

Article

Sentinel-2 Retrieval of Chlorophyll-a and Total Suspended Solids in Optically Complex Inland Waters: A Case Study of Lake Palic

Mihajlo Miroslavljević ¹ , Mirjana Horvat ^{1,2,*} , Zoltan Horvat ^{1,2}  and Koch Dániel ²

¹ Faculty of Civil Engineering Subotica, University of Novi Sad, Kozaračka 2a, 24000 Subotica, Serbia; mihajlomiroslavljevic@gmail.com (M.M.); horvatz@gf.uns.ac.rs (Z.H.)

² Faculty of Water Sciences, Ludovika University of Public Service, 6500 Baja, Hungary; koch.daniel@uni-nke.hu

* Correspondence: isicm@gf.uns.ac.rs or isic.mirjana@gmail.com

Abstract

This study focuses on the development of satellite-image-based models for estimating Chl-a and TSS concentrations in Lake Palic. Using Sentinel-2 images, we ensured high spatial resolution. Image processing was conducted using the SNAP (Sentinel Application Platform version 10.0.0, developed by the European Space Agency (ESA), Frascati, Italy) software with the C2RCC algorithm. A series of environmental indices was calculated and correlated with measured Chl-a and TSS values to support the development of a reliable predictive model. Two modeling approaches were applied: simple linear regression based on selected spectral indices and measured Chl-a and TSS concentrations, and machine learning methods to improve predictive performance and capture potentially nonlinear relationships. The models were evaluated using standard metrics and verified against field measurements. The overall process was implemented at two locations within the lake, the middle and the outflow. Spatial variability in optical conditions prevented the development of a single unified model. While Chl-a concentrations could be estimated with reasonable accuracy, TSS retrieval remained limited by optical complexity and spatial variability in water quality conditions. This suggests that the lake exhibits spatially variable optical regimes, in which different optically active constituents dominate the spectral response, thereby directly influencing model performance.

Keywords: remote sensing; optical complexity; Chl-a; TSS; machine learning; linear regression



Academic Editor: Nir Krakauer

Received: 9 May 2026

Revised: 5 June 2026

Accepted: 20 June 2026

Published: 2 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

High human impact on most aquatic ecosystems [1] leads to notable changes in nutrient transfer from terrestrial to aquatic environments, ultimately altering ecosystem functioning [2]. Extensive agricultural practices [3,4] lead to increased nutrient loading (nitrogen (N) and phosphorus (P)) and accelerated eutrophication of surrounding aquatic ecosystems [2,5]. Eutrophication affects both freshwater and coastal marine environments and can devalue the benefits and services these aquatic systems provide [6].

Lake Palic suffered severe eutrophication during the 20th century, culminating in a mass fish kill in 1970. Following this event, the lake was drained, the sludge removed, and refilled in 1976. Pollution persisted in subsequent decades. From 2019 to 2022, revitalization measures were implemented, including the removal of the invasive fish species (Prussian carp), the construction of sewerage infrastructure, and the implementation of a

lake protection zone. This resulted in improved water quality from class V to class III [7], but issues due to excess nutrients and algal blooms remain.

The findings from Simeunović [8] provide evidence regarding the presence and severity of cyanobacterial toxins in Lake Palic. Her study analyzed the toxicity of isolated cyanobacterial strains, revealing a significant level of Microcystin-LR. Lake Palic contained high levels of this toxin, exceeding limits suggested by the World Health Organization's (WHO) [9] long-term recreational exposure guidelines.

Given that traditional field measurements are spatially or temporally limited by economic, organizational, or environmental constraints, satellite remote sensing has emerged as a powerful tool to overcome these limitations. However, the application of remote sensing in inland waters remains challenging due to the optical complexity, as multiple optically active constituents simultaneously contribute to the spectral signal.

Aquatic remote sensing is based on the analysis of the spectral signature of a given water body, governed by the concentration and optical properties of Optically Active Components (OACs), expressed as the Remote Sensing Reflectance [10–14]. The dominant OACs in most inland and coastal waters are chlorophyll *a* (Chl-*a*), total suspended solids (TSS), and colored dissolved organic matter (CDOM). The strong absorption of light in one spectral region and scattering in another result in the development of the characteristic spectral “valley-peak” pattern [15–17]. In optically complex Case 2 waters, typical of shallow and turbid lakes, high concentrations of multiple OACs complicate the separation of individual spectral contributions.

A major advancement in satellite-based water quality monitoring occurred in 2015 with the launch of the Sentinel-2 mission. Equipped with the Multispectral Imager (MSI), Sentinel-2 provides high spatial resolution (10–60 m) and a short revisit period (5 days), making it particularly suitable for monitoring small- to medium-sized inland water bodies such as Lake Palic [18].

The application of satellite remote sensing for the estimation of Chl-*a* concentration—a primary indicator of phytoplankton biomass—has been extensively validated due to its strong optical signal [19–23]. Similarly, TSS retrieval has been successfully applied to characterize particulate matter dynamics and water clarity in both marine [24] and inland aquatic environments [25].

Recent studies have increasingly demonstrated the potential of machine learning approaches for retrieving water quality parameters from satellite imagery in optically complex aquatic environments. Ibrahim et al. [26] investigated the performance of several machine learning algorithms, including Random Forest Regression (RFR), Support Vector Regression (SVR), Extreme Gradient Boosting (XGB), and Partial Least Squares Regression (PLSR), for estimating chlorophyll-*a* (Chl-*a*) and total suspended solids (TSS) concentrations from Sentinel-2 imagery in the coastal waters of Abu Dhabi. Their results showed that ensemble learning approaches, particularly XGB and RFR, outperformed traditional regression methods, highlighting the ability of machine learning algorithms to capture complex nonlinear relationships between spectral reflectance and water quality parameters. Similarly, Guo et al. [27] compared Mathematical Regression Analysis (MRA) and Artificial Neural Networks (ANN) for Chl-*a* retrieval using Landsat-8 OLI imagery in the eutrophic Haihe River, China. Although both approaches produced acceptable results, the ANN model substantially outperformed the traditional regression model, demonstrating the advantages of machine learning techniques for modeling nonlinear spectral responses in productive inland waters. Xu et al. [28] evaluated Random Forest (RF), Support Vector Regression (SVR), and Back Propagation Neural Networks (BPNN) for Chl-*a* estimation in Xinkai Lake and reported superior performance of the RF model compared to the other evaluated approaches. Collectively, these studies indicate that machine learning methods

are particularly suitable for Case 2 and eutrophic waters, where the optical signal is influenced by multiple interacting constituents and cannot be adequately described by simple empirical relationships.

At the same time, the performance of retrieval models remains highly dependent on local optical conditions, trophic status, constituent composition, and the spectral characteristics of the available sensors. Consequently, algorithms that perform well in one aquatic environment cannot be assumed to be directly transferable to another without site-specific evaluation and calibration.

In this context, Lake Palic represents an especially challenging test case due to its hypertrophic conditions, frequent algal blooms, and complex optical properties. This study focuses on the comparative evaluation of Sentinel-2-based modeling approaches to estimate Chl-a, an indicator of eutrophication, and TSS, a measure of particulate matter dynamics, in Lake Palic. Using Sentinel-2 Level-1C (L1C) data [20] and the C2RCC processor, a well-established atmospheric correction and water quality retrieval tool for aquatic environments, bottom-of-atmosphere reflectance values were derived for predictive model development. To better represent the optical characteristics of the study site, a range of spectral indices was evaluated because such indices can enhance spectral features associated with specific water quality parameters and have been widely applied in aquatic remote sensing studies [23,26].

Although processors such as C2RCC provide operational estimates of water quality parameters, numerous studies have demonstrated that the performance of generic retrieval algorithms can vary considerably among water bodies due to differences in trophic state, constituent composition, optical complexity, and local environmental conditions. Consequently, the development and evaluation of locally calibrated models remain an active area of aquatic remote sensing research. In this context, a range of spectral indices, including both aquatic and vegetation-derived indices, was evaluated to determine their applicability under the highly eutrophic conditions of Lake Palic. Rather than assuming transferability based on their original application domain, their suitability was assessed empirically through correlation analysis and predictive model performance.

The main objective of this study is to evaluate the applicability of Sentinel-2 remote sensing for estimating Chl-a and TSS concentrations in Lake Palic under optically complex conditions. To assess both the performance of transparent, physically interpretable approaches and the potential benefits of more flexible nonlinear methods under highly eutrophic conditions, both linear regression and machine learning models were evaluated. Random Forest (RF) and k-nearest neighbor (kNN) were selected because they are non-parametric methods capable of modeling complex nonlinear relationships without strong assumptions regarding data distribution and have demonstrated robust performance in previous water quality retrieval studies [26,28]. This objective was addressed through the following specific tasks:

- Evaluation of spectral indices for Chl-a and TSS estimation;
- Development of a simple linear regression (SLR) model;
- Assessment of the feasibility of a unified model across spatial locations;
- Development of Random Forest (RF) and k-nearest neighbor (kNN) models to improve predictive accuracy;
- Model validation using statistical performance metrics (e.g., R^2 , RMSE) and comparison with in situ measurements.

2. Materials and Methods

2.1. Study Area

The importance of Lake Palic stems from its location in the immediate vicinity of the city of Subotica and the town of Palic (Figure 1). The lake is divided into four sectors extending from the Subotica Wastewater Treatment Plant (WWTP), located in Trg Lazara Nesica 9/a, 24000 Subotica, Serbia, to the town of Palic, the region's primary tourist destination. Furthermore, water exiting Lake Palic flows toward the protected Ludas Lake, emphasizing the need for continuous monitoring and effective management of water quality within the system.

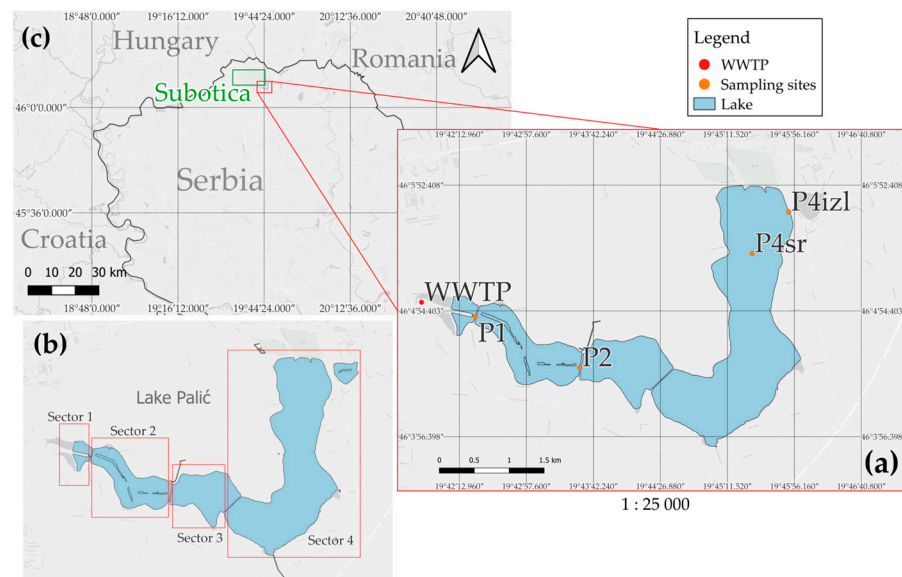


Figure 1. Disposition of Lake Palic and the sampling sites. (a) general disposition of the four sectors with the employed labels, (b) boundaries of specific sectors, (c) placement of the lake within Serbia.

To support these activities, the Public Health Institute Subotica (PHI), located in Zmaj Jovina 30, 24000 Subotica, Serbia, conducts regular water quality monitoring. The sampling network comprises one monitoring location in each of Segments 1, 2, and 3 (P1, P2, and P3) and two locations in Segment 4 (P4sr and P4izl), reflecting its larger size and environmental significance. The coordinates of the monitoring locations are P1 (46.081031° N, 19.706585° E), P2 (46.074523° N, 19.725869° E), P3 (46.073052° N, 19.741102° E), P4sr (46.089140° N, 19.757801° E), and P4izl (46.094470° N, 19.764555° E). The resulting monitoring dataset represents a valuable long-term record of water quality conditions and provides an important basis for remote-sensing-based analyses and model development [29].

The present study focuses on Segment 4, where the monitoring locations P4sr and P4izl were used for model development and validation. P4sr is located near the center of the sector and is surrounded predominantly by agricultural land. The site is characterized by relatively sheltered conditions, where water circulation is primarily influenced by wind-induced mixing and occasional sediment resuspension. In contrast, P4izl is situated at the lake outflow within the urban area of Palic and provides information on the quality of water leaving the lake. Due to their contrasting positions within the lake and differences in the surrounding land use, the two locations are exposed to distinct environmental pressures and may exhibit different water quality dynamics.

2.2. Remote Sensing, Image Processing, and Field Measurements

The Public Health Institute Subotica conducts systematic monitoring of surface water quality at multiple locations throughout the Palic-Ludas lake system [30]. Some parameters

are available monthly, while others are measured quarterly; therefore, the analysis was constrained by the availability of the quarterly data.

The research was conducted using L1C (later supplemented with Level-2A imagery for TSS) satellite images obtained from the Copernicus Sentinel-2 MSI [31]. The images were selected as the closest match to available measurements of Chl-a and TSS concentrations, spanning 2015 to 2024. Although the time interval spans over 10 years, the available data consist of 92 images at the outflow, P4izl, and 31 at the middle of the lake, P4sr. The reason for such a small sample size is the limited applicability of satellite imagery (images are lost due to poor weather conditions or heavy cloud cover) and the scarcity of field measurements, which are scarcer in the middle of the lake than at the outflow. Image processing was performed using the European Space Agency's (ESA) Sentinel Application Platform [32].

Satellite images were selected based on dates and location, downloaded, and then subset to the Lake Palic region. Sentinel-2 MSI provides spectral bands at three spatial resolutions (10 m, 20 m, and 60 m); therefore, all bands were resampled to a common 10 m spatial resolution. L1C provides Top-of-Atmosphere (TOA) radiance values that were converted to Remote Sensing Reflectance using the C2RCC processor. This processor is specifically designed for optically complex Case 2 waters, such as Lake Palic, where Chl-a, TSS, and CDOM strongly influence the optical signal [33]. Using L1C data and applying atmospheric correction via C2RCC allows the user to incorporate important environmental data for the specific location (atmospheric pressure, air temperature, turbidity, etc.), thereby yielding a more precise atmospheric correction than relying on preprocessed L2A images.

The atmospherically corrected satellite images were used to extract reflectance values from the representative pixels. To ensure pixel reliability, after locating the exact coordinates of the sampling locations, we applied a multi-pixel averaging approach using at least three surrounding control pixels [34] to the extracted reflectance values.

The gathered reflectance values were employed to calculate a set of spectral indices. Spectral indices are used because certain band ratios provide better visibility of one reflectance value over another or reduce the impact of a currently non-focal bandwidth [18,35]. The chosen indices are presented in detail in Sections 2.3.1 and 2.3.2.

Exploratory analysis of the Chl-a dataset revealed a limited number of observations exceeding 700 mg/m³. These measurements corresponded to extreme bloom conditions and exhibited spectral behavior substantially different from the remainder of the dataset, suggesting the presence of a distinct optical regime. To evaluate the influence of these observations on model development, a refined dataset was created by excluding values above this threshold. Both the full and refined datasets were subsequently analyzed, allowing assessment of model performance under typical conditions as well as evaluation of the impact of extreme bloom events on the observed spectral relationships.

The determined spectral indices were correlated with the corresponding field measurements to evaluate their ability to describe the observed Chl-a and TSS concentrations. Correlation strength was assessed using *Pearson's r*, *Spearman's rs*, and *Kendall's tau* coefficients. Initial analyses for both parameters were performed using reflectance values derived from the C2RCC processor. While satisfactory relationships were identified for Chl-a, substantially weaker correlations were observed for TSS. Consequently, an additional analysis was conducted for TSS using Sentinel-2 (European Space Agency, Paris, France.) L2A bottom-of-atmosphere reflectance products to investigate whether alternative atmospheric correction and reflectance inputs could improve the representation of suspended sediment dynamics under the optically complex conditions of Lake Palic. Since the Chl-a analysis produced satisfactory results using C2RCC-derived reflectance, this additional evaluation was limited to TSS.

Based on the measured concentrations and the calculated spectral indices, predictive models for Chl-a and TSS were developed using two alternative approaches. The first relied on simple linear regression (SLR), which provides transparent and easily interpretable relationships between environmental indices and water quality parameters. The second employed machine learning techniques to evaluate whether nonlinear approaches could improve predictive performance. Two machine learning algorithms were tested: Random Forest (RF) and k-Nearest Neighbors (kNN). All machine learning models were developed using the complete set of evaluated environmental indices and compared against the corresponding linear regression models.

2.3. Environmental Indices

The spectral indices used for Chl-a and TSS estimation were investigated separately due to differences in their spectral signatures. For the evaluation of Chl-a, we considered fifteen spectral indices (Table 1). The description of these indices and the reasoning behind their selection is given below.

Table 1. Environmental indices used for the Chl-a estimation models.

Name	Expression	Employed Bands (nm)	Reference
RGI	R_{690}/R_{550}	$R_{690} = R_{705}$ $R_{550} = R_{560}$	[35]
BRI	R_{430}/R_{680}	$R_{430} = R_{490}$ $R_{680} = R_{705}$	[36]
MI1	R_{700}/R_{670}	$R_{700} = R_{705}$ $R_{670} = R_{665}$	[37]
RARSa	R_{675}/R_{700}	$R_{675} = R_{665}$ $R_{700} = R_{705}$	[38]
Crt2	R_{695}/R_{670}	$R_{695} = R_{705}$ $R_{670} = R_{665}$	[39]
ZTM	R_{750}/R_{710}	$R_{750} = R_{740}$ $R_{710} = R_{705}$	[40]
VARIGreen	$(R_{550} - R_{665})/(R_{550} + R_{665})$	$R_{550} = R_{560}$ $R_{665} = R_{665}$	[41]
TCARI	$3 \times [(R_{700} - R_{670}) - 0.2 \times (R_{700} - R_{550})] \times (R_{700}/R_{670})$	$R_{550} = R_{560}$ $R_{670} = R_{665}$ $R_{700} = R_{705}$	[42]
PSSRc	R_{800}/R_{470}	$R_{800} = R_{740}$ $R_{470} = R_{443}$	[43]
SR2	R_{750}/R_{705}	$R_{750} = R_{740}$ $R_{705} = R_{705}$	[44]
NDI	$(R_{750} - R_{705})/(R_{750} + R_{705})$	$R_{750} = R_{740}$ $R_{705} = R_{705}$	[45]
OSAVI	$(1 + 0.16) \cdot (R_{801} - R_{670})/(R_{801} + R_{670} + 0.16)$	$R_{801} = R_{740}$ $R_{670} = R_{665}$	[46]
MSAVI	$0.5 \times [2 \times R_{800} + 1 - \sqrt{(2 \times R_{800} + 1)^2 - 8 \times (R_{800} - R_{670})}]$	$R_{800} = R_{740}$ $R_{670} = R_{665}$	[46]
Datt99	$(R_{850} - R_{710})/(R_{850} - R_{680})$	$R_{850} = R_{740}$ $R_{710} = R_{705}$ $R_{680} = R_{665}$	[47]
NDVI	$(R_{800} - R_{670})/(R_{800} + R_{670})$	$R_{800} = R_{740}$ $R_{670} = R_{665}$	[48]

2.3.1. Environmental Indices Used for Chl-a Analysis

The Red-Green Index (RGI) is commonly applied for Chl-a monitoring and is defined as a simple reflectance ratio between the red and green bands. It was first introduced by Gitelson et al. [35]. Chlorophyll a produces moderate reflectance in the green band while strongly absorbing light in the red wavelengths. However, in turbid or highly productive waters, suspended matter may increase red-band backscattering while reducing green reflectance, thereby elevating RGI values. Analogous to the RGI, the Blue Ratio Index (BRI) compares reflectance in the blue and red-edge spectral regions. Peñuelas et al. [36] developed several semi-empirical indices, including BRI, to estimate the carotenoid-to-Chl-a ratio in higher plants, demonstrating their effectiveness for monitoring pigment dynamics and physiological stress.

The MI1 index was introduced by Kim et al. [37] using hyperspectral reflectance measurements in the red and red-edge spectral regions to quantify absorbed photosynthetically active radiation (APAR). By evaluating specific reflectance ratios, MI1 was proposed as a sensitive metric for estimating APAR and, consequently, Chl-a concentration, which motivated its inclusion in this study. The Ratio Analysis of Reflectance Spectra (RARSa) was introduced by Chapelle et al. [38] for the simultaneous quantification of Chl-a, chlorophyll-b, and carotenoids. Their work demonstrated that carefully selected reflectance ratios can reliably discriminate between individual pigment concentrations. Carter [39] proposed the Crt2 index for early stress detection in vegetation. By analyzing reflectance at 670 nm (red) and 695 nm (red-edge), both sensitive to early reductions in Chl-a, Carter demonstrated that Crt2 effectively captures stress-induced pigment decline. Zarco-Tejada et al. [40] developed the ZTM index to decouple pigment signals from canopy scattering effects. By combining a red-edge band sensitive to Chl-a variation (710 nm) with a near-infrared band (750 nm), ZTM enhances sensitivity while reducing structural interference.

The VARI index was introduced by Gitelson et al. [41] to estimate vegetation fraction using only visible bands. By combining green (550 nm) and red (665 nm) reflectance, VARI minimizes atmospheric and soil background effects, enabling reliable Chl-a estimation even without near-infrared data. Haboudane et al. [42] proposed the Transformed Chlorophyll Absorption in Reflectance Index (TCARI) to enhance Chl-a sensitivity while suppressing background noise. TCARI integrates differences in green, red, and red-edge reflectance to compensate for heterogeneous backgrounds, thereby improving prediction accuracy. Jordan [43] established the Simple Ratio (SR), later adapted into the PSSRc index for carotenoid estimation. By combining near-infrared (800 nm) and blue (470 nm) bands, PSSRc captures pigment-related reflectance changes. The Simple Ratio 2 (SR2) index was developed by Gitelson et al. [44] for Chl-a estimation in dense vegetation. It exploits the contrast between the highly sensitive red-edge band (705 nm) and the near-infrared band (750 nm), where chlorophyll absorption is minimal. McNairn and Protz [45] introduced the Normalized Difference Index (NDI), which constrains values between -1 and 1 , facilitating comparison across sites. Using red-edge bands (705 and 750 nm), NDI effectively tracks pigment variations. Values closer to 1 indicate higher Chl-a concentrations, values closer to 0 indicate lower Chl-a content, and non-vegetated surfaces yield values near -1 . Daughtry et al. [46] applied the OSAVI and MSAVI indices to mitigate soil background interference. OSAVI stabilizes reflectance in sparsely vegetated areas, while MSAVI includes a self-adjusting factor to further reduce soil effects. By focusing on the contrast between the 800 nm and 670 nm bands, these indices reflect the actual chlorophyll content instead of the background color. Datt [47] introduced the Datt99 index, which incorporates near-infrared (850 nm), red-edge (710 nm), and red (680 nm) absorption bands to estimate chlorophyll content while minimizing the effects of leaf structure. Instead of a simple ratio, this formula uses three specific points: 850 nm, where plants reflect light based on their structure;

710 nm, the sensitive red edge; and 680 nm, the point of maximum Chl-a absorption. We evaluated the behavior of individual spectral bands and found no evidence of reflectance clipping, indicating it is not a standard sensor saturation issue. Instead, we noticed peaks occurring simultaneously across bands, suggesting extreme optical events. The identified pattern is consistent with nonlinear spectral responses under high-concentration or optically complex conditions.

Finally, Rouse et al. [48] proposed NDVI as a general indicator of vegetation health. NDVI contrasts red absorption and near-infrared reflectance and, in aquatic environments, serves as a proxy for phytoplankton biomass and Chl-a concentration. Healthy leaves absorb red light (670 nm) for photosynthesis, while reflecting infrared light (800 nm) like a mirror. Quantifying the contrast between the maximum absorption of pigments and the light scattering from particles in the water enables the differentiation between clear water and areas with high phytoplankton density or active blooms.

After adapting the published index formulations to the spectral bands available in Sentinel-2 MSI imagery, MI1 and SR2 became mathematically identical to Crt2 and ZTM, respectively, and were therefore excluded from further analysis.

2.3.2. Environmental Indices Used for TSS Analysis

In developing the TSS estimation model, we considered seven spectral indices presented in Table 2.

Table 2. Environmental indices used for the TSS estimation models.

Name	Expression	Employed Bands (nm)	Reference
NDWI	$(R_{NIR} - R_R) / (R_{NIR} + R_R)$	$R_{NIR} = R_{842}$ $R_R = R_{665}$	[49]
NDSSI	$(R_B - R_{NIR}) / (R_B + R_{NIR})$	$R_B = R_{490}$ $R_{NIR} = R_{842}$	[50]
NDTI	$(R_R - R_G) / (R_R + R_G)$	$R_R = R_{665}$ $R_G = R_{560}$	[49]
NSMI	$(R_R + R_G - R_B) / (R_R + R_G + R_B)$	$R_R = R_{665}$ $R_G = R_{560}$ $R_B = R_{490}$	[51]
SWRC	R_R / R_B	$R_R = R_{665}$ $R_B = R_{490}$	[49]
SR	R_R / R_{NIR}	$R_R = R_{665}$ $R_{NIR} = R_{842}$	[49]
R_{705} / R_{665}	R_{705} / R_{665}	$R_{705} = R_{705}$ $R_{665} = R_{665}$	[52]
$(R_{665}^{-1} - R_{705}^{-1})R_{740}$	$(R_{665}^{-1} - R_{705}^{-1})R_{740}$	$R_{705} = R_{705}$ $R_{665} = R_{665}$ $R_{740} = R_{740}$	[53]
hybrid	$R_{B8A} / R_{B4} - 0.5(R_{B5} / R_{B4})$	$R_{B8A} = R_{865}$ $R_{B4} = R_{665}$ $R_{B5} = R_{705}$	

In the study by Mo et al. [49], four water quality indices—the Normalized Difference Water Index (NDWI), the Normalized Difference Turbidity Index (NDTI), the Simple Ratio (SR), and the Simple Ratio Water Color (SWRC)—were derived from Sentinel-2 spectral bands and included among the candidate predictor variables for TSS inversion. These indices were combined with single-band reflectance values to train six machine learning

algorithms. The authors demonstrated that models incorporating both single bands and spectral indices yielded the highest accuracy. NDWI and SR-type ratios are designed to capture contrasts in visible and near-infrared reflectance associated with scattering and water color, while NDTI exploits the relative increase in red (B4) reflectance relative to green (B3) as turbidity increases. This behavior proved effective in the Pinglu Canal estuary and contributed significantly to model performance. Consequently, the same set of four spectral indices (NDWI, NDTI, SR, and SWRC) was adopted in the present study, as they have been shown to complement single-band Sentinel-2 data and improve the retrieval of suspended particulate matter in optically complex aquatic environments.

In addition to these indices, two further TSS-related indices—the Normalized Difference Suspended Sediment Index (NDSSI) and the Normalized Suspended Matter Index (NSMI)—were included based on their independent development and successful application in previous remote sensing studies of suspended sediments [50,51]. NDSSI was originally proposed to enhance the contrast between sediment-laden and clearer waters by exploiting the normalized difference between visible and near-infrared reflectance, thereby emphasizing increases in backscattering associated with higher sediment concentrations. NSMI follows a similar conceptual approach but employs an alternative band combination to enhance sensitivity to suspended matter under varying optical conditions. In this study, NDSSI and NSMI were incorporated alongside the other indices to evaluate whether their distinct spectral formulations could more effectively isolate the suspended sediment signal in optically complex, phytoplankton-dominated lake environments.

Another widely used proxy for Case 2 water monitoring is the R_{705}/R_{665} ratio [52], which isolates the red-edge peak relative to red absorption. Additionally, the index $(R_{665}^{-1} - R_{705}^{-1})R_{740}$ [53] was included to reduce the influence of non-algal particles, such as TSS and CDOM, on Chl-a estimation. By incorporating the red band (665 nm), the index captures the peak of Chl-a absorption, while the red-edge band (705 nm) reflects Chl-a variability. The near-infrared band (740 nm) helps reduce the influence of TSS-related scattering.

Because initial correlation screenings between spectral indices and in situ TSS measurements produced unreliable results, an additional custom red-band-normalized index, referred to as the “hybrid” index, was developed. The hybrid index is defined as $(B8A/B4) - 0.5(B5/B4)$ and was designed as an exploratory attempt to improve TSS representation under the extreme optical conditions of Lake Palic. The first ratio, $B8A/B4$, enhances sensitivity to suspended particles, as near-infrared reflectance generally increases with turbidity due to scattering effects. In contrast, the ratio $B5/B4$ captures the red-edge response commonly associated with elevated phytoplankton biomass. The subtractive term was therefore introduced to partially reduce the influence of chlorophyll-related optical effects that may obscure the TSS signal under highly eutrophic conditions. Given the extremely high Chl-a concentrations observed in Lake Palic (exceeding 1250 mg/m^3), it was hypothesized that phytoplankton-related reflectance could partially mask TSS-related variability. The weighting coefficient (0.5) was selected empirically and should be viewed as an exploratory correction factor rather than a universally applicable parameter.

2.3.3. Evaluation of the Considered Indices

Correlation analyses were performed for the two monitoring locations to quantify linear (*Pearson's*, r) and monotonic (*Spearman's*, r_s , and *Kendall's* τ) relationships between the target variables, Chl-a and TSS, and the Sentinel-2-derived spectral indices. The results are presented in Table 3. In the correlation matrices, red indicates negative correlations, and blue indicates positive correlations. The color intensity represents the strength of the correlation; darker shades suggest stronger relationships (closer to -1 or $+1$).

Table 3. Chl-a correlation analysis results for the two observed locations.

Location P4izl						
	Full Dataset N = 92			Refined Dataset N = 77		
RGI	−0.665	−0.731	−0.525	−0.639	−0.623	−0.438
BRI	0.717	0.756	0.546	−0.789	−0.761	−0.566
RARSa	0.803	0.824	0.625	0.810	0.761	0.566
Crt2	−0.777	−0.824	−0.625	−0.789	−0.761	−0.566
ZTM	−0.766	−0.773	−0.568	−0.750	−0.683	−0.496
VARIGreen	0.724	0.710	0.506	0.692	0.592	0.413
PSSRc	−0.671	−0.773	−0.563	0.501	0.789	0.588
TCARI	0.467	0.830	0.622	−0.677	−0.695	−0.491
NDI	−0.772	−0.773	−0.568	−0.755	−0.683	−0.496
OSAVI	0.548	0.844	0.645	0.588	0.808	0.605
Datt99	−0.753	−0.822	−0.619	−0.768	−0.762	−0.563
MSAVI	0.513	0.845	0.645	0.555	0.808	0.608
NDVI	−0.777	−0.812	−0.609	−0.780	−0.742	−0.546
	<i>r</i> *	<i>r_S</i> **	<i>τ</i> ***	<i>r</i> *	<i>r_S</i> **	<i>τ</i> ***
Location P4sr						
	Full Dataset N = 31			Refined Dataset N = 27		
RGI	−0.571	−0.671	−0.490	−0.532	−0.514	−0.356
BRI	−0.648	−0.753	−0.591	−0.261	−0.548	−0.379
RARSa	0.647	0.753	0.591	0.607	0.661	0.499
Crt2	−0.648	−0.753	−0.591	−0.650	−0.661	−0.499
ZTM	−0.666	−0.773	−0.586	−0.631	−0.644	−0.453
VARIGreen	0.579	0.636	0.462	0.468	0.464	0.322
PSSRc	0.555	0.742	0.591	−0.630	−0.601	−0.419
TCARI	−0.596	−0.688	−0.517	0.686	−0.670	−0.521
NDI	−0.668	−0.773	−0.586	−0.625	−0.644	−0.453
OSAVI	0.590	0.725	0.591	0.730	0.692	0.550
Datt99	−0.642	−0.734	−0.559	−0.667	−0.658	−0.487
MSAVI	0.564	0.749	0.617	0.714	0.692	0.556
NDVI	−0.658	−0.782	−0.614	−0.649	−0.652	−0.487
	<i>r</i> *	<i>r_S</i> **	<i>τ</i> ***	<i>r</i> *	<i>r_S</i> **	<i>τ</i> ***

*—Pearson's *r*; **—Spearman's *r_S*; ***—Kendall's *τ*.

2.4. Model Development and Evaluation

2.4.1. Model Development

The implemented methodology for Chl-a and TSS estimation is as follows:

- correlation screening;
- preparing the reference model—using simple linear regression (SLR);
- preparing the machine learning models—nonlinear ensemble modeling using Random Forest (RF) and distance-based modeling using k-Nearest Neighbors (kNN);

Correlation screening is essential for assessing the strength of the relationship between the selected spectral indices and the measured data, guiding the forthcoming model development.

Parameter estimation using simple linear regression is considered the baseline approach that provides a general guide for model interpretation. Although it is relatively rigid, a limiting factor for modeling complex environments such as high-turbidity waters, its interpretability makes it an important part of these studies. Given the limitations of linear regression models, specifically the inability to model nonlinear behavior caused by the fluctuating conditions in the environment, machine learning algorithms were added as a way to evaluate whether they could better model the measured data. With that in mind, we selected two machine learning approaches considered standard methods in remote sensing—Random Forest [54] and k-Nearest Neighbors [55]. RF was introduced by Breiman [54]. It is an ensemble-based learning method that improves prediction accuracy and robustness by aggregating the outputs of multiple independent decision trees. Each tree is trained on a bootstrap sample of the original dataset (bagging), and at each node, only a randomly selected subset of predictors is considered. For regression tasks, the final predicted Chl-a value is the average across all trees. Although it is very well suited for handling complex datasets, such as remote sensing data, it is limited because it cannot predict outside its training range.

In contrast, kNN is a non-parametric, instance-based algorithm. Fix and Hodges presented a mathematical analysis of this model in their seminal work [56]. Unlike RF, this approach assumes that observations with similar feature characteristics are close to one another in the multidimensional feature space. For an unknown sample, kNN computes the distance to all training samples, and the predicted Chl-a concentration is derived as the mean of the values associated with the k nearest neighbors. The principal advantage of kNN lies in its conceptual simplicity and its ability to operate without assumptions regarding the underlying data distribution. However, it is sensitive to outliers and can produce unreliable predictions if these are not properly handled.

The joint application of RF and kNN, therefore, allows a direct comparison between a high-performance ensemble regression approach and a simpler, distance-based method, providing a comprehensive assessment of Chl-a estimation for the given dataset.

2.4.2. Model Evaluation

The presented research employs two fundamentally different modeling approaches: simple linear regression (SLR), a parametric statistical method, and Random Forest (RF) and k-nearest neighbors (kNN), which are non-parametric machine learning (ML) models. As a result, the evaluation framework must account for differences in model assumptions. While both approaches are assessed using common performance metrics, SLR additionally requires diagnostic tests to verify underlying statistical assumptions.

The SLR models were evaluated using the coefficient of determination (R^2), root mean squared error (RMSE) [57], model significance (p -value), Shapiro–Wilk test for residual normality, and the Durbin–Watson statistic for autocorrelation. The coefficient of determination (R^2) quantifies the proportion of variance in the observed data explained by the model, whereas RMSE measures the average magnitude of prediction error in the original units of the variable, allowing for direct physical interpretation [57].

The performance of the machine learning model was evaluated using multiple error metrics. Mean Squared Error (MSE) quantifies the average squared prediction error, placing greater emphasis on large deviations and thus being sensitive to outliers [57]. Mean Absolute Error (MAE) represents the average absolute deviation between observed and predicted values and is less sensitive to extreme values, providing a measure of typical

model error [57]. Given the wide range of observed values in the studied lake, the Mean Absolute Percentage Error (MAPE) was used to normalize errors relative to the magnitude of observations, thereby enabling more meaningful comparisons across varying conditions [58]. Additionally, the coefficient of variation in the RMSE (CVRMSE) was employed to facilitate comparisons between locations (e.g., the middle of the lake versus the out-flow) and between different water quality parameters [59]. The employed equations are the following:

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2, \text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}, \text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|,$$

$$\text{MAPE} = \frac{1}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right| \cdot 100\%, R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}, \text{CVRMSE} = \frac{\text{RMSE}}{\bar{y}} \cdot 100\%,$$

where y_i represents the observed in situ Chl-a or TSS concentration, $\hat{y}_i$ is the corresponding model prediction for the i -th observation, $\bar{y}$ is the mean of the observed values, and n is the number of samples.

Model selection was based primarily on maximizing R^2 and minimizing RMSE, while the remaining metrics were used to evaluate model assumptions and provide complementary information regarding prediction error characteristics.

To enable comparison between parametric and non-parametric approaches, overlapping performance metrics—specifically R^2 and RMSE—were used as common evaluation criteria across all models.

3. Results

3.1. Evaluation of Environmental Indices

3.1.1. Correlation Analysis for Chl-a

Exploratory analysis of the Chl-a dataset revealed a limited number of observations exceeding 700 mg/m³. These measurements corresponded to extreme bloom conditions and exerted a disproportionate influence on the observed spectral relationships and subsequent model development. Similar observations were reported by Ibrahim et al. [26], who found that the exclusion of locations characterized by exceptionally high Chl-a concentrations substantially improved model performance, highlighting the influence of extreme bloom conditions on retrieval-model reliability.

To evaluate the impact of such events on model performance, all analyses were conducted using both the full dataset (including all observations; e.g., $N = 92$ for P4izl) and a refined dataset (e.g., $N = 77$ for P4izl), in which observations exceeding 700 mg/m³ were excluded. This approach enabled assessment of the influence of extreme bloom events and associated nonlinear spectral responses on the derived relationships. The correlation results between the evaluated spectral indices and measured Chl-a concentrations are presented in Table 3.

The results show that several indices had a reversal in correlation direction following the removal of extreme Chl-a values. For example, index BRI at P4izl went from a strong positive to a strong negative correlation (at P4izl: $r_{92} = 0.717$ to $r_{77} = -0.789$, $rs_{92} = 0.756$ to $rs_{77} = -0.761$, $\tau_{92} = 0.546$ to $\tau_{77} = -0.566$). The magnitude of this shift indicates that it is not a minor variation but a complete reversal in the direction of the correlation. This change is primarily due to the removal of extreme Chl-a values, which exerted a disproportionate influence on the correlation structure. These values likely introduced nonlinear effects, leading to the observed reversal.

In contrast, the RARSa index maintained a consistently strong and stable monotonic association, with only a minor decrease in *Kendall's tau* from 0.625 (full dataset) to 0.566

(refined dataset). Similarly, BRI reversed from a positive correlation ($r_{92} = 0.717$) to a strong negative correlation ($r_{77} = -0.789$). The selective correlation reversal demonstrates the strong influence of extreme eutrophic conditions on indices that rely on specific spectral bands (e.g., blue at 490 nm and red-edge at 705 nm in BRI). Under highly nonlinear optical conditions, such indices may lose monotonicity and yield misleading correlations. The observed reversals in the refined dataset indicate that data filtering reduced the influence of extreme observations, thereby revealing more stable spectral-biophysical relationships.

The RARSa index, based on the red and red-edge band ratio (665 and 705 nm), showed robust performance across both datasets ($r_{92} = 0.803$ for the full dataset and $r_{77} = 0.810$ for the refined dataset). This stability is likely related to the enhanced robustness of red-edge features under optically complex and high-concentration conditions [18], supporting the selection of RARSa as the most reliable linear predictor of Chl-a.

At location P4sr, the correlation analysis was constrained by a relatively small initial sample size ($N = 31$), which was further reduced to $N = 27$ after excluding extreme Chl-a values, inherently lowering coefficient stability. The refined results reveal a location-specific response to data filtering. In contrast to the stronger, more systematic reversals observed at P4izl, the P4sr site exhibited mixed behavior, with some indices showing moderate strengthening of existing correlations, while others underwent complete sign reversals.

For example, OSAVI improved from $r_{31} = 0.590$ to $r_{27} = 0.730$, and MSAVI from $r_{31} = 0.564$ to $r_{27} = 0.714$, indicating that removing extreme observations enhanced the underlying spectral relationships. The RARSa index also remained relatively stable, with only a minor reduction from $r_{31} = 0.647$ to $r_{27} = 0.607$. However, several indices exhibited substantial structural changes following refinement. Most notably, PSSRc reversed from a positive to a negative correlation ($r_{31} = 0.555$ to $r_{27} = -0.630$), while TCARI changed from a negative to a positive Pearson correlation ($r_{31} = -0.596$ to $r_{27} = 0.686$). These shifts indicate that certain indices are highly sensitive to extreme eutrophic conditions and may respond nonlinearly under varying optical regimes.

Overall, the results suggest that removing extreme observations improved the interpretability of several spectral relationships while simultaneously revealing the instability of indices strongly affected by optical complexity and changing spectral dominance.

3.1.2. Correlation Analysis for TSS

Table 4 shows the correlation coefficients between measured TSS concentrations and the selected indices. In the case of TSS, there were no extreme outliers; therefore, there was no need to examine how data refinement would affect the overall procedure. The attained results display two important patterns. First, correlation values are notably lower than those observed for Chl-a estimation. Second, there is a pronounced difference between the correlations obtained at the middle of the lake and those obtained at the outflow location. These findings indicate that TSS retrieval performance is strongly location-dependent and suggest that the two sampling locations are characterized by distinct optical conditions, which may influence the observed relationships between spectral indices and measured TSS concentrations.

Overall, the results reveal weak associations, with the strongest relationships observed for NDTI ($r = -0.389$; $r_s = -0.427$) and R705/R665 ($r = -0.380$; $r_s = -0.434$) at P4izl.

At the P4sr location, correlations are even weaker, with most indices approaching zero (e.g., $r = -0.150$; $r_s = -0.146$ for R705/R665). This suggests that the spectral signal associated with suspended sediment was likely masked by the dominant optical influence of Chl-a and other optically active constituents. The limited performance of standard TSS indices—primarily based on visible and conventional near-infrared bands (B2–B5)—is

therefore likely related to strong optical interference caused by high phytoplankton biomass and overall optical complexity.

Table 4. TSS correlation analysis results for the two observed locations L1C.

	Location P4izl			Location P4sr		
NDSSI	0.347	0.410	0.280	0.084	0.097	0.083
NDTI	−0.389	−0.427	−0.311	−0.084	−0.002	−0.009
NSMI	−0.376	−0.387	−0.263	−0.069	−0.066	−0.041
SWRC	−0.340	−0.415	−0.289	−0.143	−0.075	−0.078
R_{705}/R_{665}	−0.380	−0.434	−0.276	−0.150	−0.146	−0.110
$(R_{665}^{-1} - R_{705}^{-1})R_{740}$	−0.332	−0.362	−0.229	−0.112	−0.114	−0.087
	r *	rs **	τ ***	r *	rs **	τ ***

*—Pearson's r ; **—Spearman's rs ; ***—Kendall's τ .

This limitation motivated the use of Level-2A bottom-of-atmosphere (BOA) reflectance data to exploit additional near-infrared bands (B7 and B8A) and to develop a custom hybrid index to isolate the subtle sediment signal from overwhelming spectral interference.

As shown in Table 5, the correlation analysis using L2A BOA reflectance produced only limited changes in correlation strength compared with the corresponding L1C-based analysis. Although marginal increases were observed for certain indices at P4izl (e.g., NDWI and NDSSI), correlation strengths remained below acceptable thresholds for empirical modeling. The hybrid index produced only modest responses ($r = -0.245$ at P4izl and $r = 0.100$ at P4sr), confirming the persistent difficulty in retrieving TSS under extreme eutrophic and optically complex conditions. Further analysis of the two red-red edge-NIR indices, R_{705}/R_{665} and $(R_{665}^{-1} - R_{705}^{-1})R_{740}$, demonstrates a notable reduction in Pearson's r at the middle of the lake location. These indices are sensitive to biological activity and scattering-related optical variability, particularly in optically complex waters. Because these indices respond strongly to scattering processes, the observed reduction in correlation at P4sr likely reflects the transition from the more dynamic outflow zone—characterized by mixing, resuspension, and phytoplankton accumulation—to the comparatively stable, diluted conditions in the middle of the lake.

Table 5. TSS correlation analysis results for the two observed locations L2A.

	Location P4izl			Location P4sr		
NDWI	0.366	0.438	0.304	0.296	0.133	0.123
NDSSI	0.293	0.451	0.296	0.156	0.076	0.054
NDTI	−0.292	−0.349	−0.190	0.046	−0.131	−0.123
NSMI	−0.052	0.017	0.020	−0.262	−0.105	−0.094
SWRC	−0.061	−0.044	−0.045	−0.145	−0.270	−0.163
R_{705}/R_{665}	0.385	0.494	0.344	0.049	0.038	0.031
$(R_{665}^{-1} - R_{705}^{-1})R_{740}$	0.395	0.486	0.344	0.132	0.001	0.014
SR	−0.344	−0.438	−0.304	−0.293	−0.133	−0.123
hybrid	−0.245	−0.378	−0.267	0.100	−0.027	−0.009
	r *	rs **	τ ***	r *	rs **	τ ***

*—Pearson's r ; **—Spearman's rs ; ***—Kendall's τ .

3.2. Evaluation of the Developed Models

3.2.1. Chlorophyll-a at Location P4izl

Regardless of the calculated correlations, all indices were used to develop SLR models. Table 6 shows the metrics for those that resulted in $R^2 > 0.5$. The highest predictive capability was observed with RARSa, with the highest determination coefficient ($R^2 = 0.645$) and the lowest RMSE (174.99). The two weakest-performing models were VARIGreen ($R^2 = 0.525$) and BRI ($R^2 = 0.514$), both with high RMSE values (>200).

Table 6. Error analysis for SLR models developed using unrefined data at P4izl.

N = 92		Model Performance		Statistical Signif.	Normality Shapiro–Wilk		Autocorrel. Durbin–Watson	
Index	Model Chl – a	R^2	RMSE	Slope p -Value	W	p -Value	DW	p -Value
RARSa	$-664.6 + 795 \cdot \text{RARSa}$	0.645	174.99	<0.0001	0.95	0.0029	1.89	0.542
NDVI	$-766.5 - 1918 \cdot \text{NDVI}$	0.604	184.68	<0.0001	0.96	0.0066	1.85	0.408
Datt99	$1238 - 1143 \cdot \text{Datt99}$	0.567	193.24	<0.0001	0.95	0.0011	1.78	0.292
NDI	$-1465 - 3518 \cdot \text{NDI}$	0.595	186.74	<0.0001	0.96	0.0137	1.79	0.258
VARIGreen	$193.2 + 600.8 \cdot \text{VARIGreen}$	0.525	202.38	<0.0001	0.96	0.0090	1.78	0.240
ZTM	$1672 - 4112 \cdot \text{ZTM}$	0.586	188.78	<0.0001	0.96	0.0131	1.78	0.262
Crt2	$1469 - 1366 \cdot \text{Crt2}$	0.604	184.76	<0.0001	0.95	0.0029	1.83	0.382
BRI	$282.6 + 5.631 \cdot \text{BRI}$	0.514	204.65	<0.0001	0.89	<0.0001	1.91	0.622

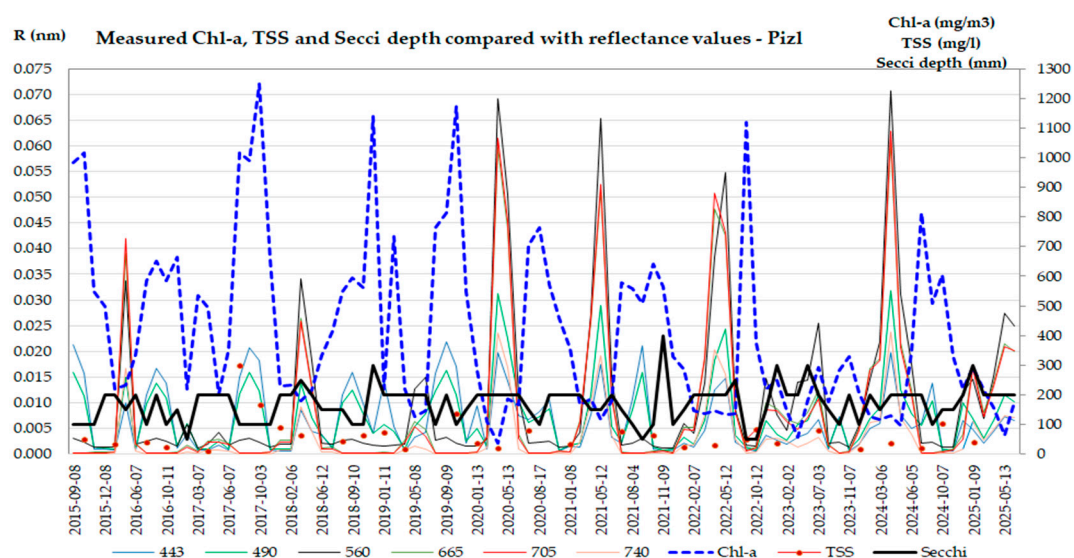
All models in Table 6 show significant regression slopes ($p < 0.0001$), confirming a meaningful relationship between the considered indices and Chl-a concentrations. The main limitation across all models is the lack of residual normality, as indicated by Shapiro–Wilk p -values below 0.05, suggesting that the observed Chl-a variability is not fully captured by the developed models.

After excluding extreme Chl-a values ($>700 \text{ mg/m}^3$), Table 7, all models satisfied the normality assumption, indicating that the previously observed deviations were primarily driven by high outliers. Among the evaluated models, RARSa demonstrated the best performance, with a noticeable improvement in R^2 (from 0.645 to 0.655) and a substantial reduction in RMSE (from 174.99 to 110.13). Similar reductions in RMSE were observed across all models, suggesting that prediction errors were largely dominated by extreme values. Although R^2 values showed only modest improvement, this indicates that the underlying relationships between the indices and Chl-a were already well established and that outliers predominantly affected model error rather than correlation strength. The most significant improvement was observed in the residual normality, which is now satisfied for all models, thereby confirming the statistical validity of the regression analyses. Model significance remained consistently high ($p < 0.0001$ for all models), confirming strong relationships between the selected spectral indices and Chl-a concentrations. Furthermore, Durbin–Watson statistics (values close to 2) indicate the absence of autocorrelation in the residuals.

Figure 2 presents the relationship between measured chlorophyll a (Chl-a), total suspended solids (TSS), and the corresponding spectral reflectance values. The primary objective of this analysis was to investigate the cause of the observed reversal in the correlation.

Table 7. Error analysis for SLR models developed using refined data at P4izl.

N = 77		Model Performance		Statistical Signif.	Normality Shapiro–Wilk		Autocorrel. Durbin–Watson	
Index	Model Chl – a	R ²	RMSE	Slope p-Value	W	p-Value	DW	p-Value
RARSa	$-399.5 + 553 \cdot \text{RARSa}$	0.655	110.13	<0.0001	0.98	0.4180	2.23	0.356
NDVI	$-452.2 - 1299 \cdot \text{NDVI}$	0.608	117.38	<0.0001	0.98	0.2759	2.22	0.394
Datt99	$897.9 - 768 \cdot \text{Datt99}$	0.589	120.22	<0.0001	0.98	0.3086	2.24	0.354
NDI	$-915.5 - 2364 \cdot \text{NDI}$	0.570	122.95	<0.0001	0.98	0.2547	1.98	0.826
ZTM	$1186 - 2745 \cdot \text{ZTM}$	0.563	124.07	<0.0001	0.98	0.3155	1.99	0.878
Crt2	$1065 - 929.8 \cdot \text{Crt2}$	0.623	115.24	<0.0001	0.98	0.1964	2.27	0.282

**Figure 2.** Measurement—reflectance plot at P4izl.

The visual assessment of the figure indicates that the phenomenon is unlikely to be caused by simple pixel saturation, as no clear evidence of signal clipping (i.e., plateauing or horizontal flattening of reflectance values) is observed. Instead, reflectance values exhibit sharp and pronounced peaks, suggesting dynamic optical variability rather than simple signal clipping. Furthermore, reflectance peaks do not systematically coincide with peaks in measured Chl-a concentrations, suggesting that Chl-a alone does not explain the observed optical variability at the P4izl location. Analysis of TSS shows that several major reflectance peaks align with elevated TSS values, particularly during high-intensity events. However, this relationship is not consistent across the entire dataset, indicating that TSS contributes to several high-reflectance events but does not consistently explain the observed variability. The inclusion of Secchi depth provides clearer insight, revealing a consistent inverse relationship between water transparency and reflectance. Periods of reduced Secchi depth (i.e., lower water transparency) correspond closely to increased reflectance across multiple spectral bands. This pattern suggests that scattering processes associated with suspended particulate matter and overall turbidity play a major role in shaping the observed spectral response. Overall, the observed patterns suggest that the optical properties of the lake at P4izl are influenced by a complex interaction of multiple constituents, including suspended particles, phytoplankton (Chl-a), and likely colored dissolved organic matter (CDOM). While turbidity-related processes appear to be an important contributor to the observed

variability, the available data indicate that no single constituent fully explains the spectral behavior of the lake.

The time series in Figure 3 compares the SLR model's predictions based on RARSa and NDVI indices with Chl-a measurements, using the refined dataset ($N = 77$). The SLR models successfully capture the primary temporal rhythm and magnitude of the Chl-a peaks. This visual consistency validates the quantitative results of SLR models. However, the chart also highlights the inherent limitation of linear models attempting to fit a complex, nonlinear environmental signal. While the SLR predictions closely follow the trend, they still exhibit noticeable deviations during peak bloom periods, such as late 2017 and the summers of 2020 and 2021, where the models underestimate the observed concentration peaks. On the other hand, in late 2023 and 2024, the SLR models appear to overestimate the bloom's magnitude relative to the measured data. This subtle variability confirms that while data refinement significantly enhanced the performance of the SLR model, improving its precision considerably, the underlying relationship between a single spectral index and Chl-a is not perfectly linear, justifying the continued investigation of nonlinear methods like Random Forest.

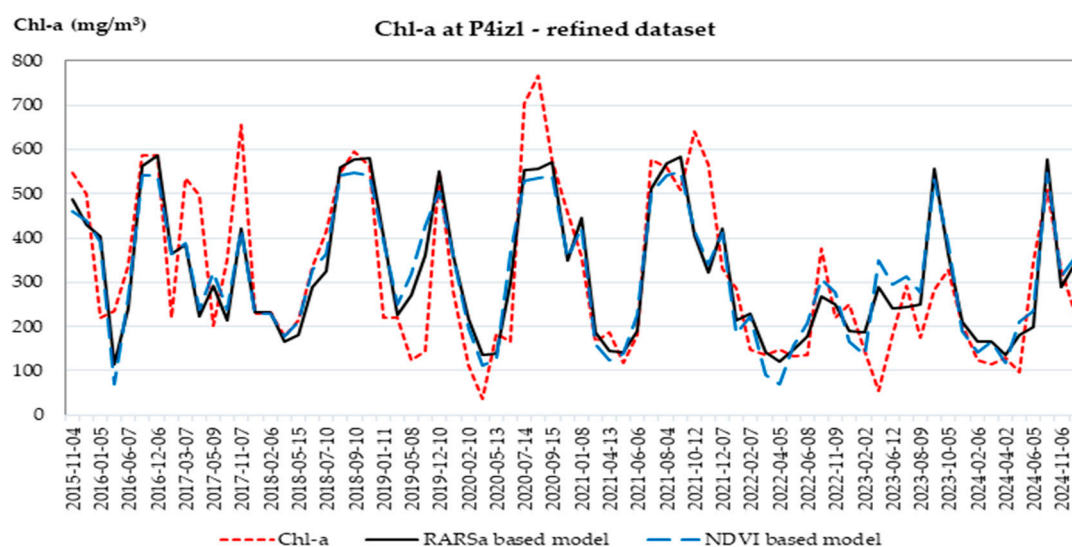


Figure 3. Chl-a measurements compared with SLR predictions using the refined dataset.

The Random Forest model, Figure 4, demonstrated superior generalization performance compared to k-Nearest Neighbors when applied to the optically complex full dataset at sampling location P4izl (Table 8). The RF model achieved a coefficient of determination of $R^2 = 0.701$, whereas kNN explained only 60.5% of the variance ($R^2 = 0.605$). Despite this relatively high explanatory power, the absolute prediction errors were substantial. The RF model yielded an RMSE of 158.693 mg/m^3 , while the kNN model performed considerably worse with an RMSE of 182.419 mg/m^3 . A Mean Squared Error (MSE)-based comparison clearly favored RF over kNN, confirming RF as the preferred nonlinear approach for the full dataset.

Table 8. Error analysis results using the full dataset.

	Model	MSE	RMSE	MAE	MAPE	R^2	CVRMSE
full dataset	RF	25,183.60	158.69	115.22	39.69	0.701	38.93
	kNN	33,276.64	182.42	131.34	47.80	0.605	44.75

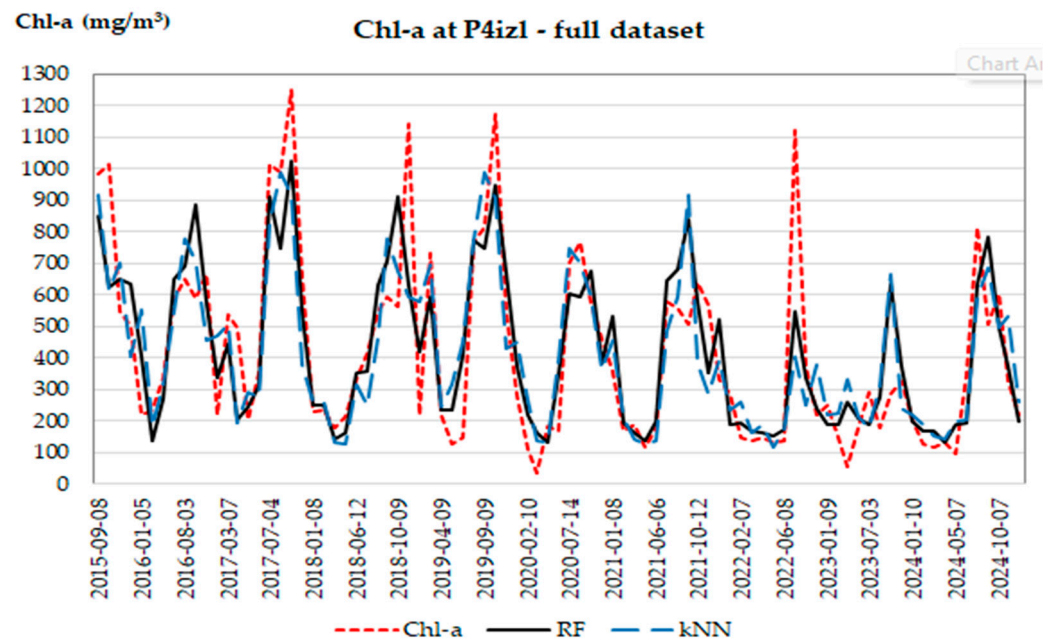


Figure 4. Chl-a measurements compared with RF and kNN model results—full dataset.

The time-series comparison shown in Figure 2 confirms that both models captured the seasonal periodicity and interannual variability of Chl-a concentrations. However, both RF and kNN systematically underestimated extreme bloom events. Observed peak concentrations of approximately 1200 mg/m³ were consistently underestimated, resulting in a pronounced clipping effect. This behavior likely reflects nonlinear optical responses under extreme bloom conditions, where the spectral signal becomes increasingly difficult to interpret reliably. Consequently, despite the relatively high R^2 obtained by RF, the large RMSE values are primarily driven by residual errors during peak bloom periods, particularly above the ~700 mg/m³ threshold.

After excluding extreme observations, the models were retrained and revalidated on the refined dataset to assess the effect of reduced optical variability. Both models exhibited improved predictive precision. The RF model RMSE decreased from 158.693 mg/m³ to 112.994 mg/m³, representing a 28.8% reduction in average prediction error, while MAE declined to 80.706 mg/m³. Although the RF R^2 value decreased slightly from 0.701 to 0.627, both RMSE and MAE decreased substantially after filtering, indicating improved predictive precision despite the modest reduction in R^2 . The kNN model also improved but remained clearly inferior, achieving an R^2 of only 0.449. MSE-based comparison again favored RF for the refined data.

Figure 5 demonstrates that, unlike the results obtained from the full dataset, the models trained on the refined dataset more successfully estimated Chl-a concentrations across the operational range. The RF model in particular captured subtle temporal variations in Chl-a during 2019 and 2020, with close agreement between predicted and observed values during most bloom periods. The reduction in prediction error is consistent with the measured RMSE of 112.99 mg/m³ for RF, Table 9, representing an approximately 29% improvement relative to the unfiltered model. While kNN also showed improvement, its predictions remained less stable, with larger deviations during peak concentration periods (e.g., late 2018 and 2021). These results indicate that RF is better suited for modeling nonlinear Chl-a dynamics within the non-saturated concentration range.

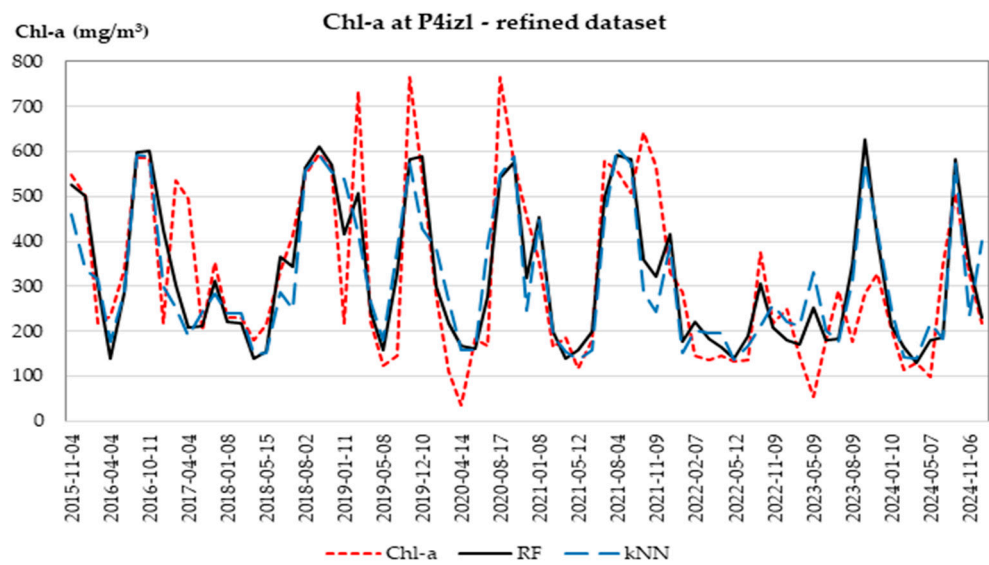


Figure 5. Chl-a measurements compared with RF and kNN model results—refined dataset.

Table 9. Error analysis results using the refined dataset.

	Model	MSE	RMSE	MAE	MAPE	R ²	CVRMSE
refined dataset	RF	12,767.67	112.99	80.71	39.01	0.627	36.39
	kNN	18,902.48	137.49	101.76	47.78	0.449	44.28

3.2.2. Chlorophyll-a at Location P4sr

Trying to establish a model with $R^2 > 0.5$ using unrefined data from the middle of the lake failed; hence, we used the same principle as before, excluding $\text{Chl-a} > 700 \text{ mg/m}^3$. Table 10 shows the results. Only two indices produced models meeting the predefined performance and diagnostic criteria.

Table 10. Error analysis for SLR models developed using refined data at P4sr.

N = 27		Model Performance		Statistical Signif.	Normality Shapiro–Wilk		Autocorrel. Durbin–Watson	
Index	Model Chl – a	R ²	RMSE	p-Value	W	p-Value	DW	p-Value
MSAVI	422.4 + 6089·MSAVI	0.510	128.46	<0.0001	0.97	0.4907	2.20	0.520
OSAVI	430.6 + 2292·OSAVI	0.533	125.51	<0.0001	0.98	0.7367	2.19	0.592

Following the same analytical approach, a comparative visualization of measured parameters and reflectance values was also performed for the central lake location (P4sr), Figure 6. At this site, an observable relationship between Secchi depth and reflectance can be identified: periods of lower water transparency generally correspond to higher reflectance. However, this relationship is less pronounced and more variable than at the outflow location. Although individual parameters such as TSS and Chl-a occasionally align with reflectance peaks, no single variable consistently explains the observed optical variability. This indicates that the spectral response is not governed by a dominant constituent.

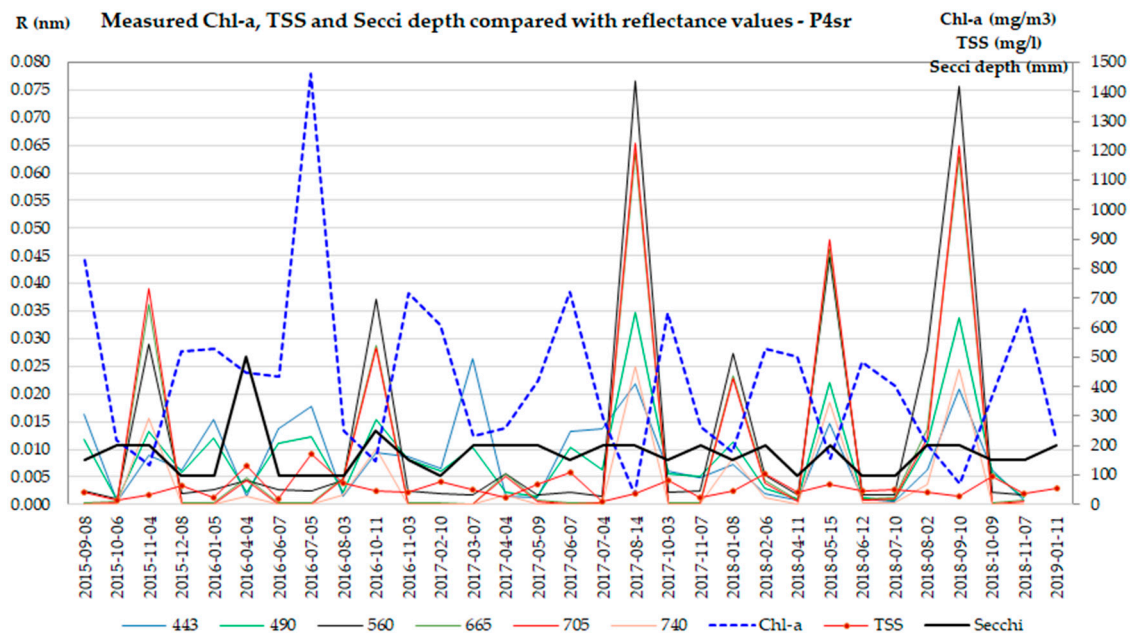


Figure 6. Measurement—reflectance plot at P4sr.

Overall, the results suggest that reflectance at the P4sr location arises from the combined influence of multiple optically active constituents, including suspended matter, phytoplankton, and likely dissolved organic substances, resulting in a moderately complex optical system. Consequently, we investigated the possibility of jointly analyzing all refined observations by combining the P4izl and P4sr datasets into a single dataset representing Sector 4 of Lake Palic. The models listed in Table 11 produced apparently acceptable results ($R^2 > 0.5$ and strong significance); however, all models consistently exhibited non-normal residuals.

Table 11. Error analysis for SLR models developed using refined data at both sites.

N = 119		Model Performance		Statistical Signif.	Normality Shapiro–Wilk		Autocorrel. Durbin–Watson	
Index	Model Chl – a	R^2	RMSE	p -Value	W	p -Value	DW	p -Value
RARSa	$-545.5 + 695.6 \cdot \text{RARSa}$	0.572	178.30	<0.0001	0.95	0.0005	1.94	0.7100
NDVI	$-641.4 - 1689 \cdot \text{NDVI}$	0.547	183.40	<0.0001	0.95	0.0004	1.95	0.7640
Datt99	$1121 - 1001 \cdot \text{Datt99}$	0.519	189.00	<0.0001	0.94	<0.0001	1.88	0.4900
NDI	$-1273 - 3126 \cdot \text{NDI}$	0.538	185.15	<0.0001	0.96	0.0009	1.91	0.6140
ZTM	$1514 - 3653 \cdot \text{ZTM}$	0.532	186.46	<0.0001	0.96	0.0009	1.91	0.5980
Crt2	$1324 - 1198 \cdot \text{Crt2}$	0.546	183.53	<0.0001	0.95	0.0002	1.92	0.6040

The identified behavior confirms that these two parts of the lake act as different systems. As a result, using data within their system, only data from P4izl or P4sr provides statistically and physically consistent models.

Although the pooled models achieved moderate predictive performance ($R^2 = 0.519$ – 0.572), they consistently violated the normality assumption and produced substantially larger prediction errors than the best location-specific models. This indicates that combining observations from P4izl and P4sr introduces additional variability arising from different optical and environmental conditions, thereby reducing model robustness

and interpretability. Consequently, site-specific models were considered more appropriate than a single unified regression framework.

Figure 7 compares temporal in situ Chl-a measurements with model predictions from RF and kNN algorithms using the full dataset. While both models generally follow the temporal trends of the in situ Chl-a concentrations, their accuracy varies. The models tend to underestimate the peak Chl-a concentrations, particularly in late 2017, where in situ measurements exceed 1400 mg/m³. The model predictions mostly do not surpass 800 mg/m³. In contrast, during periods of low Chl-a concentrations, both RF and kNN exhibit a tendency to overestimate the measured values. The kNN model produced smoother predictions, sometimes failing to capture rapid fluctuations present in the measurements. Using the refined dataset led to an unambiguous improvement in model performance (Figure 8). In general, the RF proved slightly more reliable than the kNN model.

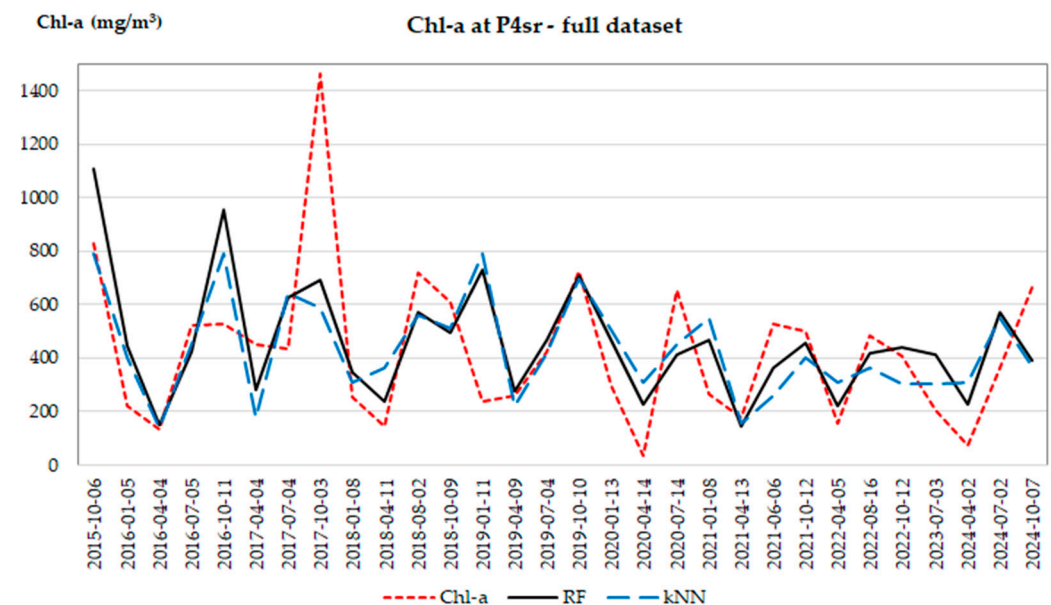


Figure 7. Chl-a measurements compared with RF and kNN model results—full dataset.

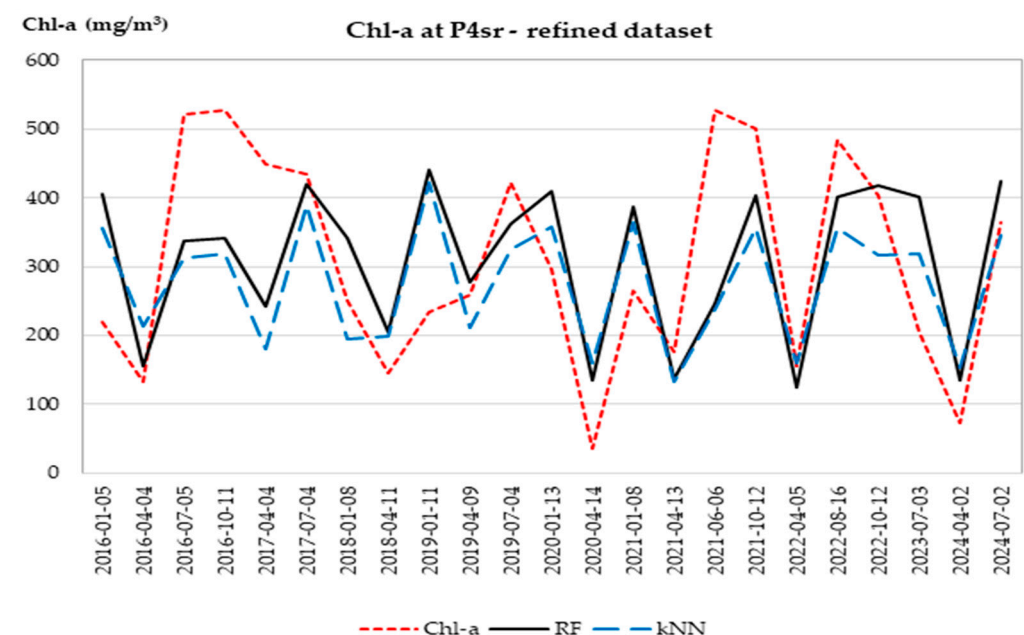


Figure 8. Chl-a measurements compared with RF and kNN model results—refined dataset.

To assess the model complexity, an SLR was developed using OSAVI (Optimized Soil Adjusted Vegetation Index), the strongest linear predictor for the refined dataset. OSAVI showed a strong correlation with Chl-a after filtering ($r = 0.730$). The resulting SLR model ($\text{Chl-a} = 430.6 + 2292 \cdot \text{OSAVI}$) achieved an R^2 of 0.533 and an RMSE of 125.51 mg/m^3 .

Figure 9 presents a comparison of SLR model OSAVI and MSAVI results obtained using the refined data with the in situ Chl-a measurements at P4sr. Despite the small sample size, the OSAVI-based model successfully captured the key trends within the limited range, while MSAVI depicts a notable shift compared to the measurements. The high compliance between the OSAVI-based model and measured data confirms that data filtering successfully isolated a very strong, stable linear relationship at the mid-lake site. Despite the reduced statistical power, the SLR model provided the most stable and precise predictions at P4sr. These findings indicate that, under conditions of severe data limitation, a parsimonious linear model may outperform more complex nonlinear approaches.

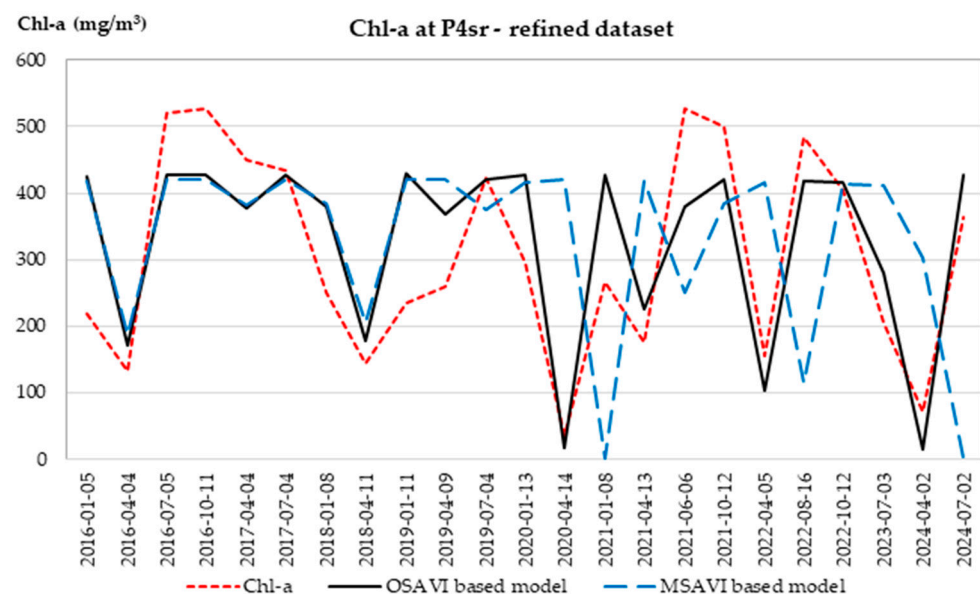


Figure 9. Chl-a measurements compared with SLR predictions using the refined dataset (P4sr).

3.2.3. Total Suspended Solids at Location P4izl

Correlation screening for TSS revealed consistently weak relationships between spectral indices and in situ measurements ($r < 0.3$). This pattern persisted even when using Level-2A bottom-of-atmosphere reflectance in addition to C2RCC-corrected Level-1C imagery. These results indicate that TSS retrieval is unreliable under the prevailing optical conditions of the highly eutrophic Lake Palic.

The TSS resulting SLR models exhibited insufficient predictive performance and were therefore not considered suitable for further analysis. Therefore, an additional analysis using Sentinel-2 L2A bottom-of-atmosphere reflectance products was performed to evaluate whether alternative reflectance inputs and additional near-infrared information could improve the representation of suspended sediment dynamics under the optically complex conditions of Lake Palic.

Qualitative validation at P4izl showed limited predictive skill for both RF and kNN models, Figure 10. RF tended to overestimate TSS at low concentrations ($< 50 \text{ mg/L}$) and to underestimate higher values, whereas kNN showed greater dispersion and systematic underestimation above approximately 30 mg/L . These patterns likely reflect both the limited representation of high-TSS conditions within the training data and the strong optical interference from coexisting water constituents.

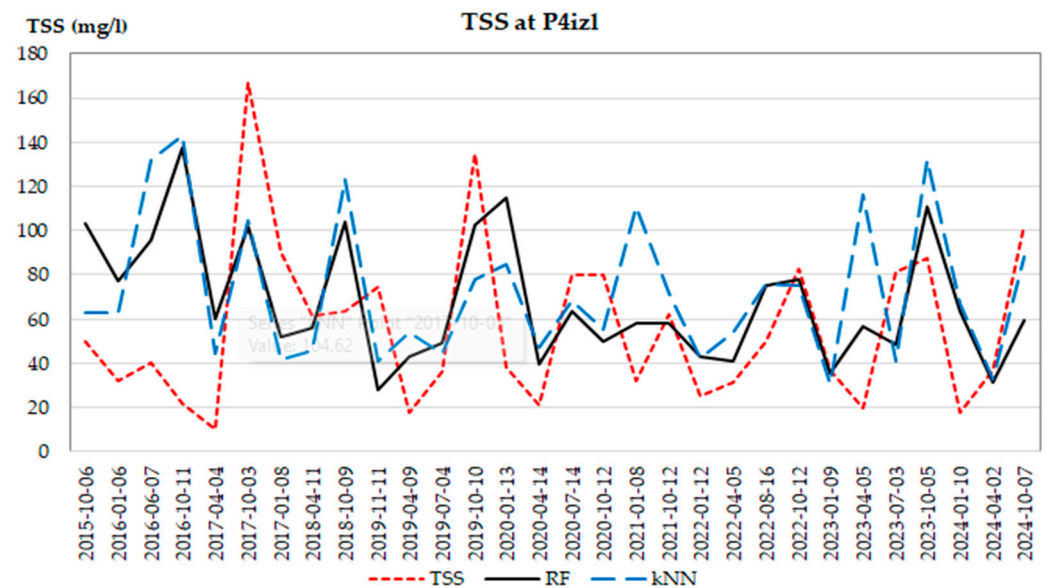


Figure 10. TSS measurements compared with RF and kNN predictions at P4izl.

3.2.4. Total Suspended Solids at Location P4sr

A comparative validation at P4sr revealed improved model performance relative to P4izl. As shown in Figure 11, RF predictions clustered more tightly around the in situ measurements, and kNN also showed reduced dispersion. These results suggest site-specific differences in optical conditions and environmental characteristics that better align with the models' training data.

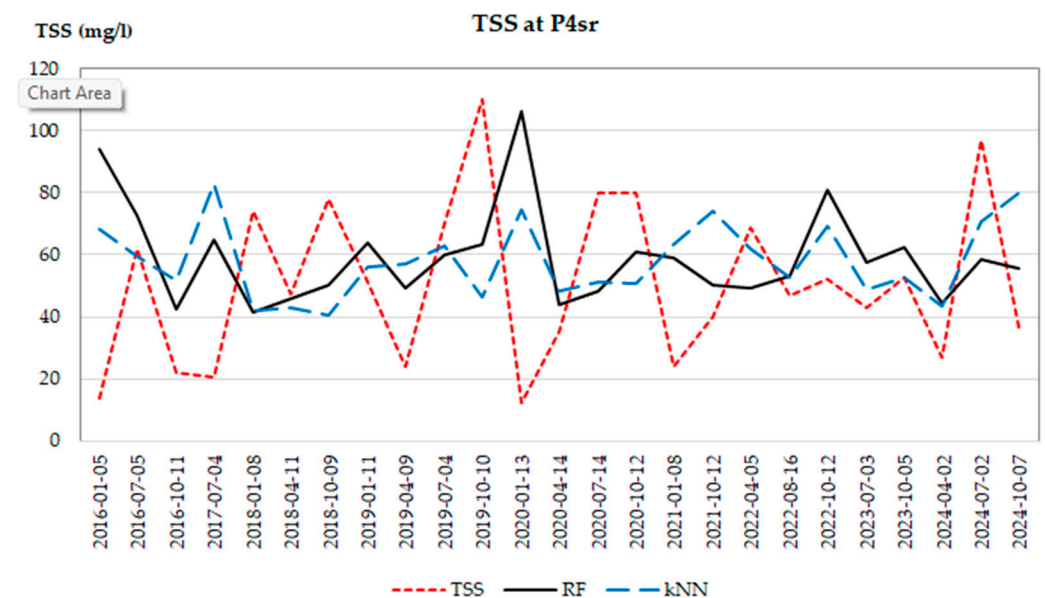


Figure 11. TSS measurements compared with RF and kNN predictions at P4sr.

Overall, RF consistently outperformed kNN at both sites, with higher predictive accuracy at P4sr. The findings highlight the site-specific nature of TSS retrieval and the strong suppressive effect of extreme Chl-a concentrations on the TSS spectral signal. Future work should apply this framework to limnologically similar water bodies with lower Chl-a concentrations, where the TSS scattering signal may be more effectively isolated under less optically complex conditions.

4. Discussion

This study explored the applicability of Sentinel-2 MSI imagery for operational monitoring of chlorophyll-a and total suspended solids in highly eutrophic inland waters. To provide a comprehensive evaluation of model behavior and predictive performance, three different modeling approaches were implemented: simple linear regression, Random Forest, and k-nearest neighbors. The analyzed dataset comprised a continuous nine-year record (2016–2024) of water quality measurements collected from Lake Palic. Preliminary inspection of the Chl-a dataset revealed several extremely high concentrations that exceeded the typical range reported for comparable inland water bodies. Consequently, Chl-a analyses were performed using both the complete dataset and a refined dataset excluding extreme bloom observations, allowing assessment of their influence on spectral relationships and model performance. TSS analyses were conducted separately using the complete dataset and additional reflectance products to evaluate the effect of alternative processing approaches on model development.

The study followed a standard remote sensing workflow consisting of correlation analysis, model calibration, validation, and performance evaluation. Although the primary objective was to establish predictive models for both Chl-a and TSS, the obtained results demonstrated a substantial difference in retrieval success between the two parameters. While Chl-a estimation produced statistically meaningful and operationally relevant models, TSS retrieval remained unreliable across all evaluated approaches.

The initial correlation analysis revealed direction reversals in several spectral indices at location P4izl, particularly for BRI, TCARI, and PSSRc. These reversals are likely associated with nonlinear optical responses under extreme eutrophic conditions, where complex interactions among phytoplankton, suspended matter, and other optically active constituents alter the expected spectral–biophysical relationships. Future studies should investigate whether alternative atmospheric correction approaches and additional spectral information could improve parameter retrieval in highly eutrophic and optically complex inland waters.

The simple linear regression model was implemented as a transparent, physically interpretable baseline method, enabling direct assessment of relationships between spectral indices and measured water quality parameters. Because optically complex inland waters often exhibit nonlinear spectral behavior, RF and kNN were additionally employed to improve predictive flexibility and capture nonlinear relationships within the dataset.

The SLR models were developed exclusively for the estimation of Chl-a values, using the refined datasets because the full datasets violated the normality assumptions for residuals. At P4izl, the strongest SLR model was obtained using the RARSa index, achieving $R^2 = 0.655$ and $RMSE = 110.13 \text{ mg/m}^3$. At P4sr, the best-performing linear model was based on OSAVI, yielding $R^2 = 0.582$ and $RMSE = 100.96 \text{ mg/m}^3$. Interestingly, despite the greater flexibility of the machine learning approaches, the refined SLR models produced the lowest absolute prediction errors at both locations. This indicates that stable linear relationships emerged after the exclusion of extreme observations.

In contrast to the relatively successful Chl-a retrieval, all evaluated approaches produced unreliable TSS predictions. Correlation analysis revealed consistently weak relationships between measured TSS concentrations and the tested spectral indices, regardless of whether Level-1C or Level-2A imagery was used. The strongest relationships remained well below acceptable thresholds for empirical modeling, with maximum values observed for NDTI at P4izl using Level-1C data ($r = -0.389$) and NDWI at P4izl using Level-2A data ($r = 0.366$).

The systematic failure of both standard and custom TSS indices motivated the incorporation of Level-2A bottom-of-atmosphere reflectance products and the development of a hybrid index utilizing additional near-infrared bands. However, even the hybrid index

produced only weak responses, indicating that the optical complexity of Lake Palic could not be adequately resolved using the available Sentinel-2 MSI spectral configuration. The persistent inability to reliably retrieve TSS is likely associated with the combined optical influence of extremely high Chl-a concentrations, suspended particles, and other optically active constituents. Under such conditions, strong phytoplankton-related optical effects may dominate the spectral response, partially obscuring the TSS scattering signal typically expressed in the red and near-infrared spectral regions. Furthermore, the simultaneous influence of CDOM and other OACs introduces additional spectral interference and nonlinearity, further reducing the separability of TSS-related signals. Collectively, these effects limit the applicability of conventional TSS indices and highlight an important challenge for MSI-based TSS retrieval in highly eutrophic and optically complex inland waters.

The results further demonstrated that model performance was strongly governed by spatially variable optical regimes within Lake Palic. The outflow location (P4izl), characterized by greater optical variability, produced the highest-performing Chl-a models after refinement, particularly using the RARSa index and RF regression. In contrast, the mid-lake location (P4sr) exhibited weaker, less stable relationships in the full dataset but showed improved linearity after filtering extreme observations, with OSAVI emerging as the strongest predictor. These findings are consistent with the pooled-model analysis, where combining observations from both locations resulted in persistent non-normal residuals despite moderate predictive performance.

Attempts to construct unified regression models using combined observations from both sampling locations yielded persistent non-normal residuals despite acceptable R^2 values. This indicates that merging observations governed by distinct optical regimes introduces structural heterogeneity that cannot be adequately represented by a single linear relationship. The result further supports the need for site-specific calibration in optically complex inland waters.

Among the evaluated ML approaches, RF produced the strongest overall performance for Chl-a retrieval. At location P4izl, RF achieved $R^2 = 0.701$ and $RMSE = 158.693 \text{ mg/m}^3$ using the full dataset, while the refined dataset produced $R^2 = 0.627$ and $RMSE = 112.994 \text{ mg/m}^3$. At location P4sr, RF yielded substantially lower predictive performance: $R^2 = 0.301$ and $RMSE = 234.491 \text{ mg/m}^3$ for the full dataset, and $R^2 = 0.202$ and $RMSE = 133.568 \text{ mg/m}^3$ for the refined dataset. These results suggest that RF may have benefited from its ability to accommodate complex and potentially nonlinear relationships at the outflow location (P4izl), whereas performance at the mid-lake location (P4sr) remained limited, likely due to weaker and less consistent spectral relationships.

The kNN model showed moderate improvement after data refinement, but consistently underperformed RF. At P4izl, kNN achieved $R^2 = 0.605$ and $RMSE = 182.419 \text{ mg/m}^3$ for the full dataset, while refinement improved the model to $R^2 = 0.449$ and $RMSE = 137.486 \text{ mg/m}^3$. At P4sr, the model produced $R^2 = 0.178$ and $RMSE = 254.384 \text{ mg/m}^3$ for the full dataset and $R^2 = 0.156$ with $RMSE = 137.381 \text{ mg/m}^3$ after refinement. The comparatively weaker performance of kNN likely reflects the sensitivity of distance-based methods to outliers and the limited representativeness of samples in optically heterogeneous environments.

Quantitative evaluation of the TSS machine learning models frequently yielded negative R^2 values, reflecting weak spectral sensitivity to TSS, limited TSS variability, strong optical interference, and limited sample representativeness. Consequently, TSS model performance was primarily assessed through qualitative comparisons between measured and predicted values, rather than relying exclusively on statistical metrics.

From a practical perspective, the developed Chl-a models provide a promising operational framework for systematic water quality monitoring in lakes characterized by frequent bloom events and strong eutrophication pressure. Given the current absence of

a comprehensive satellite-based monitoring framework in Serbian environmental regulation, the presented methodology may support future monitoring strategies to track bloom severity and assess ecological risk in protected inland waters.

5. Conclusions

This study evaluated the applicability of Sentinel-2 MSI imagery for the retrieval of chlorophyll-a (Chl-a) and total suspended solids (TSS) under the highly eutrophic and optically complex conditions of Lake Palic. The evaluated spectral indices demonstrated moderate to strong relationships with Chl-a, particularly after the removal of extreme bloom observations, whereas all tested indices exhibited weak relationships with TSS. These findings indicate that Chl-a can be reliably monitored using spectral-index-based approaches, while TSS retrieval remains challenging under conditions of severe eutrophication and optical complexity.

The developed SLR models produced statistically significant results for Chl-a estimation at both monitoring locations. Among the evaluated indices, RARSa provided the strongest performance at the outflow location (P4izl), whereas OSAVI emerged as the most reliable predictor at the central lake location (P4sr). The substantial reduction in prediction error following data refinement highlights the influence of extreme bloom events on model stability and calibration, consistent with the findings of Ibrahim et al. [26].

Among the machine learning approaches, Random Forest consistently outperformed k-nearest neighbors and achieved the highest overall predictive performance for Chl-a retrieval. These findings further support previous studies [26–28], which demonstrated the ability of machine learning methods to capture complex nonlinear relationships between spectral reflectance and water quality parameters in optically complex aquatic environments. Nevertheless, both RF and kNN showed limitations when predicting extreme bloom conditions.

The assessment of spatial transferability demonstrated that a unified model applicable across both sampling locations could not be established. The results indicate that P4sr and P4izl represent distinct optical regimes influenced by different environmental pressures, requiring site-specific calibration for reliable remote-sensing-based water quality retrieval.

Model validation using R^2 , RMSE, residual diagnostics, and comparison with in situ measurements confirmed the applicability of Sentinel-2 imagery for Chl-a monitoring while simultaneously highlighting the limitations of MSI-based TSS retrieval under highly eutrophic conditions.

Future research should investigate alternative atmospheric correction approaches, including ACOLITE, incorporate synchronized field sampling during satellite overpasses, and evaluate additional Earth observation missions and water quality parameters. Extending the framework to limnologically similar lakes with lower Chl-a concentrations may also improve the feasibility of TSS retrieval.

Author Contributions: Conceptualization, M.M. and M.H.; methodology, M.M. and M.H.; software, M.M.; validation, M.M.; formal analysis, M.M. and K.D.; investigation, M.M.; resources, M.M., M.H. and Z.H.; data curation, M.M.; writing—original draft preparation, M.M.; writing—review and editing, M.H. and K.D.; visualization, M.M. and M.H.; supervision, M.H. and Z.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All utilized data are available at the official web page: <https://subotica.ls.gov.rs/zastita-zivotne-sredine/monitoring-stanja-zivotne-sredine/kvalitet-povrsinskih-voda-natitoriji-grada-subotice/>.

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

References

1. Han, G.; Liu, X. Environmental Effects of Natural Processes and Human Activities on the Water Environment in Watershed. *Water* **2024**, *16*, 2929. [CrossRef]
2. Grimm, N.B.; Gergel, S.E.; McDowell, W.H.; Boyer, E.W.; Dent, C.L.; Groffman, P.; Hart, S.C.; Harvey, J.; Johnston, C.; Mayorga, E.; et al. Merging Aquatic and Terrestrial Perspectives of Nutrient Biogeochemistry. *Oecologia* **2003**, *137*, 485–501. [CrossRef] [PubMed]
3. Xie, H.; Huang, Y.; Chen, Q.; Zhang, Y.; Wu, Q. Prospects for Agricultural Sustainable Intensification: A Review of Research. *Land* **2019**, *8*, 157. [CrossRef]
4. Madjar, R.M.; Vasile Scăteanu, G.; Sandu, M.A. Nutrient Water Pollution from Unsustainable Patterns of Agricultural Systems, Effects and Measures of Integrated Farming. *Water* **2024**, *16*, 3146. [CrossRef]
5. Zheng, T.; Cao, H.; Liu, W.; Xu, J.; Yan, Y.; Lin, X.; Huang, J. Characteristics of Atmospheric Deposition during the Period of Algal Bloom Formation in Urban Water Bodies. *Sustainability* **2019**, *11*, 1703. [CrossRef]
6. Cantonati, M.; Poikane, S.; Pringle, C.M.; Stevens, L.E.; Turak, E.; Heino, J.; Richardson, J.S.; Bolpagni, R.; Borrini, A.; Cid, N.; et al. Characteristics, Main Impacts, and Stewardship of Natural and Artificial Freshwater Environments: Consequences for Biodiversity Conservation. *Water* **2020**, *12*, 260. [CrossRef]
7. City of Subotica. *Plan for the Improvement of the Ecological Condition of Lake Palić and Its Surroundings*; City of Subotica: Subotica, Serbia, 2014. Available online: https://subotica.ls.gov.rs/web/wp-content/uploads/2024/02/8253_2.pdf (accessed on 21 December 2025).
8. Simenunović, J. *Cijanobakterije i Cijanotoksini u Površinskim Vodama Vojvodine*; Univerzitet u Novom Sadu: Novi Sad, Serbia, 2010.
9. World Health Organization. *Guidelines for Safe Recreational Water Environments: Volume 1: Coastal and Fresh Waters*; World Health Organization: Geneva, Switzerland, 2021. Available online: <https://www.who.int/publications/i/item/9241545801> (accessed on 21 December 2025).
10. Wesley, M. Satellite-Based Estimation of Chlorophyll-a Concentration in Turbid Productive Waters. Master's Thesis, University of Nebraska Lincoln, Lincoln, NE, USA, 2009.
11. Schalles, J. *Optical Remote Sensing Techniques to Estimate Phytoplankton Chlorophyll a Concentrations in Coastal*. In *Remote Sensing of Aquatic Coastal Ecosystem Processes*; Richardson, L., LeDrew, E.F., Eds.; Remote Sensing and Digital Image Processing Series 9; Springer: Dordrecht, The Netherlands, 2006; Volume 9, pp. 27–79.
12. Gordon, H.R.; Brown, O.B.; Jacobs, M.M. Computed Relationships between the Inherent and Apparent Optical Properties of a Flat Homogeneous Ocean. *Appl. Opt.* **1975**, *14*, 417–427. [CrossRef] [PubMed]
13. Ferrão-Filho, A.d.S.; Kozłowsky-Suzuki, B. Cyanotoxins: Bioaccumulation and Effects on Aquatic Animals. *Mar. Drugs* **2011**, *9*, 2729–2772. [CrossRef] [PubMed]
14. Dekker, A.G. Detection of Optical Water Quality Parameters for Eutrophic Waters by High Resolution Remote Sensing. Ph.D. Thesis, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands, 1993. Available online: <https://research.vu.nl/en/publications/detection-of-optical-water-quality-parameters-for-eutrophic-water/> (accessed on 1 June 2024).
15. Dorji, P.; Fearn, P.; Broomhall, M. A Semi-Analytic Model for Estimating Total Suspended Sediment Concentration in Turbid Coastal Waters of Northern Western Australia Using MODIS-Aqua 250 m Data. *Remote Sens.* **2016**, *8*, 556. [CrossRef]
16. Zarco-Tejada, P.J.; Miller, J.R.; Morales, A.; Berjón, A.; Agüera, J. Hyperspectral Indices and Model Simulation for Chlorophyll Estimation in Open-Canopy Tree Crops. *Remote Sens. Environ.* **2004**, *90*, 463–476.
17. Yadav, T. *Habitable Exoplanets: A Literature Review of Potential Signatures of Life*; ResearchGate: Berlin, Germany, 2018.
18. Toming, K.; Kutser, T.; Laas, A.; Sepp, M.; Paavel, B.; Nöges, T. First Experiences in Mapping Lake Water Quality Parameters with Sentinel-2 MSI Imagery. *Remote Sens.* **2016**, *8*, 640. [CrossRef]
19. Abbas, M.M.; Melesse, A.M.; Scinto, L.J.; Rehage, J.S. Satellite Estimation of Chlorophyll-a Using Moderate Resolution Imaging Spectroradiometer (MODIS) Sensor in Shallow Coastal Water Bodies: Validation and Improvement. *Water* **2019**, *11*, 1621. [CrossRef]
20. Barraza-Moraga, F.; Alcayaga, H.; Pizarro, A.; Félez-Bernal, J.; Urrutia, R. Estimation of Chlorophyll-a Concentrations in Lanalhue Lake Using Sentinel-2 MSI Satellite Images. *Remote Sens.* **2022**, *14*, 5647. [CrossRef]
21. 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. [CrossRef]

22. 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. [CrossRef]
23. Yan, Y.; Wang, Y.; Yu, C.; Zhang, Z. Multispectral Remote Sensing for Estimating Water Quality Parameters: A Comparative Study of Inversion Methods Using Unmanned Aerial Vehicles (UAVs). *Sustainability* **2023**, *15*, 10298. [CrossRef]
24. Patricio-Valerio, L.; Schroeder, T.; Devlin, M.J.; Qin, Y.; Smithers, S. A Machine Learning Algorithm for Himawari-8 Total Suspended Solids Retrievals in the Great Barrier Reef. *Remote Sens.* **2022**, *14*, 3503. [CrossRef]
25. Pelevin, V.; Koltsova, E.; Molkov, A.; Fedorov, S.; Alymkulov, S.; Kononov, B.; Alymkulova, M.; Jumaliev, K. Regional Models for Sentinel-2/MSI Imagery of Chlorophyll a and TSS, Obtained for Oligotrophic Issyk-Kul Lake Using High-Resolution LIF LiDAR Data. *Remote Sens.* **2023**, *15*, 4443. [CrossRef]
26. Ibrahim, A.; Alkarbi, N.; Alsaadi, M.; Almazrouei, A.; Alshamsi, S.; Hamouda, M.A. Satellite-based machine learning models for chlorophyll-a and TSS retrieval in Abu Dhabi's coastal waters. *Front. Mar. Sci.* **2026**, *13*, 1787597. [CrossRef]
27. Guo, Q.; Wu, X.; Bing, Q.; Pan, Y.; Wang, Z.; Fu, Y.; Wang, D.; Liu, J. Study on Retrieval of Chlorophyll-a Concentration Based on Landsat OLI Imagery in the Haihe River, China. *Sustainability* **2016**, *8*, 758. [CrossRef]
28. Xu, S.; Li, S.; Tao, Z.; Song, K.; Wen, Z.; Li, Y.; Chen, F. Remote Sensing of Chlorophyll-a in Xinkai Lake Using Machine Learning and GF-6 WFV Images. *Remote Sens.* **2022**, *14*, 5136. [CrossRef]
29. City Administration of Subotica. Surface Water Quality in the Territory of the City of Subotica. Available online: <https://subotica.rs.gov.rs/zastita-zivotne-sredine/monitoring-stanja-zivotne-sredine/kvalitet-povrsinskih-voda-na-teritoriji-grada-subotice/> (accessed on 7 February 2026).
30. Horvat, M.; Horvat, Z. Implementation of a monitoring approach: The Palic-Ludas lake system in the Republic of Serbia. *Environ. Monit. Assess.* **2020**, *192*, 2. [CrossRef] [PubMed]
31. European Space Agency. *Sentinel-2 MSI (Multi-Spectral Instrument)*; European Space Agency: Paris, France, 2021. Available online: <https://sentinels.copernicus.eu/web/sentinel/missions/sentinel-2> (accessed on 21 December 2025).
32. European Space Agency. *SNAP—ESA Sentinel Application Platform*, version 10.0.0; European Space Agency: Paris, France, 2024. Available online: <https://step.esa.int/main/toolboxes/snap/> (accessed on 21 December 2025).
33. Morel, A.; Prieur, L. Analysis of variations in ocean color. *Limnol. Oceanogr.* **1977**, *22*, 709–722. [CrossRef]
34. Bailey, S.W.; Werdell, P.J. A multi-sensor approach for the on-orbit validation of ocean color satellite data products. *Remote Sens. Environ.* **2006**, *102*, 12–23. [CrossRef]
35. 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. *J. Plant Physiol.* **2003**, *160*, 271–282. [CrossRef] [PubMed]
36. Peñuelas, J.; Baret, F.; Filella, I. Semi-empirical Indexes to Assess Carotenoids/Chlorophyll a Ratio from Leaf Spectral Reflectance. *Photosynthetica* **1995**, *31*, 221–230.
37. Kim, M.S.; Daughtry, C.S.T.; Chappelle, E.W.; McMurtrey, J.E.; Walthall, C.L. The use of high spectral resolution bands for estimating absorbed photosynthetically active radiation (A par). In Proceedings of the 6th International Symposium on Physical Measurements and Signatures in Remote Sensing, Val D'Isere, France, 17–21 January 1994; pp. 299–306. Available online: [https://www.semanticscholar.org/paper/The-use-of-high-spectral-resolution-bands-for-\(A-Kim-Daughtry/4e13fcc53f78c6b95ba163a0cf448a553e46ce23?utm_source=direct_link](https://www.semanticscholar.org/paper/The-use-of-high-spectral-resolution-bands-for-(A-Kim-Daughtry/4e13fcc53f78c6b95ba163a0cf448a553e46ce23?utm_source=direct_link) (accessed on 16 August 2025).
38. Chappelle, E.W.; Kim, M.S.; McMurtrey, J.E. Ratio analysis of reflectance spectra (RARS) as a novel method for the remote estimation of chlorophyll a and b content in higher plant leaves. *Remote Sens. Environ.* **1992**, *39*, 219–229. [CrossRef]
39. Carter, G.A. Ratios of Leaf Reflectances in Narrow Wavebands as Indicators of Plant Stress. *Int. J. Remote Sens.* **1994**, *15*, 697–703. [CrossRef]
40. Zarco-Tejada, P.J.; Miller, J.R.; Noland, T.L.; Angel, J.N.N.; Mohammed, G.H. Needle chlorophyll content estimation through model inversion using hyperspectral data from boreal conifer forest canopies. *Remote Sens. Environ.* **2001**, *77*, 295–307. [CrossRef]
41. Gitelson, A.A.; Kaufman, Y.J.; Stark, R.; Rundquist, D. Vegetation and soil lines in visible spectral space: A concept and technique for remote estimation of vegetation fraction. *Int. J. Remote Sens.* **2002**, *23*, 2537–2562.
42. Haboudane, D.; Miller, J.R.; Pattey, E.; Zarco-Tejada, P.J.; Strachan, I.B. Integrated narrow-band vegetation indices for prediction of crop chlorophyll content for application to precision agriculture. *Remote Sens. Environ.* **2002**, *81*, 211–225. [CrossRef]
43. Jordan, C.F. Derivation of Leaf-Area Index from Quality of Light on the Forest Floor. *Ecology* **1969**, *50*, 663–666. [CrossRef]
44. Gitelson, A.A.; Merzlyak, M.N. Remote Sensing of Chlorophyll Concentration in Higher Plant Leaves. *Adv. Space. Res.* **1998**, *22*, 689–692. [CrossRef]
45. McNairn, H.; Protz, R. Mapping corn residue cover on agricultural fields in Oxford County, Ontario using Thematic Mapper. *Can. J. Remote Sens.* **1993**, *19*, 152–158. [CrossRef]
46. Daughtry, C.S.T.; Walthall, C.L.; Kim, M.S.; Brown de Colstoun, E.; McMurtrey, J.E. Estimating Corn Leaf Chlorophyll Concentration from Leaf and Canopy Reflectance. *Remote Sens. Environ.* **2000**, *74*, 229–239. [CrossRef]

47. Datt, B. A new reflectance index for remote estimation of chlorophyll content in higher plants. *Remote Sens. Environ.* **1999**, *68*, 321–328. [CrossRef]
48. Rouse, J.W.; Haas, R.H.; Schell, J.A.; Deering, D.W. Monitoring vegetation systems in the Great Plains with ERTS. In *Proceedings of the Third Earth Resources Technology Satellite-1 Symposium, Washington, DC, USA, 10–14 December 1973*; Volume 1: Technical Presentations, NASA SP-351; National Aeronautics and Space Administration: Washington, DC, USA, 1974; pp. 309–317. Available online: <https://ntrs.nasa.gov/api/citations/19740022614/downloads/19740022614.pdf> (accessed on 19 September 2025).
49. Mo, J.; Tian, Y.; Wang, J.; Zhang, Q.; Zhang, Y.; Tao, J.; Lin, J. Remote sensing inversion of suspended particulate matter in the estuary of the Pinglu Canal in China based on machine learning algorithms. *Front. Mar. Sci.* **2024**, *11*, 1473104. [CrossRef]
50. Hossain, A.; Jia, Y.; Chao, X. Development of remote sensing based index for estimating/mapping suspended sediment concentration in river and lake environments. In *Proceedings of the 8th International Symposium on ECOHYDRAULICS (ISE 2010)*, Seoul, Republic of Korea, 12–16 September 2010; pp. 578–585.
51. Montalvo, L.G. Spectral Analysis of Suspended Material in Coastal Waters: A Comparison Between Band Math Equations. 2010. Available online: https://www.uprm.edu/gerslab/wp-content/uploads/sites/214/2023/07/montalvo_2011.pdf (accessed on 22 September 2025).
52. Pérez-González, R.; Sòria-Perpinyà, X.; Soria, J.; Delegido, J.; Urrego, P.; Sendra Cabrera, M.D.; Ruiz-Verdu, A.; Vicente, E.; Moreno, J. Phycocyanin Monitoring in Some Spanish Water Bodies with Sentinel-2 Imagery. *Water* **2021**, *13*, 2866. [CrossRef]
53. Ansper, A.; Alikas, K. Retrieval of Chlorophyll a from Sentinel-2 MSI Data for the European Union Water Framework Directive Reporting Purposes. *Remote Sens.* **2019**, *11*, 64. [CrossRef]
54. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
55. Cover, T.M.; Hart, P.E. Nearest neighbor pattern classification. *IEEE Trans. Inf. Theory* **1967**, *13*, 21–27. [CrossRef]
56. Fix, E.; Hodges, J.L. Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties. *Int. Stat. Rev.* **1989**, *57*, 238–247. [CrossRef]
57. Azubuike, B.N.; Chlingaryan, A.; Correa-Luna, M.; Clark, C.E.F.; Garcia, S.C. Data Augmentation and Interpolation Improves Machine Learning-Based Pasture Biomass Estimation from Sentinel-2 Imagery. *Remote Sens.* **2025**, *17*, 3787. [CrossRef]
58. Pascal, C.; Ferrant, S.; Peña-Luque, S.; Selles, A.; Marechal, J.C.; Wijmer, T.; Gascoin, S.; Merlin, O. Monitoring Indian ungauged small reservoirs volume from remote sensing: Feasibility, bias and perspectives. *PLoS Water* **2024**, *3*, e0000260. [CrossRef]
59. Song, J.; Liu, X.; Adingo, S.; Guo, Y.; Li, Q. A Comparative Analysis of Remote Sensing Estimation of Aboveground Biomass in Boreal Forests Using Machine Learning Modeling and Environmental Data. *Sustainability* **2024**, *16*, 7232. [CrossRef]

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.