial processes and features, like vegetation, water, urban development, and agriculture [58,60,61]. A comprehensive investigation of 16 different indices and their impact over the model's result was performed in the present study. Their mathematical formulations are displayed in equations 1 to 16.

$$NDVI = \frac{Band\ 8 - Band\ 4}{Band\ 8 + Band\ 4} \quad (1)$$

In Equation 1, we have the normalized difference vegetation index (NDVI) formulation. This index is vastly used in remote sensing, being primarily used for the evaluation of green areas and the changes related to it, and is a valuable input on different remote sensing applications [62,63].

$$GNDVI = \frac{Band\ 8 - Band\ 3}{Band\ 8 + Band\ 3} \quad (2)$$

Equation 2 shows the formulation for the Green Normalized Difference Vegetation Index (GNDVI), an adapted version of the NDVI index specifically aimed at detecting Chla in the vegetation [64,65].

$$EVI = 2.5 \cdot \frac{(Band\ 8 - Band\ 4)}{(Band\ 8 + 6 \cdot Band\ 4 - 7.5 \cdot Band\ 2 + 1)} \quad (3)$$

The enhanced vegetation index (EVI) is similar to the previous NDVI but removes the impacts of the atmosphere and the soil over the vegetation signal [66,67].

$$SAVI = \frac{Band\ 8 - Band\ 4}{Band\ 8 + Band\ 4 + 0.428} \cdot 1.428 \quad (4)$$

The soil-adjusted vegetation index (SAVI) improves the NDVI index by considering the soil effects due to multiple scattering of soil [67,68].

$$NDMI = \frac{Band\ 8 - Band\ 11}{Band\ 8 + Band\ 11} \quad (5)$$

The normalized difference moisture index (NDMI) is used to verify changes in vegetation physiology via determination of its water content [69,70].

$$MSI = \frac{Band\ 11}{Band\ 8} \quad (6)$$

The moisture stress index (MSI) is used to evaluate changes in the water content in the vegetation via canopy stress analysis. It is also used to indicate water concentration in the soil [71,72].

$$GCI = \frac{Band\ 9}{Band\ 3} - 1 \quad (7)$$

The green chlorophyll vegetation Index (GCI), as its name implies, is applied to the remote sensed data to estimate chlorophyll concentration in the vegetation, and, consequently,determine the health of the analyzed vegetation [73,74].

$$NBRI = \frac{Band\ 8 - Band\ 12}{Band\ 8 + Band\ 12} \quad (8)$$

The normalized burned ratio index (NBRI) seeks to identify the occurrence and severity of natural or man-made fires in vegetation areas [75,76].

$$BSI = \frac{(Band\ 11 + Band\ 4) - (Band\ 8 + Band\ 2)}{(Band\ 11 + Band\ 4) + (Band\ 8 + Band\ 2)} \quad (9)$$

The bare soil index (BSI) formulated in Equation 9 retrieves information from the vegetation in cases where its coverage is less than half of the assessed area. This index allows us to determine the vegetation health concerning the exposed soil area [77,78].

$$NDWI = \frac{Band\ 3 - Band\ 8}{Band\ 3 + Band\ 8} \quad (10)$$

The normalized difference water index (NDWI) retrieves information from water bodies effectively from remote sensing data [79,80].

$$NDSI = \frac{Band\ 3 - Band\ 11}{Band\ 3 + Band\ 11} \quad (11)$$

The Normalized Difference Snow Index (NDSI) is a tool that detects snow cover in a specific area by analyzing the light reflection properties of ice. This index retrieves information by distinguishing snow coverage from other surfaces and adjusting for atmospheric and terrain effects [81–83].

$$NDGI = \frac{Band\ 3 - Band\ 4}{Band\ 3 + Band\ 4} \quad (12)$$

The normalized difference glacier index (NDGI), similar to NDSI, identifies glacier coverage in a region mainly composed of snow, ice, and debris [84,85].

$$ARVI = \frac{Band\ 8 - 2 \cdot Band\ 4 + Band\ 2}{Band\ 8 + 2 \cdot Band\ 4} \quad (13)$$

The atmospherically resistant vegetation index (ARVI) is an improvement over the NDVI index by implementing atmospheric corrections. The ARVI is especially useful for regions under dense aerosol coverage [86,87].

$$SIPI = \frac{Band\ 8 - Band\ 2}{Band\ 8 - Band\ 4} \quad (14)$$

The structure-insensitive pigment index (SIPI) was initially proposed to identify vegetation stress via the ratio between carotenoid and chlorophyll in vegetation. It is also useful for analyzing vegetation structures with different canopy configurations [88,89] .

$$SWM = \frac{Band\ 2 + Band\ 3}{Band\ 8 + Band\ 11} \quad (15)$$

The sentinel water mask (SWM) specifically seeks to detect water data from the Sentinel-2 constellation [90].

$$AWEI = 4 \cdot (Band\ 3 - Band\ 11) - (0.25 \cdot Band\ 8 + 2.75 \cdot Band\ 12) \quad (16)$$

The automated water extraction index (AWEI) is used to detect water given various environmental interferences [96,97] accurately.

A first analysis of the indexes may result in the conclusion that some of them, e.g., NDSI and NDGI, may not be suitable for the study location due to the semi-arid characterization of the Ceara state. However, a careful examination of Equations 11 and 12, together with Table 1, elucidates that the mentioned indexes result from bands within the spectral range that favors Chla identification, as well as being of high spatial resolution [13,29,59]. Therefore, these indexes may still carry relevant spatiotemporal information that may help uncover the latent relationship between the input parameters and Chla, thus improving the performance of the ML models. This potential improvement provides a compelling reason for further investigation into the participation of these indices in Chla modeling. Figures 2 and 3 show the correlation between the satellite bands and Chla and the indexes and Chla, respectively. In both images, the lighter the color, the more correlated the variables are.

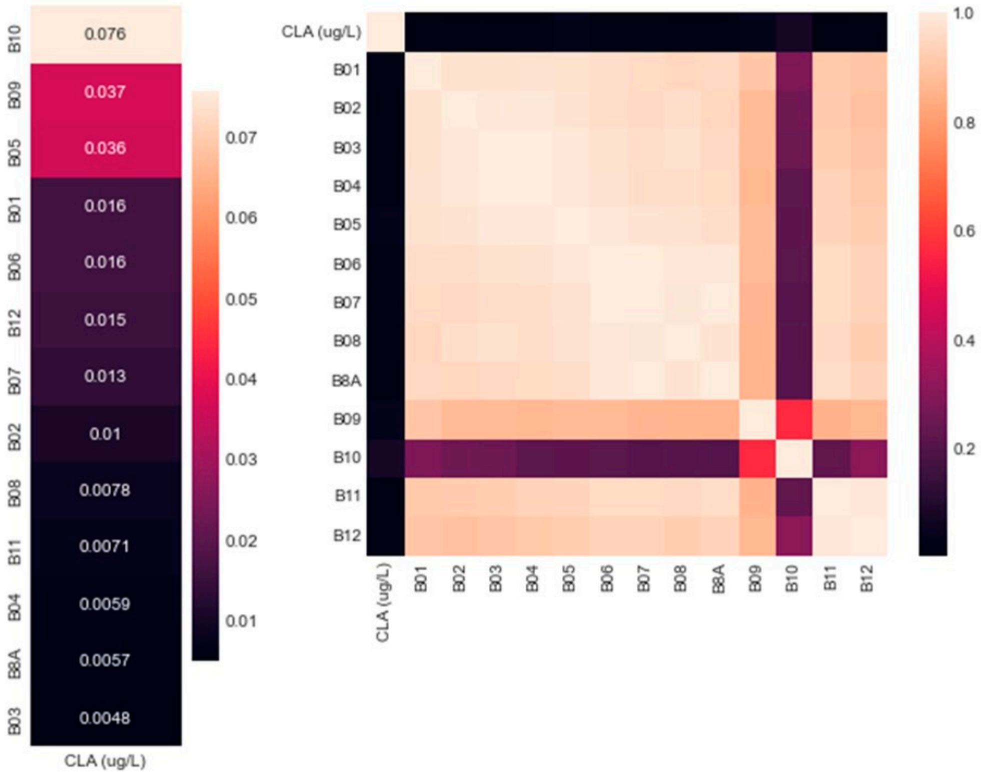

Figure 2. Correlation matrix for the satellite spectral bands and the chlorophyll a.

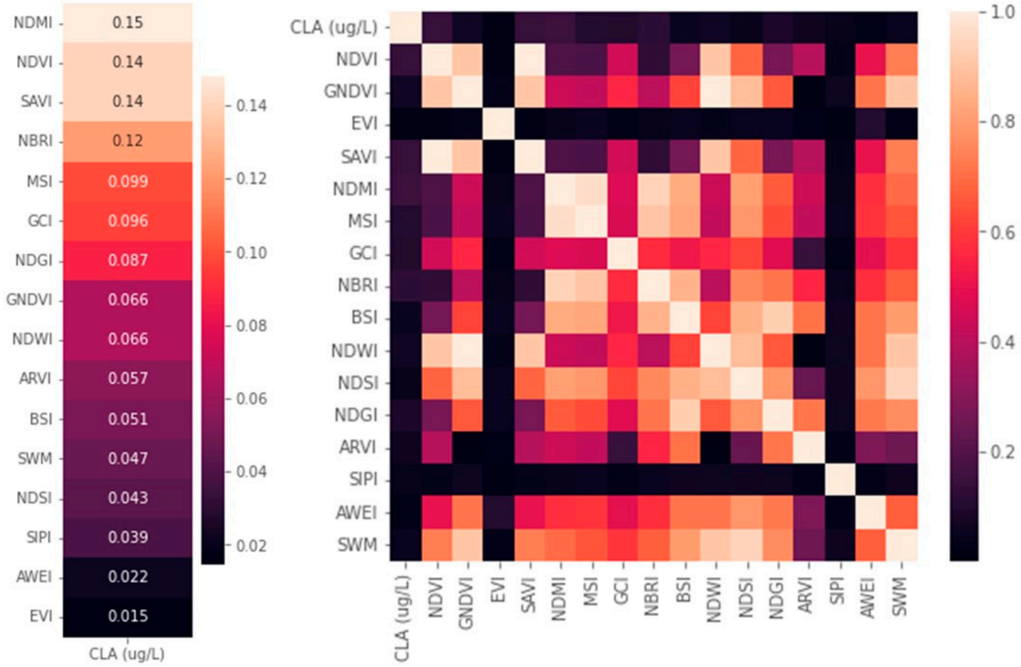

Figure 3. Correlation matrix for the satellite indexes and the chlorophyll a.

Figure 2 shows that the bands share a significant correlation between themselves, but band 10. In contrast, in Figure 3, the indexes show less correlation with each other, indicating potential to be used by ML models due to their low collinearity. Regarding the correlation with the Chla attribute, the bands present significantly lower values, up to ten times smaller than the correlation between the indexes and Chla.

The elevated correlation within the bands' data does not necessarily mean a positive impact on the performance of the forecasting model. In fact, highly correlated variables, when used as inputs on a predictive model, may add noise into the dataset, increasing the model's variance and thus reducing its performance [91,92]. However, discarding highly and low correlated variables altogether may also be detrimental to the model's performance since they may still carry relevant spatiotemporal information relating to the inputs and outputs attributes, leading to improved forecasting performance [30].

2.4. Machine learning models

Chlorophyll a was estimated using satellite data and machine learning models. The first implemented approach was the random forest (RF). The RF is a model composed of trees trained using the bagging method of resampling considering just a subset of predictors (Figures 2 and 3) [100,101], deeming it an ensemble model. The trained trees are low correlated, reducing the ensemble model variance and improving performance [93–95]. Different published works have investigated the RF methodology [96–98].

XGBoost [99] was also implemented as a standalone model and used as benchmark for the models' performance. This tree-based approach is an ensemble model and an extreme improvement over the random forest. It consists of bag sampling smaller tree models, which are then combined into a larger and more robust tree-model, reducing the model variance while improving its generalization and reducing the model's tendency to overfit [93,100]. The XGBoost model can handle missing data and manipulate increasing dataset size, keeping its generalization. This approach reached excellent results when applied to different time-series forecasting tasks [101–103].

In this work, another ML used in Chla prediction was the k-nearest neighbors (k-NN). The k-NN algorithm is a supervised ML model, which uses non-parametric vectors to determine an unknow point [104]. However, given that it is also based on the distance between points distributed on a possible multi-dimensional space, a distance metric must be implemented, often the Euclidean distance [104–106]. The k-NN architecture can be applied to data classification and regression model cases [104,107]. For the former application, the model classifies an unknow datapoint considering its neighbors by a simple voting system, assigning the unknow point the label of the most common class around it. The latter approach has a similar implementation, but this time, a continuous value is assigned to an unknow point given the average value of the target attribute of its neighbors [104,107,108]. Despite its simplicity, the regression performed by the k-NN approach offers competitive results within the ML field and has been explored for different scenarios in previous studies [109–111].

The Support Vector Machine (SVM) was also implemented in this study. The SVM is a flexible ML approach with diverse applications for classification, regression, and outlier detection [112]. Initially proposed by Cortes and Vapnik in 1995 [113], with previous formulations dating back to the 1960s, the SVM is a generalization of the maximal margin classifier [107]. The SVM learns a boundary function discriminating the dataset when implemented for classification. At the same time, for regression, it provides a best-fitting function describing the data behavior by taking into consideration its extreme attribute vectors. A unique feature of SVM is the use of kernel functions that allow the dataset to be manipulated into higher-dimensional spaces, making it possible for the model to learn complex non-linear relationships by applying convex optimization without being computationally expensive reducing the training error [104,114,115]. However, one drawback of the SVM approach is that it does not handle large datasets efficiently, requiring extended computational time to be trained [116,117].

The least absolute shrinkage and selection operator (LASSO) regression [118] was another implemented ML methodology. This approach is a more straightforward ML methodology, which seeks to implement the best linear regression line in the dataset. Besides that, the LASSO paradigm is also a regularization and parameter selection approach, making its results more interpretable compared to other traditional ML models [93,118].

Lastly, we investigated the GMDH ML model for the Chla modeling using satellite data. It was first proposed by Ivakhnenko [119] as an alternative to address the challenges of linear dependency and equation complexity for higher dimensional problems and small data sequences [38,120]. This methodology is a feedforward unidirectional ML model, similar to a multilayer perceptron [121,122]. It is also a self-organizing model, indicating that its parameters are selected automatically, not needing parameter tuning [123]. The resulting value by the GMDH model is a quadratic approximation, using pair combinations of the input variables [124,125], which are used to model the relationship between input and the output parameters [126]. Differently than other artificial neural networks paradigms, the GMDH does not need vast amounts of training data, as its parameters estimation is automatically determined without recursion [38]. The performance of the GMDH model was verified in previous studies for time-series challenges, including hydrological applications [127–129].

2.5. Evaluation metrics

In order to assess the forecasts by the proposed models, we opted to calculate the metrics root mean squared error (RMSE), normalized RMSE (nRMSE), mean absolute error (MAE), mean absolute percentage error (MAPE), mean bias error (MBE), and coefficient of determination (R^2). Their equations can be verified in [130] for R^2, and [131] for the remaining metrics.

2.6. Dataset preprocessing and attribute selection.

The dataset standardization seeks to restrain the dataset to withing the same dimensionless scale, setting the mean value equal to zero and the standard deviation to one, often improving the model's performance [32]. Besides the data standardization, in this study, the Yeo-Johnson transformation [132] was implemented for some of the ML models. This transformation is an improvement over the Box-Cox approach, which is restricted to handling positive numbers only. The Yeo-Johnson transformation is based on the power transformation with different parameters to positive and negative values and is presented in Equation 18 [132].

$$\psi(y,\lambda) = \begin{cases} \dfrac{(y+1)^{\lambda}+1}{\lambda} & y \geq 0, \lambda \neq 0 \\ \log(y+1) & y \geq 0, \lambda = 0 \\ -\dfrac{(-y+1)^{2-\lambda}-1}{2-\lambda} & y < 0, \lambda \neq 2 \\ -\log(-y+1) & y < 0, \lambda = 2 \end{cases} \quad (18)$$

where the transformed value ψ is a function of the original attribute value, y, and a parameter λ, which is determined via maximum likelihood [101, 102]. This transformation seeks to reduce the data skewness, approximating the original dataset distribution to a normal distribution as $\psi(y,\lambda) \sim N(\mu, \sigma^2)$ [133,134].

The best parameters were selected among the bands and indexes. To this end, using a step-by-step approach [107], we started investigating the influence of the bands only. First, we investigated the performance of just one band at a time, from band 1 to 12, and we selected the best one which returned the best R^2, using the XGBoost model as the benchmarking approach. After that, we tested the combination of the already selected bands, with the remaining ones, again, one-by-one, and kept the band that returned the highest coefficient of determination value (forward variable selection). This process was repeated until all the bands were investigated and their best combination was kept [107].

The same procedure was repeated for the selection of the indexes. This time, we started the investigation with the best band selection previously set. Then, we added one index at a time to the input parameters, selecting the index which yielded the highest R². Again, the process was repeated until all the indexes were assessed. The selection of the best bands and indexes are presented in Figures 4 and 5, respectively.

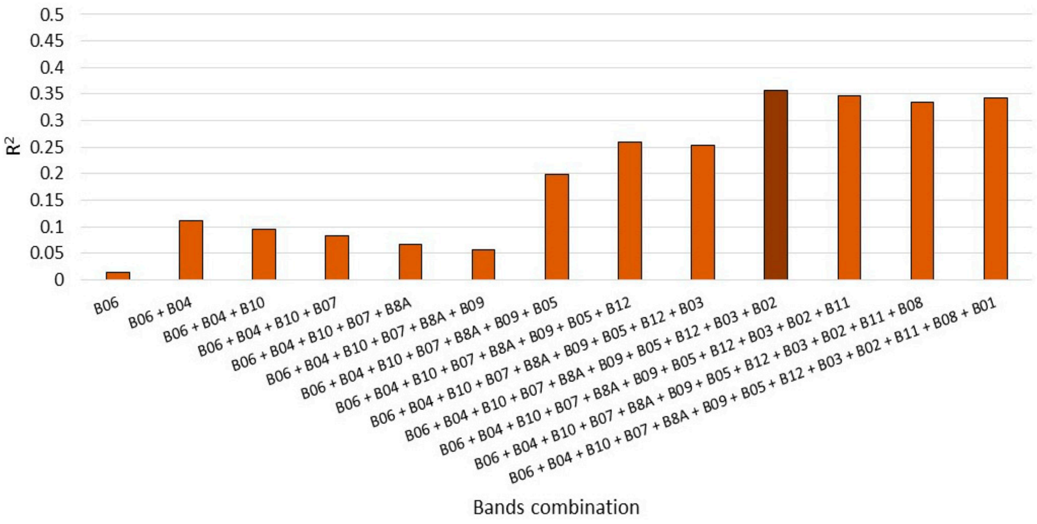

Figure 4. Selected spectral bands combination. The darker color indicates the best result in terms of R².

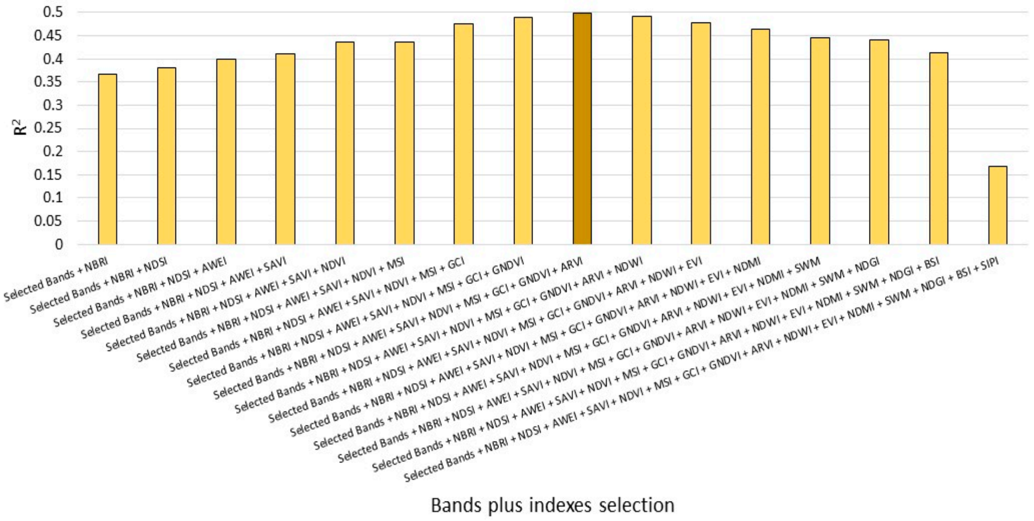

Figure 5. Selected bands plus the selected spectral indexes combination. The darker color indicates the best result in terms of R².

As depicted in Figure 4, the modeling for Chla substantially improved with the combination of bands 6, 4, 10, 7, 8a, 9, 5, 12, and 3, resulting in an R² of 0.36. Figure 5 shows that the addition of the indexes also managed to increase the coefficient of determination to 0.50 after including NBRI, NDSI, AWEI, SAVI, NDVI, MSI, GCI, GNDVI, and ARVI in the selected input data. Interestingly, the NDSI, an index related to snow coverage, promoted improvements over the Chla identification, as indicated in Figure 5, proving that it conveyed spatiotemporal information for the modeling approach. Therefore, the results discussed in the subsequent section are based on grouping the aforementioned spectral bands and indexes.

3. Results

3.1. Limnological behavior

The mean value for Chlorophyll varied widely among reservoirs and according to seasonality. Considering the complete dataset, Chla values ranged from a minimum of 1 µg/L to a maximum of 1001.78 µg/L, with a mean value of 39.62 µg/L and a standard deviation of 65.78 µg/L. The boxplot in Figure 6-A shows the inter-annual and intra-annual distribution of the dataset. Mean values decreased throughout the years (from 81.34 µg/L in 2015 to 27.84 µg/L in 2021), while maximum values did not reflect this tendency, with recorded values increasing, for example, from 707.06 in 2015 to 1001.78 in 2016, indicating the occurrence of concentration blooms. Regarding the intra-annual distribution of Chla, higher concentrations were detected in the rainy seasons for most of the studied years, except in 2017.

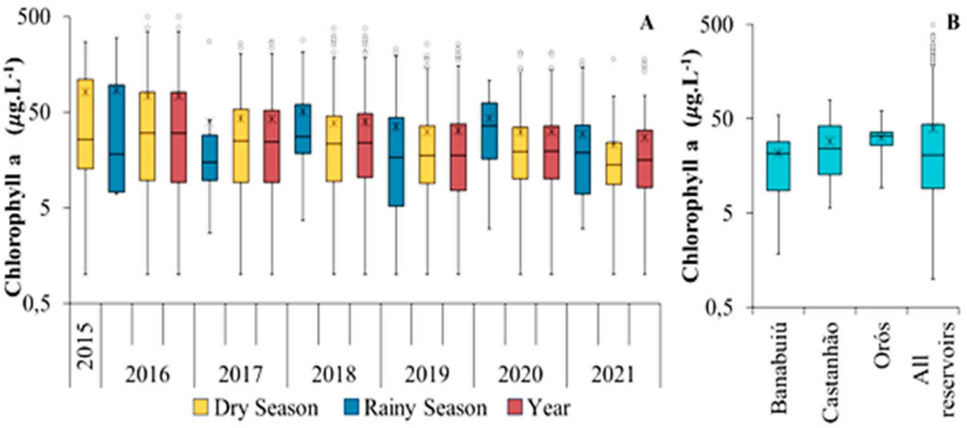

Figure 6. Boxplots presenting basic statistics of Chla concentrations in logarithmic scale: (A) grouped by sampling year and categorized by season (Rainy: January from April, Dry: remaining months); (B) grouped by the state's largest reservoirs (Castanhão, Banabuiú and Orós) and all studied reservoirs. The lower and upper limits represent minimum and maximum values. The bottom and top of the box represent the first and third quartiles. Inner lines, asterisks, and circles indicate media, means, and outliers.

The state's three largest reservoirs were also analyzed (Figure 6-B). The mean recorded values for Orós were the highest (31.89 µg/L), between 9.20 and 60.94 µg/L, while Chla for Banabuiú was found in lower mean concentrations (21.64 µg/L) between 1.82 and 54.55 µg/L. For Castanhão, values ranged from 5.65 and 78.72 µg/L with mean Chla value of 29.03 µg/L.

3.2. Results for Chla estimation by the ML models

The proposed models were built using as input parameters the variables present in Figures 4 and 5. We utilized data from the Sentinel-2 satellite, combined with on-site measurements from the 149 reservoirs, to train and test these models. We adopted a 70/30 split for the training and testing datasets. These datasets were randomly assembled using data spanning from 2015 to 2021. The outcomes for Chla modeling by each one of the assessed models are presented in Table 2.

Table 2. Results for Chla estimation for the standalone models.

Model	RMSE (µg/L)	nRMSE (%)	MAE (µg/L)	MAPE (%)	MBE (µg/L)	R^2	Yeo-Johnson Transformation
k-NN	61.82	146.07	30.90	260.60	-4.91	0.38	Yes
XGBoost	55.60	131.36	29.41	288.34	-2.53	0.50	No
RF	56.75	134.10	29.92	311.58	-1.54	0.48	No

doi:10.20944/preprints202312.1854.v1

SVR	50.64	119.64	25.07	182.60	-6.97	0.58	Yes
LASSO	89.87	212.34	47.41	466.35	-3.60	0.41	Yes
GMDH	20.38	53.20	14.09	102.34	-4.86	0.91	Yes

Table 2 shows that the ML approaches reached similar results regarding the RMSE metric, but for the GMDH which was able to surpass all of the assessed paradigms, achieving RMSE of 20.38 µg/L. The highest error value for RMSE was achieved by the LASSO model, with 89.87 µg/L, followed by the k-NN model, with RMSE value of 61.82 µg/L. The tree-based models XGBoost and RF achieved similar RMSE values equal to 55.60 µg/L and 56.75 µg/L, respectively. The SVR output scored an RMSE value equal to 50.64 µg/L. It is essential to disclose that an analysis of the RMSE alone may be misleading to assess the performance of the models. The coefficient of determination is another crucial factor to consider in this scenario. This parameter indicates the total variance of the dependent variable Chla, which can be adequately forecasted by the input parameters and may be understood as an indication of the accuracy achieved by the model [135]. This behavior can be better visualized in Figure 7.

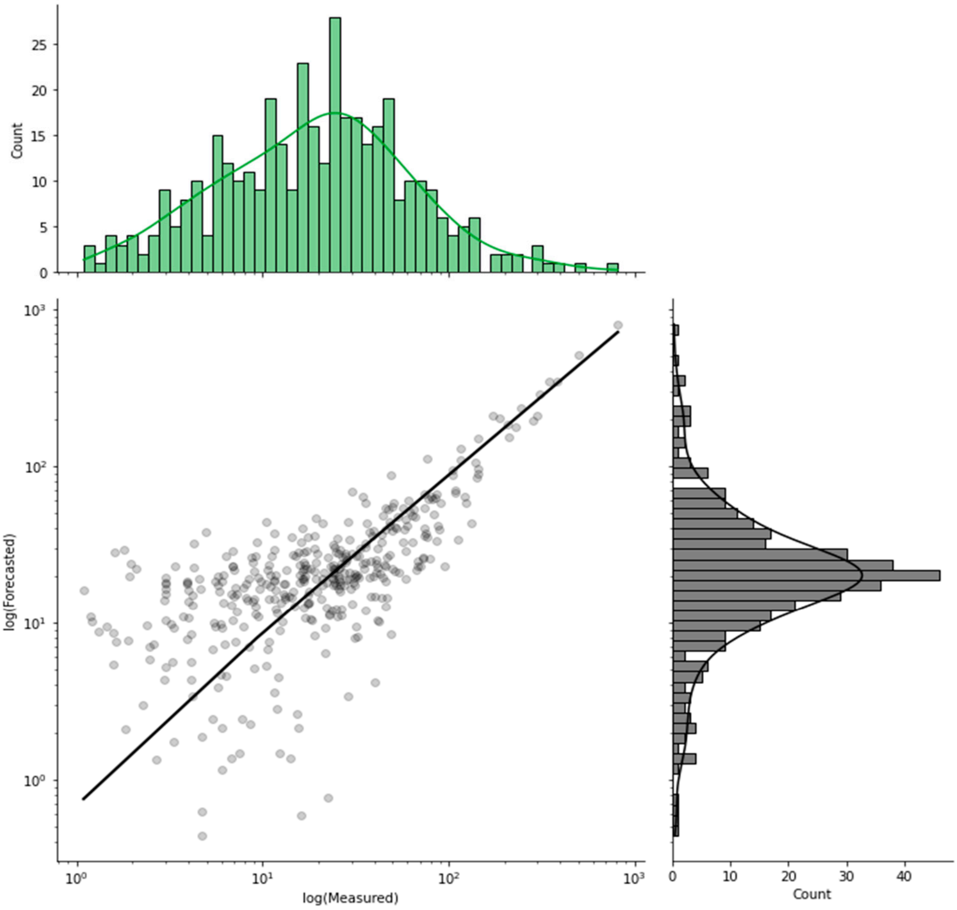

Figure 7. Scatter plot with marginal distribution for the Chla forecasting by the GMDH model.

Figure 7 shows how the predicted data compares with the actual measured Chla values. We used log values to facilitate the comparison due to differences in the variables' scales. The histograms showcase the normal distribution of the data after the logarithm transformation. It is possible to observe that both predicted and real data present similar distributions, which indicates the GMDH model's good performance in predicting Chla levels. The regression line displayed in the plotting area shows a positive correlation between the predicted and real Chla data, further attesting the robust GMDH performance. The points clustered around the regression line are also an indication of the superior performance of the GMDH algorithm, especially for extreme values as depicted by the top-rightmost points in the graph area. Comparing the results in Table 2, we can observe that the highest R^2 value reached by the GMDH model was 0.91, representing a significant improvement of 57% over the SVR result, the second best-performing approach regarding the same parameter. The third- and fourth-best performing models were the tree-based ones XGBoost and RF, which achieved R^2 equal to 0.50 and 0.48, respectively. The k-NN performance, R^2 equal to 0.38, and the LASSO model, R^2 equal to 0.38, achieving a coefficient of determination equal to 0.41, were the worst-performing modeling paradigms.

4. Discussion of the results

Mainly in tropical regions, the mechanisms that regulate phytoplankton growth require an advanced analysis due to the complexity and nonlinearity of the relationship between chlorophyll and physicochemical/environmental factors [13,136]. The approach presented in this study achieves this analysis using a highly heterogeneous collection of in-situ observations, investigating the performance of different ML models to determine Chla levels in 149 dams in the Brazilian state of Ceará. Although most dams were predominantly eutrophic throughout the years, different human activities and pollution sources have contributed to eutrophication processes lead to Chla spatiotemporal fluctuations and algal blooms [17,20,137,138].

We applied the forward stepwise approach to select the parameters, including bands and spectral indexes. To the best of our knowledge, this method has not yet been applied in previous remote sensing studies. The forward stepwise approach showed significant improvement for the Chla determination for the ML models. Existing literature regarding the influence of spectral bands for Chla determination can be found. In work [139], the authors applied the SHAP analysis to explain the influence of the Sentinel-2 spectral bands for estimating Chla values. Their results elucidated that band 3, band 2, and band 8 were the top-three most influential parameters for Chla determination. Bands 3 and 2 exhibited a positive correlation with the Chla, while band 8 showed a relevant negative correlation with the same parameter. A similar approach was conducted by Kim et al. [140]. In their work, the SHAP analysis showed relevant participation on Chla prediction of red bands, i.e., bands 4, 5 and 6, as well as blue and green bands (Table 1), a similar conclusion found by Ha et al. [141].

The influence of different spectral indexes for Chla modeling was investigated in previous works. The work of Castro et al. [142] showed that indexes merging red and NIR bands yielded the best outcomes for the determination of Chla concentrations in small reservoirs. Similar conclusions were drawn by Buma and Lee [143] and by Aubriot et al. [144], which attested the importance of bands within the red spectral range for Chla characterization in a lake in Chad, and for the Rio de La Plata, respectively. On the other hand, Viso-Vazquez et al. [145], showed that the green band, i.e., band 3, promoted the highest correlation between the remote sensing data and Chla levels.

Our results show the GMDH methodology could efficiently identify the latent non-linear ties ruling input and output attributes. Moreover, the GMDH approach proved to be a more robust model for analyzing satellite imagery, being more resilient to noises and satisfactory results when handling different bands information, which also can be attributed to its improved performance.

To better understand where the GMDH results lie within the literature, we compared our results with those found in previously published works. Such evaluation, however, may not be completely representative given the different methodologies used in different studies, the different models, input attributes, and the study area. Yet, a comparison between their evaluation metrics is still a viable approach to assess the performance of different models [146]. Table 4 compiles the results for the GMDH model, while Table 5 gathers the results found in the literature.

Table 4. Results for the GMDH model estimation of Chla.

Model	RMSE (μg/L)	MAE (μg/L)	MBE (μg/L)	R^2
GMDH	20.38	14.09	-4.86	0.91

From Table 5, we observe that the models based on the deep learning methodology all performed remarkably well in determining Chla levels. Compared to the results found in our study, it is possible to notice that the R-squared values for references [147] and [148] are in the same range, over 90%. Nevertheless, regarding the work by Guo et al. [147], the combined utilization of Sentinel-2 and Landsat-8 data significantly enhances the machine learning model's performance. This improvement is primarily due to the reduction in the revisit time, which subsequently minimizes the variance in the dataset. Consequently, this leads to a more robust and accurate machine-learning model. It is important to note that, although their results are slightly superior, the study location is limited to only one place.

Table 5. Compilation of results for Chla modeling found in the literature.

Model	Location	Dataset	RMSE μg/L	R^2	Reference
Multimodal Deep Learning	Lake Simcoe, Canada	Sentinel-2 and Landsat-8 imagery	60	0.92	[147]
Convolutional Neural Network	Lake Balik, Turkey	Sentinel-2 imagery	2.9	0.95	[148]
Convolutional Neural Network	11 Lakes in Karlsruhe, Germany	Simulated Chla data used for training, data from SpecWa used for evaluation	12.4	0.82	[149]
SVR	45 Lakes across China	Sentinel-2 imagery	6.3	0.88	[150]
SVR	Buffalo Pound Lake, Canada	Sentinel-2 imagery	13.9	0.66	[151]
Toming's index	A Baxe reservoir, Spain	Sentinel-2 imagery		0.86	[145]
3BSI index	5 Reservoirs in Ceará, Brazil	Sentinel-2 imagery		0.80	[13]

The addition of more water bodies is expected to add variance to the dataset. The model in reference [149] was built using modeled data, including 11 times more water bodies than the previous studies. Their results proved to be slightly inferior to the other two previous works, and more relatable to the ones found in this present assessment, with RMSE in the same magnitude order.

Regarding the R^2, our GMDH model showed a value nearly 10% superior. However, it must be stated that since the data used for training the ML model was simulated, it may not consider several natural occurring situations. This would lead to a more homogeneous dataset, with less variance, improving the ML model performance compared to the model proposed in our study. Another significant difference is the time window used for testing the developed model in reference [149], which is significantly smaller than the one used in our model [152], which reduces the dataset variance improving the ML performance.

The DL approaches presented in Table 5 still require a huge amount of data to yield reliable outputs. This characteristic may pose a major drawback depending on the data availability for the study location. On the other hand, the GMDH paradigm can be promptly implemented with less information available, and not requiring extensive training dataset [38], deeming them a more flexible usage for different situations.

References [150] and [151] applied SVR to determine Chla. It is noticeable that although the models are the same, their methodology was disparate. Regarding reference [150], the RMSE values were within the same order of magnitude. Besides that, the R^2 values were relatively close to each other. Contrarywise, reference [151] performed a much broader study regarding several lakes spread across the Chinese territory. As previously mentioned, the inclusion of different lakes allows the forecasting model to better generalize its results, and thus provide more robust outcomes. Another critical difference between these two works is that the former implemented as input information just the spectral bands of the MSI onboard of Sentinel-2, while the latter used both bands and indexes. In this aspect, the present study achieved superior performance, attesting the GMDH model benefited from the inclusion of spectral indexes, which improved their Chla outcomes.

The work conducted by Aranha et al. [13] shares the same location as the one used in this study. However, they used only a subset of 5 reservoirs of 149. In their approach, the authors fitted a regression line to their dataset using the three-band spectral index (3BSI), showing a good agreement between the index and the Chla values. A similar methodology was implemented in reference [145], where the Toming's index was used to fit a regression line for Chla values. These two studies implemented spectral bands to estimate Chla concentrations. By evaluating the R^2 metric values for [13] and [145], the proposed GMDH was superior over both studies with significant improvements of 12% and 5%, respectively. Furthermore, a major difference between these two studies and the methodology presented in this work, is the data handling. The other studies used methodologies which consisted of fitting a regression line using the proposed indexes over the entire dataset, making no distinction between training and testing datasets. This is analogous to assessing our ML model's performance considering the training dataset only. Therefore, their methodologies lack generalization, being bound to a particular time and geographic location. In addition, while these approaches are considerably less complex than the proposed ML model, they provide valuable insights into Chla's behavior when analyzed using the evaluated indexes.

It is important to disclose that our study was conducted in a tropical semi-arid region, considering 149 dams, which is considerably larger than any of the presented literature. In addition, a broader investigation of the spectral indexes for the Sentinel-2 constellation was assessed compared to the references in Table 5. This allows us to gather further insight into the impact of the spatiotemporal influence of both bands and indexes on the Chla prediction. From the results for parameters selection in Figures 4 and 5, we observe that bands 8 and 11 do not compose the selected bands set. However, they are still present in the form of indexes. Moreover, careful assessment of the GMDH results (available at the Supplementary Spreadsheet 1) showcase relevant contribution of bands 3, 4, 5, 7, 8 and 11 to the Chla prediction. This indicates that the predictive models benefit from a higher spatial resolution, as well as from green, red and infra-red bands composing the indexes, accordingly with the literature [13,29,59].

5. Conclusions

In this study, several input parameters were evaluated for Chla modeling in a vast spatial coverage of 149 freshwater reservoirs spanning over a semi-arid tropical region across the Brazilian state of Ceará. This evaluation was conducted based on satellite remotely sensed data and a set of ground-truth Chla concentrations measurements that reflected the temporal and spatial distribution, notably impacted by interannual rainfall variability. To this end, we investigated the performance of several ML approaches, namely the k-nearest neighbours, random forests, extreme gradient boosting, the least absolute shrinkage and selection operator, support vector machine, and GMDH.

Forward stepwise parameter selection was implemented to determine the best input attribute selection among the 13 bands from the Sentinel-2 constellation and 16 spectral indexes derived from such bands. The usage of this methodology applied to the remote sensing field is, to the best knowledge of the authors, new and proved to improve the outcomes of the investigated Chla forecasting models. Figure 7 elucidates the performance of GMDH model, showing that the model under investigation yielded excelent values for the Chla concentrations, achieving R^2 value greater than 90%.

It is important to emphasize that our methodology ensures a clear separation between training and testing data, unlike some studies found in the literature. The approach is substantial for applying ML models, as a way to promote more robust results and capacity to generalize unseen data. This approach is especially crucial for a large dataset such as the one used in the present work, comprising 149 reservoirs in a semi-arid region similar in size to England, where the spatiotemporal variation of the data is substantial. Regarding the spectral bands and indexes, we showed that the Chla modeling benefited mostly from red, NIR, and green bands.

Extensive comparison with the literature found studies showed that the used models in this work offer competitive results. It is important to note that such comparison may be misleading due to the disparate methodologies and study locations, which directly influence the ML outcomes.

In future works, the addition of more indexes, as well as the merge of Sentinel-2 data with Landsat-8 could be investigated. Including more spectral indexes and the Landsat-8 MODIS data would provide more spatiotemporal information and reduce data variance due to the finer temporal resolution. Furthermore, the implementation of atmosphere correction preprocessing could also be beneficial to the predictive paradigms, as it would reduce data noise, diminish the error variance, and improve the forecasting of Chla concentration.

Supplementary Materials: The following supporting information can be downloaded at the website of this paper posted on Preprints.org.

Author Contributions: Conceptualization, I.E.L.N. and P.A.C.R.; methodology, I.E.L.N., P.A.C.R., F.A.S.F., J.V.G.T., and B.G.; software, P.A.C.R.; validation, I.E.L.N., P.A.C.R., F.A.S.F., J.V.G.T., and B.G.; formal analysis, I.E.L.N. and P.A.C.R.; investigation, I.E.L.N., P.A.C.R., F.A.S.F., J.V.G.T., and B.G.; resources, I.E.L.N. and P.A.C.R.; data curation, I.E.L.N. and P.A.C.R.; writing—original draft preparation, B.M.D.M.G., V.O.S., I.E.L.N. and P.A.C.R.; writing—review and editing, B.M.D.M.G., V.O.S., I.E.L.N., F.A.S.F., J.V.G.T., and B.G.; visualization, B.M.D.M.G., V.O.S., I.E.L.N. and P.A.C.R.; supervision, I.E.L.N., P.A.C.R., J.V.G.T. and B.G.; project administration, I.E.L.N., P.A.C.R., J.V.G.T. and B.G.; funding acquisition, B.G. and J.V.G.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research study was funded by the Natural Sciences and Engineering Research Council of Canada (NSERC) Alliance, grant No. 401643, in association with Lakes Environmental Software Inc., and by the Conselho Nacional de Desenvolvimento Científico e Tecnológico—Brasil (CNPq), grant no. 303585/2022-6 and grant no. 312308/2020-5, and by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior—Brasil (CAPES), grant no. PROEX 04/2021.

Data Availability Statement: Publicly available datasets were analyzed in this study. This data can be found here: https://drive.google.com/drive/folders/1ALvoqdtj1nbfdNWmna0PpJnD1xUgzLqw (accessed on 20 December 2023).

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

References

1. Mamun, M.; Kwon, S.; Kim, J.-E.; An, K.-G. Evaluation of Algal Chlorophyll and Nutrient Relations and the N:P Ratios along with Trophic Status and Light Regime in 60 Korea Reservoirs. *Science of The Total Environment* **2020**, *741*, 140451. https://doi.org/10.1016/j.scitotenv.2020.140451.
2. Zhang, X.; Chen, X.; Zheng, G.; Cao, G. Improved Prediction of Chlorophyll-a Concentrations in Reservoirs by GRU Neural Network Based on Particle Swarm Algorithm Optimized Variational Modal Decomposition. *Environmental Research* **2023**, *221*, 115259. https://doi.org/10.1016/j.envres.2023.115259.
3. Rocha, M.D.J.D.; Lima Neto, I.E. Internal Phosphorus Loading and Its Driving Factors in the Dry Period of Brazilian Semiarid Reservoirs. *Journal of Environmental Management* **2022**, *312*, 114983. https://doi.org/10.1016/j.jenvman.2022.114983.
4. Rodríguez-López, L.; Usta, D.B.; Duran-Llacer, I.; Alvarez, L.B.; Yépez, S.; Bourrel, L.; Frappart, F.; Urrutia, R. Estimation of Water Quality Parameters through a Combination of Deep Learning and Remote Sensing Techniques in a Lake in Southern Chile. *Remote Sensing* **2023**, *15*, 4157. https://doi.org/10.3390/rs15174157.
5. Rolim, S.B.A.; Veettil, B.K.; Vieiro, A.P.; Kessler, A.B.; Gonzatti, C. Remote Sensing for Mapping Algal Blooms in Freshwater Lakes: A Review. *Environ Sci Pollut Res* **2023**, *30*, 19602–19616. https://doi.org/10.1007/s11356-023-25230-2.
6. Rocha, P.A.C.; Santos, V.O.; Thé, J.V.G.; Gharabaghi, B. New Graph-Based and Transformer Deep Learning Models for River Dissolved Oxygen Forecasting. **2023**.
7. Kayastha, P.; Dzialowski, A.R.; Stoodley, S.H.; Wagner, K.L.; Mansaray, A.S. Effect of Time Window on Satellite and Ground-Based Data for Estimating Chlorophyll-a in Reservoirs. *Remote Sensing* **2022**, *14*, 846. https://doi.org/10.3390/rs14040846.
8. Zhu, W.-D.; Qian, C.-Y.; He, N.-Y.; Kong, Y.-X.; Zou, Z.-Y.; Li, Y.-W. Research on Chlorophyll-a Concentration Retrieval Based on BP Neural Network Model—Case Study of Dianshan Lake, China. *Sustainability* **2022**, *14*, 8894. https://doi.org/10.3390/su14148894.
9. 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 Sensing of Environment* **2020**, *248*, 111974. https://doi.org/10.1016/j.rse.2020.111974.

10. Fu, L.; Zhou, Y.; Liu, G.; Song, K.; Tao, H.; Zhao, F.; Li, S.; Shi, S.; Shang, Y. Retrieval of Chla Concentrations in Lake Xingkai Using OLCI Images. *Remote Sensing* **2023**, *15*, 3809. https://doi.org/10.3390/rs15153809.
11. Dzurume, T.; Dube, T.; Shoko, C. Remotely Sensed Data for Estimating Chlorophyll-a Concentration in Wetlands Located in the Limpopo Transboundary River Basin, South Africa. *Physics and Chemistry of the Earth, Parts A/B/C* **2022**, *127*, 103193. https://doi.org/10.1016/j.pce.2022.103193.
12. Karimian, H.; Huang, J.; Chen, Y.; Wang, Z.; Huang, J. A Novel Framework to Predict Chlorophyll-a Concentrations in Water Bodies through Multi-Source Big Data and Machine Learning Algorithms. *Environ Sci Pollut Res* **2023**, *30*, 79402–79422. https://doi.org/10.1007/s11356-023-27886-2.
13. Aranha, T.R.B.T.; Martinez, J.-M.; Souza, E.P.; Barros, M.U.G.; Martins, E.S.P.R. Remote Analysis of the Chlorophyll-a Concentration Using Sentinel-2 MSI Images in a Semiarid Environment in Northeastern Brazil. *Water* **2022**, *14*, 451. https://doi.org/10.3390/w14030451.
14. Barros, M.U.G.; Wilson, A.E.; Leitão, J.I.R.; Pereira, S.P.; Buley, R.P.; Fernandez-Figueroa, E.G.; Capelo-Neto, J. Environmental Factors Associated with Toxic Cyanobacterial Blooms across 20 Drinking Water Reservoirs in a Semi-Arid Region of Brazil. *Harmful Algae* **2019**, *86*, 128–137. https://doi.org/10.1016/j.hal.2019.05.006.
15. Lu, K.; Gao, X.; Yang, F.; Gao, H.; Yan, X.; Yu, H. Driving Mechanism of Water Replenishment on DOM Composition and Eutrophic Status Changes of Lake in Arid and Semi-Arid Regions of Loess Area. *Science of The Total Environment* **2023**, *899*, 165609. https://doi.org/10.1016/j.scitotenv.2023.165609.
16. Raulino, J.B.S.; Silveira, C.S.; Lima Neto, I.E. Assessment of Climate Change Impacts on Hydrology and Water Quality of Large Semi-Arid Reservoirs in Brazil. *Hydrological Sciences Journal* **2021**, *66*, 1321–1336. https://doi.org/10.1080/02626667.2021.1933491.
17. Guimarães, B.M.D.M.; Neto, I.E.L. Chlorophyll-a Prediction in Tropical Reservoirs as a Function of Hydroclimatic Variability and Water Quality. *Environ Sci Pollut Res* **2023**, *30*, 91028–91045. https://doi.org/10.1007/s11356-023-28826-w.
18. Rocha Junior, C.A.N.D.; Costa, M.R.A.D.; Menezes, R.F.; Attayde, J.L.; Becker, V. Water Volume Reduction Increases Eutrophication Risk in Tropical Semi-Arid Reservoirs. *Acta Limnol. Bras.* **2018**, *30*. https://doi.org/10.1590/s2179-975x2117.
19. Rocha, M.D.J.D.; Lima Neto, I.E. Modeling Flow-Related Phosphorus Inputs to Tropical Semiarid Reservoirs. *Journal of Environmental Management* **2021**, *295*, 113123. https://doi.org/10.1016/j.jenvman.2021.113123.
20. Wiegand, M.C.; Do Nascimento, A.T.P.; Costa, A.C.; Lima Neto, I.E. Trophic State Changes of Semi-Arid Reservoirs as a Function of the Hydro-Climatic Variability. *Journal of Arid Environments* **2021**, *184*, 104321. https://doi.org/10.1016/j.jaridenv.2020.104321.
21. Freire, L.L.; Costa, A.C.; Neto, I.E.L. Effects of Rainfall and Land Use on Nutrient Responses in Rivers in the Brazilian Semiarid Region. *Environ Monit Assess* **2023**, *195*, 652. https://doi.org/10.1007/s10661-023-11281-y.
22. Rabelo, U.P.; Dietrich, J.; Costa, A.C.; Simshäuser, M.N.; Scholz, F.E.; Nguyen, V.T.; Lima Neto, I.E. Representing a Dense Network of Ponds and Reservoirs in a Semi-Distributed Dryland Catchment Model. *Journal of Hydrology* **2021**, *603*, 127103. https://doi.org/10.1016/j.jhydrol.2021.127103.
23. Rabelo, U.P.; Costa, A.C.; Dietrich, J.; Fallah-Mehdipour, E.; Van Oel, P.; Lima Neto, I.E. Impact of Dense Networks of Reservoirs on Streamflows at Dryland Catchments. *Sustainability* **2022**, *14*, 14117. https://doi.org/10.3390/su142114117.
24. Li, J.; Pei, Y.; Zhao, S.; Xiao, R.; Sang, X.; Zhang, C. A Review of Remote Sensing for Environmental Monitoring in China. *Remote Sensing* **2020**, *12*, 1130. https://doi.org/10.3390/rs12071130.
25. 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 Sensing of Environment* **2020**, *240*, 111604. https://doi.org/10.1016/j.rse.2019.111604.
26. Song, K.; Wang, Q.; Liu, G.; Jacinthe, P.-A.; Li, S.; Tao, H.; Du, Y.; Wen, Z.; Wang, X.; Guo, W.; et al. A Unified Model for High Resolution Mapping of Global Lake (>1 Ha) Clarity Using Landsat Imagery Data. *Science of The Total Environment* **2022**, *810*, 151188. https://doi.org/10.1016/j.scitotenv.2021.151188.
27. Shi, J.; Shen, Q.; Yao, Y.; Li, J.; Chen, F.; Wang, R.; Xu, W.; Gao, Z.; Wang, L.; Zhou, Y. Estimation of Chlorophyll-a Concentrations in Small Water Bodies: Comparison of Fused Gaofen-6 and Sentinel-2 Sensors. *Remote Sensing* **2022**, *14*, 229. https://doi.org/10.3390/rs14010229.
28. Segarra, J.; Buchaillot, M.L.; Araus, J.L.; Kefauver, S.C. Remote Sensing for Precision Agriculture: Sentinel-2 Improved Features and Applications. *Agronomy* **2020**, *10*, 641. https://doi.org/10.3390/agronomy10050641.
29. Bramich, J.; Bolch, C.J.S.; Fischer, A. Improved Red-Edge Chlorophyll-a Detection for Sentinel 2. *Ecological Indicators* **2021**, *120*, 106876. https://doi.org/10.1016/j.ecolind.2020.106876.
30. Oliveira Santos, V.; Costa Rocha, P.A.; Thé, J.V.G.; Gharabaghi, B. Graph-Based Deep Learning Model for Forecasting Chloride Concentration in Urban Streams to Protect Salt-Vulnerable Areas. *Environments* **2023**, *10*, 157. https://doi.org/10.3390/environments10090157.

31. Oliveira Santos, V.; Costa Rocha, P.A.; Scott, J.; Van Griensven Thé, J.; Gharabaghi, B. Spatiotemporal Air Pollution Forecasting in Houston-TX: A Case Study for Ozone Using Deep Graph Neural Networks. *Atmosphere* **2023**, *14*, 308. https://doi.org/10.3390/atmos14020308.
32. Rocha, P.A.C.; Santos, V.O. Global Horizontal and Direct Normal Solar Irradiance Modeling by the Machine Learning Methods XGBoost and Deep Neural Networks with CNN-LSTM Layers: A Case Study Using the GOES-16 Satellite Imagery. *Int J Energy Environ Eng* **2022**, *13*, 1271–1286. https://doi.org/10.1007/s40095-022-00493-6.
33. Shirmard, H.; Farahbakhsh, E.; Müller, R.D.; Chandra, R. A Review of Machine Learning in Processing Remote Sensing Data for Mineral Exploration. *Remote Sensing of Environment* **2022**, *268*, 112750. https://doi.org/10.1016/j.rse.2021.112750.
34. Mohajane, M.; Costache, R.; Karimi, F.; Bao Pham, Q.; Essahlaoui, A.; Nguyen, H.; Laneve, G.; Oudija, F. Application of Remote Sensing and Machine Learning Algorithms for Forest Fire Mapping in a Mediterranean Area. *Ecological Indicators* **2021**, *129*, 107869. https://doi.org/10.1016/j.ecolind.2021.107869.
35. Virnodkar, S.S.; Pachghare, V.K.; Patil, V.C.; Jha, S.K. Remote Sensing and Machine Learning for Crop Water Stress Determination in Various Crops: A Critical Review. *Precision Agric* **2020**, *21*, 1121–1155. https://doi.org/10.1007/s11119-020-09711-9.
36. Shi, X.; Gu, L.; Jiang, T.; Zheng, X.; Dong, W.; Tao, Z. Retrieval of Chlorophyll-a Concentrations Using Sentinel-2 MSI Imagery in Lake Chagan Based on Assessments with Machine Learning Models. *Remote Sensing* **2022**, *14*, 4924. https://doi.org/10.3390/rs14194924.
37. Hu, C.; Feng, L.; Guan, Q. A Machine Learning Approach to Estimate Surface Chlorophyll *a* Concentrations in Global Oceans From Satellite Measurements. *IEEE Trans. Geosci. Remote Sensing* **2021**, *59*, 4590–4607. https://doi.org/10.1109/TGRS.2020.3016473.
38. Alizamir, M.; Heddam, S.; Kim, S.; Mehr, A.D. On the Implementation of a Novel Data-Intelligence Model Based on Extreme Learning Machine Optimized by Bat Algorithm for Estimating Daily Chlorophyll-a Concentration: Case Studies of River and Lake in USA. *Journal of Cleaner Production* **2021**, *285*, 124868. https://doi.org/10.1016/j.jclepro.2020.124868.
39. Loc, H.H.; Do, Q.H.; Cokro, A.A.; Irvine, K.N. Deep Neural Network Analyses of Water Quality Time Series Associated with Water Sensitive Urban Design (WSUD) Features. *Journal of Applied Water Engineering and Research* **2020**, *8*, 313–332. https://doi.org/10.1080/23249676.2020.1831976.
40. Chen, J.; Yin, H.; Zhang, D. A Self-Adaptive Classification Method for Plant Disease Detection Using GMDH-Logistic Model. *Sustainable Computing: Informatics and Systems* **2020**, *28*, 100415. https://doi.org/10.1016/j.suscom.2020.100415.
41. INSA O Semiárido Brasileiro Available online: https://www.gov.br/insa/pt-br/semiarido-brasileiro/o-semiarido-brasileiro (accessed on 1 December 2023).
42. IBGE Ceará | Cidades e Estados | IBGE Available online: https://www.ibge.gov.br/cidades-e-estados/ce.html (accessed on 1 December 2023).
43. Alvalá, R.C.S.; Cunha, A.P.M.A.; Brito, S.S.B.; Seluchi, M.E.; Marengo, J.A.; Moraes, O.L.L.; Carvalho, M.A. Drought Monitoring in the Brazilian Semiarid Region. *An. Acad. Bras. Ciênc.* **2019**, *91*, e20170209. https://doi.org/10.1590/0001-3765201720170209.
44. Pontes Filho, J.D.; Souza Filho, F.D.A.; Martins, E.S.P.R.; Studart, T.M.D.C. Copula-Based Multivariate Frequency Analysis of the 2012–2018 Drought in Northeast Brazil. *Water* **2020**, *12*, 834. https://doi.org/10.3390/w12030834.
45. Beck, H.E.; Zimmermann, N.E.; McVicar, T.R.; Vergopolan, N.; Berg, A.; Wood, E.F. Present and Future Köppen-Geiger Climate Classification Maps at 1-Km Resolution. *Sci Data* **2018**, *5*, 180214. https://doi.org/10.1038/sdata.2018.214.
46. Sacramento, E.M. do; Carvalho, P.C.M.; de Araújo, J.C.; Riffel, D.B.; Corrêa, R.M. da C.; Pinheiro Neto, J.S. Scenarios for Use of Floating Photovoltaic Plants in Brazilian Reservoirs. *IET Renewable Power Generation* **2015**, *9*, 1019–1024. https://doi.org/10.1049/iet-rpg.2015.0120.
47. Raulino, J.B.S.; Silveira, C.S.; E.L. Neto, I. Eutrophication Risk Assessment of a Large Reservoir in the Brazilian Semiarid Region under Climate Change Scenarios. *An. Acad. Bras. Ciênc.* **2022**, *94*, e20201689. https://doi.org/10.1590/0001-3765202220201689.
48. COGERH Portal Hidrológico Do Ceará Available online: http://www.hidro.ce.gov.br/ (accessed on 1 December 2023).
49. COGERH Matriz Dos Usos Mútiplos Dos Açudes Available online: http://www.hidro.ce.gov.br/hidro-ce-zend/mi/midia/show/149 (accessed on 1 December 2023).
50. COGERH Qualidade Das Águas Dos Açudes Monitorados Pela COGERH;
51. *Standard Methods for the Examination of Water and Wastewater*; American public health association, American water works association, Water environment federation, Eds.; 22nd ed.; American public health association: Washington (D.C.), 2012; ISBN 978-0-87553-013-0.
52. Phiri, D.; Simwanda, M.; Salekin, S.; Nyirenda, V.; Murayama, Y.; Ranagalage, M. Sentinel-2 Data for Land Cover/Use Mapping: A Review. *Remote Sensing* **2020**, *12*, 2291. https://doi.org/10.3390/rs12142291.

53. Drusch, M.; Del Bello, U.; Carlier, S.; Colin, O.; Fernandez, V.; Gascon, F.; Hoersch, B.; Isola, C.; Laberinti, P.; Martimort, P.; et al. Sentinel-2: ESA's Optical High-Resolution Mission for GMES Operational Services. *Remote Sensing of Environment* **2012**, *120*, 25–36. https://doi.org/10.1016/j.rse.2011.11.026.
54. Zhang, T.; Su, J.; Liu, C.; Chen, W.-H.; Liu, H.; Liu, G. Band Selection in Sentinel-2 Satellite for Agriculture Applications. In Proceedings of the 2017 23rd International Conference on Automation and Computing (ICAC); IEEE: Huddersfield, United Kingdom, September 2017; pp. 1–6.
55. Ienco, D.; Interdonato, R.; Gaetano, R.; Ho Tong Minh, D. Combining Sentinel-1 and Sentinel-2 Satellite Image Time Series for Land Cover Mapping via a Multi-Source Deep Learning Architecture. *ISPRS Journal of Photogrammetry and Remote Sensing* **2019**, *158*, 11–22. https://doi.org/10.1016/j.isprsjprs.2019.09.016.
56. Zhang, T.-X.; Su, J.-Y.; Liu, C.-J.; Chen, W.-H. Potential Bands of Sentinel-2A Satellite for Classification Problems in Precision Agriculture. *Int. J. Autom. Comput.* **2019**, *16*, 16–26. https://doi.org/10.1007/s11633-018-1143-x.
57. Ma, Y.; Xu, N.; Liu, Z.; Yang, B.; Yang, F.; Wang, X.H.; Li, S. Satellite-Derived Bathymetry Using the ICESat-2 Lidar and Sentinel-2 Imagery Datasets. *Remote Sensing of Environment* **2020**, *250*, 112047. https://doi.org/10.1016/j.rse.2020.112047.
58. Prasad, A.D.; Ganasala, P.; Hernández-Guzmán, R.; Fathian, F. Remote Sensing Satellite Data and Spectral Indices: An Initial Evaluation for the Sustainable Development of an Urban Area. *Sustain. Water Resour. Manag.* **2022**, *8*, 19. https://doi.org/10.1007/s40899-022-00607-2.
59. Gitelson, A. The Peak near 700 Nm on Radiance Spectra of Algae and Water: Relationships of Its Magnitude and Position with Chlorophyll Concentration. *International Journal of Remote Sensing* **1992**, *13*, 3367–3373. https://doi.org/10.1080/01431169208904125.
60. Hamunzala, B.; Matsumoto, K.; Nagai, K. Improved Method for Estimating Construction Year of Road Bridges by Analyzing Landsat Normalized Difference Water Index 2. *Remote Sensing* **2023**, *15*, 3488. https://doi.org/10.3390/rs15143488.
61. Abd El-Sadek, E.; Elbeih, S.; Negm, A. Coastal and Landuse Changes of Burullus Lake, Egypt: A Comparison Using Landsat and Sentinel-2 Satellite Images. *The Egyptian Journal of Remote Sensing and Space Science* **2022**, *25*, 815–829. https://doi.org/10.1016/j.ejrs.2022.07.006.
62. Moravec, D.; Komárek, J.; López-Cuervo Medina, S.; Molina, I. Effect of Atmospheric Corrections on NDVI: Intercomparability of Landsat 8, Sentinel-2, and UAV Sensors. *Remote Sensing* **2021**, *13*, 3550. https://doi.org/10.3390/rs13183550.
63. Huang, S.; Tang, L.; Hupy, J.P.; Wang, Y.; Shao, G. A Commentary Review on the Use of Normalized Difference Vegetation Index (NDVI) in the Era of Popular Remote Sensing. *J. For. Res.* **2021**, *32*, 1–6. https://doi.org/10.1007/s11676-020-01155-1.
64. Gitelson, A.A.; Kaufman, Y.J.; Merzlyak, M.N. Use of a Green Channel in Remote Sensing of Global Vegetation from EOS-MODIS. *Remote Sensing of Environment* **1996**, *58*, 289–298. https://doi.org/10.1016/S0034-4257(96)00072-7.
65. Ge, Y.; Atefi, A.; Zhang, H.; Miao, C.; Ramamurthy, R.K.; Sigmon, B.; Yang, J.; Schnable, J.C. High-Throughput Analysis of Leaf Physiological and Chemical Traits with VIS–NIR–SWIR Spectroscopy: A Case Study with a Maize Diversity Panel. *Plant Methods* **2019**, *15*, 66. https://doi.org/10.1186/s13007-019-0450-8.
66. Huete, A. A Comparison of Vegetation Indices over a Global Set of TM Images for EOS-MODIS. *Remote Sensing of Environment* **1997**, *59*, 440–451. https://doi.org/10.1016/S0034-4257(96)00112-5.
67. Zhen, Z.; Chen, S.; Yin, T.; Gastellu-Etchegorry, J.-P. Globally Quantitative Analysis of the Impact of Atmosphere and Spectral Response Function on 2-Band Enhanced Vegetation Index (EVI2) over Sentinel-2 and Landsat-8. *ISPRS Journal of Photogrammetry and Remote Sensing* **2023**, *205*, 206–226. https://doi.org/10.1016/j.isprsjprs.2023.09.024.
68. Huete, A.R. A Soil-Adjusted Vegetation Index (SAVI). *Remote Sensing of Environment* **1988**, *25*, 295–309. https://doi.org/10.1016/0034-4257(88)90106-X.
69. Ghazaryan, G.; Dubovyk, O.; Graw, V.; Kussul, N.; Schellberg, J. Local-Scale Agricultural Drought Monitoring with Satellite-Based Multi-Sensor Time-Series. *GIScience & Remote Sensing* **2020**, *57*, 704–718. https://doi.org/10.1080/15481603.2020.1778332.
70. Lastovicka, J.; Svec, P.; Paluba, D.; Kobliuk, N.; Svoboda, J.; Hladky, R.; Stych, P. Sentinel-2 Data in an Evaluation of the Impact of the Disturbances on Forest Vegetation. *Remote Sensing* **2020**, *12*, 1914. https://doi.org/10.3390/rs12121914.
71. Welikhe, P.; Quansah, J.E.; Fall, S.; McElhenney, W. Estimation of Soil Moisture Percentage Using LANDSAT-Based Moisture Stress Index. *J Remote Sensing & GIS* **2017**, *06*. https://doi.org/10.4172/2469-4134.1000200.
72. Hunt, Jr., E.; Rock, B. Detection of Changes in Leaf Water Content Using Near- and Middle-Infrared Reflectances☆. *Remote Sensing of Environment* **1989**, *30*, 43–54. https://doi.org/10.1016/0034-4257(89)90046-1.
73. Gitelson, A.A.; Gritz †, Y.; Merzlyak, M.N. Relationships between Leaf Chlorophyll Content and Spectral Reflectance and Algorithms for Non-Destructive Chlorophyll Assessment in Higher Plant Leaves. *Journal of Plant Physiology* **2003**, *160*, 271–282. https://doi.org/10.1078/0176-1617-00887.

74. Vasudeva, V.; Nandy, S.; Padalia, H.; Srinet, R.; Chauhan, P. Mapping Spatial Variability of Foliar Nitrogen and Carbon in Indian Tropical Moist Deciduous Sal (Shorea Robusta) Forest Using Machine Learning Algorithms and Sentinel-2 Data. *International Journal of Remote Sensing* **2021**, *42*, 1139–1159. https://doi.org/10.1080/01431161.2020.1823043.
75. Escuin, S.; Navarro, R.; Fernández, P. Fire Severity Assessment by Using NBR (Normalized Burn Ratio) and NDVI (Normalized Difference Vegetation Index) Derived from LANDSAT TM/ETM Images. *International Journal of Remote Sensing* **2008**, *29*, 1053–1073. https://doi.org/10.1080/01431160701281072.
76. Meneses, B.M. Vegetation Recovery Patterns in Burned Areas Assessed with Landsat 8 OLI Imagery and Environmental Biophysical Data. *Fire* **2021**, *4*, 76. https://doi.org/10.3390/fire4040076.
77. Xu, N.; Tian, J.; Tian, Q.; Xu, K.; Tang, S. Analysis of Vegetation Red Edge with Different Illuminated/Shaded Canopy Proportions and to Construct Normalized Difference Canopy Shadow Index. *Remote Sensing* **2019**, *11*, 1192. https://doi.org/10.3390/rs11101192.
78. Saha, S.; Saha, M.; Mukherjee, K.; Arabameri, A.; Ngo, P.T.T.; Paul, G.C. Predicting the Deforestation Probability Using the Binary Logistic Regression, Random Forest, Ensemble Rotational Forest, REPTree: A Case Study at the Gumani River Basin, India. *Science of The Total Environment* **2020**, *730*, 139197. https://doi.org/10.1016/j.scitotenv.2020.139197.
79. McFeeters, S.K. The Use of the Normalized Difference Water Index (NDWI) in the Delineation of Open Water Features. *International Journal of Remote Sensing* **1996**, *17*, 1425–1432. https://doi.org/10.1080/01431169608948714.
80. Yang, X.; Zhao, S.; Qin, X.; Zhao, N.; Liang, L. Mapping of Urban Surface Water Bodies from Sentinel-2 MSI Imagery at 10 m Resolution via NDWI-Based Image Sharpening. *Remote Sensing* **2017**, *9*, 596. https://doi.org/10.3390/rs9060596.
81. Dozier, J. Spectral Signature of Alpine Snow Cover from the Landsat Thematic Mapper. *Remote Sensing of Environment* **1989**, *28*, 9–22. https://doi.org/10.1016/0034-4257(89)90101-6.
82. Salomonson, V.V.; Appel, I. Estimating Fractional Snow Cover from MODIS Using the Normalized Difference Snow Index. *Remote Sensing of Environment* **2004**, *89*, 351–360. https://doi.org/10.1016/j.rse.2003.10.016.
83. Gascoin, S.; Grizonnet, M.; Bouchet, M.; Salgues, G.; Hagolle, O. Theia Snow Collection: High-Resolution Operational Snow Cover Maps from Sentinel-2 and Landsat-8 Data. *Earth System Science Data* **2019**, *11*, 493–514. https://doi.org/10.5194/essd-11-493-2019.
84. Keshri, A.K.; Shukla, A.; Gupta, R.P. ASTER Ratio Indices for Supraglacial Terrain Mapping. *International Journal of Remote Sensing* **2009**, *30*, 519–524. https://doi.org/10.1080/01431160802385459.
85. Dirscherl, M.; Dietz, A.J.; Kneisel, C.; Kuenzer, C. Automated Mapping of Antarctic Supraglacial Lakes Using a Machine Learning Approach. *Remote Sensing* **2020**, *12*, 1203. https://doi.org/10.3390/rs12071203.
86. Kaufman, Y.J.; Tanre, D. Atmospherically Resistant Vegetation Index (ARVI) for EOS-MODIS. *IEEE Trans. Geosci. Remote Sensing* **1992**, *30*, 261–270. https://doi.org/10.1109/36.134076.
87. Somvanshi, S.S.; Kumari, M. Comparative Analysis of Different Vegetation Indices with Respect to Atmospheric Particulate Pollution Using Sentinel Data. *Applied Computing and Geosciences* **2020**, *7*, 100032. https://doi.org/10.1016/j.acags.2020.100032.
88. Penuelas, J.; Frederic, B.; Filella, I. Semi-Empirical Indices to Assess Carotenoids/Chlorophyll a Ratio from Leaf Spectral Reflectance. **2013**.
89. Zhang, N.; Su, X.; Zhang, X.; Yao, X.; Cheng, T.; Zhu, Y.; Cao, W.; Tian, Y. Monitoring Daily Variation of Leaf Layer Photosynthesis in Rice Using UAV-Based Multi-Spectral Imagery and a Light Response Curve Model. *Agricultural and Forest Meteorology* **2020**, *291*, 108098. https://doi.org/10.1016/j.agrformet.2020.108098.
90. Robak, A.; Gadawska, A.; Milczarek, M.; Lewiński, S. The detection of water on Sentinel-2 imagery.
91. Dawoud, I.; Abonazel, M.R. Robust Dawoud–Kibria Estimator for Handling Multicollinearity and Outliers in the Linear Regression Model. *Journal of Statistical Computation and Simulation* **2021**, *91*, 3678–3692. https://doi.org/10.1080/00949655.2021.1945063.
92. Chan, J.Y.-L.; Leow, S.M.H.; Bea, K.T.; Cheng, W.K.; Phoong, S.W.; Hong, Z.-W.; Chen, Y.-L. Mitigating the Multicollinearity Problem and Its Machine Learning Approach: A Review. *Mathematics* **2022**, *10*, 1283. https://doi.org/10.3390/math10081283.
93. Ghojogh, B.; Crowley, M. The Theory Behind Overfitting, Cross Validation, Regularization, Bagging, and Boosting: Tutorial 2023.
94. Breiman, L. Random Forests. *Machine Learning* **2001**, *45*, 5–32. https://doi.org/10.1023/A:1010933404324.
95. Liaw, A.; Wiener, M. Classification and Regression by randomForest. **2002**, *2*.
96. Gislason, P.O.; Benediktsson, J.A.; Sveinsson, J.R. Random Forests for Land Cover Classification. *Pattern Recognition Letters* **2006**, *27*, 294–300. https://doi.org/10.1016/j.patrec.2005.08.011.
97. Chen, X.; Ishwaran, H. Random Forests for Genomic Data Analysis. *Genomics* **2012**, *99*, 323–329. https://doi.org/10.1016/j.ygeno.2012.04.003.

98. Mei, J.; He, D.; Harley, R.; Habetler, T.; Qu, G. A Random Forest Method for Real-Time Price Forecasting in New York Electricity Market. In Proceedings of the 2014 IEEE PES General Meeting | Conference & Exposition; IEEE: National Harbor, MD, USA, July 2014; pp. 1–5.
99. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; ACM: San Francisco California USA, August 13 2016; pp. 785–794.
100. Bentéjac, C.; Csörgő, A.; Martínez-Muñoz, G. A Comparative Analysis of Gradient Boosting Algorithms. *Artif Intell Rev* **2021**, *54*, 1937–1967. https://doi.org/10.1007/s10462-020-09896-5.
101. Dai, H.; Huang, G.; Zeng, H.; Zhou, F. PM2.5 Volatility Prediction by XGBoost-MLP Based on GARCH Models. *Journal of Cleaner Production* **2022**, *356*, 131898. https://doi.org/10.1016/j.jclepro.2022.131898.
102. Zhang, C.; Hu, D.; Yang, T. Anomaly Detection and Diagnosis for Wind Turbines Using Long Short-Term Memory-Based Stacked Denoising Autoencoders and XGBoost. *Reliability Engineering & System Safety* **2022**, *222*, 108445. https://doi.org/10.1016/j.ress.2022.108445.
103. Sanders, W.; Li, D.; Li, W.; Fang, Z.N. Data-Driven Flood Alert System (FAS) Using Extreme Gradient Boosting (XGBoost) to Forecast Flood Stages. *Water* **2022**, *14*, 747. https://doi.org/10.3390/w14050747.
104. Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; MIT Press, 2016; ISBN 978-0-262-03561-3.
105. Song, Y.; Liang, J.; Lu, J.; Zhao, X. An Efficient Instance Selection Algorithm for k Nearest Neighbor Regression. *Neurocomputing* **2017**, *251*, 26–34. https://doi.org/10.1016/j.neucom.2017.04.018.
106. Huang, L.; Song, T.; Jiang, T. Linear Regression Combined KNN Algorithm to Identify Latent Defects for Imbalance Data of ICs. *Microelectronics Journal* **2023**, *131*, 105641. https://doi.org/10.1016/j.mejo.2022.105641.
107. James, G.; Witten, D.; Hastie, T.; Tibshirani, R.; Taylor, J. *An Introduction to Statistical Learning: With Applications in Python*; Springer International Publishing, 2023; ISBN 978-3-031-38746-3.
108. Uddin, S.; Haque, I.; Lu, H.; Moni, M.A.; Gide, E. Comparative Performance Analysis of K-Nearest Neighbour (KNN) Algorithm and Its Different Variants for Disease Prediction. *Sci Rep* **2022**, *12*, 6256. https://doi.org/10.1038/s41598-022-10358-x.
109. Cai, L.; Yu, Y.; Zhang, S.; Song, Y.; Xiong, Z.; Zhou, T. A Sample-Rebalanced Outlier-Rejected k -Nearest Neighbor Regression Model for Short-Term Traffic Flow Forecasting. *IEEE Access* **2020**, *8*, 22686–22696. https://doi.org/10.1109/ACCESS.2020.2970250.
110. Liu, W.; Wang, P.; Meng, Y.; Zhao, C.; Zhang, Z. Cloud Spot Instance Price Prediction Using kNN Regression. *Hum. Cent. Comput. Inf. Sci.* **2020**, *10*, 34. https://doi.org/10.1186/s13673-020-00239-5.
111. Ho, W.T.; Yu, F.W. Chiller System Optimization Using k Nearest Neighbour Regression. *Journal of Cleaner Production* **2021**, *303*, 127050. https://doi.org/10.1016/j.jclepro.2021.127050.
112. Géron, A. Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow; O'Reilly Media, Inc., 2022; ISBN 978-1-09-812246-1.
113. Cortes, C.; Vapnik, V. Support-Vector Networks. *Mach Learn* **1995**, *20*, 273–297. https://doi.org/10.1007/BF00994018.
114. Chollet, F. *Deep Learning with Python, Second Edition*; Simon and Schuster, 2021; ISBN 978-1-63835-009-5.
115. Tanveer, M.; Rajani, T.; Rastogi, R.; Shao, Y.H.; Ganaie, M.A. Comprehensive Review on Twin Support Vector Machines. *Ann Oper Res* **2022**. https://doi.org/10.1007/s10479-022-04575-w.
116. Bansal, M.; Goyal, A.; Choudhary, A. A Comparative Analysis of K-Nearest Neighbor, Genetic, Support Vector Machine, Decision Tree, and Long Short Term Memory Algorithms in Machine Learning. *Decision Analytics Journal* **2022**, *3*, 100071. https://doi.org/10.1016/j.dajour.2022.100071.
117. Manoharan, A.; Begam, K.M.; Aparow, V.R.; Sooriamoorthy, D. Artificial Neural Networks, Gradient Boosting and Support Vector Machines for Electric Vehicle Battery State Estimation: A Review. *Journal of Energy Storage* **2022**, *55*, 105384. https://doi.org/10.1016/j.est.2022.105384.
118. Tibshirani, R. Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* **1996**, *58*, 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x.
119. Ivakhnenko, A.G. The Group Method of Data Handling, A Rival of the Method of Stochastic Approximation. *Soviet Automatic Control* **1968**, *13*, 43–55.
120. Walton, R.; Binns, A.; Bonakdari, H.; Ebtehaj, I.; Gharabaghi, B. Estimating 2-Year Flood Flows Using the Generalized Structure of the Group Method of Data Handling. *Journal of Hydrology* **2019**, *575*, 671–689. https://doi.org/10.1016/j.jhydrol.2019.05.068.
121. Elkurdy, M.; Binns, A.D.; Bonakdari, H.; Gharabaghi, B.; McBean, E. Early Detection of Riverine Flooding Events Using the Group Method of Data Handling for the Bow River, Alberta, Canada. *International Journal of River Basin Management* **2022**, *20*, 533–544. https://doi.org/10.1080/15715124.2021.1906261.
122. Zaji, A.H.; Bonakdari, H.; Gharabaghi, B. Reservoir Water Level Forecasting Using Group Method of Data Handling. *Acta Geophys.* **2018**, *66*, 717–730. https://doi.org/10.1007/s11600-018-0168-4.
123. Azimi, H.; Bonakdari, H.; Ebtehaj, I.; Gharabaghi, B.; Khoshbin, F. Evolutionary Design of Generalized Group Method of Data Handling-Type Neural Network for Estimating the Hydraulic Jump Roller Length. *Acta Mech* **2018**, *229*, 1197–1214. https://doi.org/10.1007/s00707-017-2043-9.

124. Stajkowski, S.; Laleva, A.; Farghaly, H.; Bonakdari, H.; Gharabaghi, B. Modelling Dry-Weather Temperature Profiles in Urban Stormwater Management Ponds. *Journal of Hydrology* **2021**, *598*, 126206. https://doi.org/10.1016/j.jhydrol.2021.126206.
125. Stajkowski, S.; Hotson, E.; Zorica, M.; Farghaly, H.; Bonakdari, H.; McBean, E.; Gharabaghi, B. Modeling Stormwater Management Pond Thermal Impacts during Storm Events. *Journal of Hydrology* **2023**, *620*, 129413. https://doi.org/10.1016/j.jhydrol.2023.129413.
126. Bonakdari, H.; Ebtehaj, I.; Gharabaghi, B.; Vafaeifard, M.; Akhbari, A. Calculating the Energy Consumption of Electrocoagulation Using a Generalized Structure Group Method of Data Handling Integrated with a Genetic Algorithm and Singular Value Decomposition. *Clean Techn Environ Policy* **2019**, *21*, 379–393. https://doi.org/10.1007/s10098-018-1642-z.
127. Ashrafzadeh, A.; Kişi, O.; Aghelpour, P.; Biazar, S.M.; Masouleh, M.A. Comparative Study of Time Series Models, Support Vector Machines, and GMDH in Forecasting Long-Term Evapotranspiration Rates in Northern Iran. *J. Irrig. Drain Eng.* **2020**, *146*, 04020010. https://doi.org/10.1061/(ASCE)IR.1943-4774.0001471.
128. Ebtehaj, I.; Sammen, S.Sh.; Sidek, L.M.; Malik, A.; Sihag, P.; Al-Janabi, A.M.S.; Chau, K.-W.; Bonakdari, H. Prediction of Daily Water Level Using New Hybridized GS-GMDH and ANFIS-FCM Models. *Engineering Applications of Computational Fluid Mechanics* **2021**, *15*, 1343–1361. https://doi.org/10.1080/19942060.2021.1966837.
129. Wang, W.; Du, Y.; Chau, K.; Chen, H.; Liu, C.; Ma, Q. A Comparison of BPNN, GMDH, and ARIMA for Monthly Rainfall Forecasting Based on Wavelet Packet Decomposition. *Water* **2021**, *13*, 2871. https://doi.org/10.3390/w13202871.
130. Weisberg, S. *Applied Linear Regression*; John Wiley & Sons, 2005; ISBN 978-0-471-70408-9.
131. Yang, D.; Kleissl, J.; Gueymard, C.A.; Pedro, H.T.C.; Coimbra, C.F.M. History and Trends in Solar Irradiance and PV Power Forecasting: A Preliminary Assessment and Review Using Text Mining. *Solar Energy* **2018**, *168*, 60–101. https://doi.org/10.1016/j.solener.2017.11.023.
132. Yeo, I.-K.; Johnson, R.A. A New Family of Power Transformations to Improve Normality or Symmetry. *Biometrika* **2000**, *87*, 954–959.
133. He, Y.; Zheng, Y. Short-Term Power Load Probability Density Forecasting Based on Yeo-Johnson Transformation Quantile Regression and Gaussian Kernel Function. *Energy* **2018**, *154*, 143–156. https://doi.org/10.1016/j.energy.2018.04.072.
134. Vidal Batista, L. Turbidity classification of the Paraopeba River using machine learning and Sentinel-2 images. *IEEE Latin Am. Trans.* **2022**, *20*, 799–805. https://doi.org/10.1109/TLA.2022.9693564.
135. Chicco, D.; Warrens, M.J.; Jurman, G. The Coefficient of Determination R-Squared Is More Informative than SMAPE, MAE, MAPE, MSE and RMSE in Regression Analysis Evaluation. *PeerJ Computer Science* **2021**, *7*, e623. https://doi.org/10.7717/peerj-cs.623.
136. Mendonça, J.C.D.; Lopes, F.B.; Andrade, E.M.D.; Praxedes, C.F.; Lima, F.J.D.O.; Silva, F.H.O.D. Qualitative Vulnerability of the Waters of a Surface Reservoir Subjected to Drought in a Tropical Semi-Arid Region. *RCA* **2023**, *54*. https://doi.org/10.5935/1806-6690.20230033.
137. Nunes Carvalho, T.M.; Lima Neto, I.E.; Souza Filho, F.D.A. Uncovering the Influence of Hydrological and Climate Variables in Chlorophyll-A Concentration in Tropical Reservoirs with Machine Learning. *Environ Sci Pollut Res* **2022**, *29*, 74967–74982. https://doi.org/10.1007/s11356-022-21168-z.
138. Wilkinson, G.M.; Walter, J.A.; Buelo, C.D.; Pace, M.L. No Evidence of Widespread Algal Bloom Intensification in Hundreds of Lakes. *Frontiers in Ecology and the Environment* **2022**, *20*, 16–21. https://doi.org/10.1002/fee.2421.
139. Zhu, X.; Guo, H.; Huang, J.J.; Tian, S.; Xu, W.; Mai, Y. An Ensemble Machine Learning Model for Water Quality Estimation in Coastal Area Based on Remote Sensing Imagery. *Journal of Environmental Management* **2022**, *323*, 116187. https://doi.org/10.1016/j.jenvman.2022.116187.
140. Woo Kim, Y.; Kim, T.; Shin, J.; Lee, D.-S.; Park, Y.-S.; Kim, Y.; Cha, Y. Validity Evaluation of a Machine-Learning Model for Chlorophyll a Retrieval Using Sentinel-2 from Inland and Coastal Waters. *Ecological Indicators* **2022**, *137*, 108737. https://doi.org/10.1016/j.ecolind.2022.108737.
141. Ha, N.T.T.; Thao, N.T.P.; Koike, K.; Nhuan, M.T. Selecting the Best Band Ratio to Estimate Chlorophyll-a Concentration in a Tropical Freshwater Lake Using Sentinel 2A Images from a Case Study of Lake Ba Be (Northern Vietnam). *IJGI* **2017**, *6*, 290. https://doi.org/10.3390/ijgi6090290.
142. Cillero Castro, C.; Domínguez Gómez, J.A.; Delgado Martín, J.; Hinojo Sánchez, B.A.; Cereijo Arango, J.L.; Cheda Tuya, F.A.; Díaz-Varela, R. An UAV and Satellite Multispectral Data Approach to Monitor Water Quality in Small Reservoirs. *Remote Sensing* **2020**, *12*, 1514. https://doi.org/10.3390/rs12091514.
143. Buma, W.G.; Lee, S.-I. Evaluation of Sentinel-2 and Landsat 8 Images for Estimating Chlorophyll-a Concentrations in Lake Chad, Africa. *Remote Sensing* **2020**, *12*, 2437. https://doi.org/10.3390/rs12152437.
144. Aubriot, L.; Zabaleta, B.; Bordet, F.; Sienra, D.; Risso, J.; Achkar, M.; Somma, A. Assessing the Origin of a Massive Cyanobacterial Bloom in the Río de La Plata (2019): Towards an Early Warning System. *Water Research* **2020**, *181*, 115944. https://doi.org/10.1016/j.watres.2020.115944.

145. Viso-Vázquez, M.; Acuña-Alonso, C.; Rodríguez, J.L.; Álvarez, X. Remote Detection of Cyanobacterial Blooms and Chlorophyll-a Analysis in a Eutrophic Reservoir Using Sentinel-2. *Sustainability* **2021**, *13*, 8570. https://doi.org/10.3390/su13158570.
146. Nevo, S.; Morin, E.; Gerzi Rosenthal, A.; Metzger, A.; Barshai, C.; Weitzner, D.; Voloshin, D.; Kratzert, F.; Elidan, G.; Dror, G.; et al. Flood Forecasting with Machine Learning Models in an Operational Framework. *Hydrology and Earth System Sciences* **2022**, *26*, 4013–4032. https://doi.org/10.5194/hess-26-4013-2022.
147. Guo, H.; Zhu, X.; Jeanne Huang, J.; Zhang, Z.; Tian, S.; Chen, Y. An Enhanced Deep Learning Approach to Assessing Inland Lake Water Quality and Its Response to Climate and Anthropogenic Factors. *Journal of Hydrology* **2023**, *620*, 129466. https://doi.org/10.1016/j.jhydrol.2023.129466.
148. Aptoula, E.; Ariman, S. Chlorophyll-a Retrieval From Sentinel-2 Images Using Convolutional Neural Network Regression. *IEEE Geosci. Remote Sensing Lett.* **2022**, *19*, 1–5. https://doi.org/10.1109/LGRS.2021.3070437.
149. Maier, P.M.; Keller, S.; Hinz, S. Deep Learning with WASI Simulation Data for Estimating Chlorophyll a Concentration of Inland Water Bodies. *Remote Sensing* **2021**, *13*, 718. https://doi.org/10.3390/rs13040718.
150. 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. *Science of The Total Environment* **2021**, *778*, 146271. https://doi.org/10.1016/j.scitotenv.2021.146271.
151. Chegoonian, A.M.; Pahlevan, N.; Zolfaghari, K.; Leavitt, P.R.; Davies, J.-M.; Baulch, H.M.; Duguay, C.R. Comparative Analysis of Empirical and Machine Learning Models for Chl *a* Extraction Using Sentinel-2 and Landsat OLI Data: Opportunities, Limitations, and Challenges. *Canadian Journal of Remote Sensing* **2023**, *49*, 2215333. https://doi.org/10.1080/07038992.2023.2215333.
152. Maier, P.M.; Keller, S. SpecWa: Spectral Remote Sensing Data and Chlorophyll a Values of Inland Waters.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.